



# **EURALEX XIX**

**Congress of the  
European Association  
for Lexicography**

**Lexicography for inclusion**

**7-9 September 2021**  
**Virtual**

[www.euralex2020.gr](http://www.euralex2020.gr)

**Proceedings Book  
Volume 1**

Edited by Zoe Gavrilidou, Maria Mitsiaki, Asimakis Fliatouras



## **EURALEX Proceedings**

ISSN 2521-7100

ISBN 978-618-85138-1-5

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100

e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

**2020 Edition**





In Memory of  
**Tanneke Schoonheim**  
(1965-2020)



**Programme Committee**

Zoe Gavriilidou  
 Maria Mitsiaki  
 Tinatin Margalitadze  
 Gilles-Maurice de Schryver  
 Simon Krek  
 Annette Klosa  
 Tanara Zingano Kuhn  
 George Xydopoulos

**Reviewers**

Andrea Abel  
 Arleta Adamska-Sałaciak  
 Anna Anastasiadis  
 Battaner Arias  
 Xavier Banco  
 Gilles-Maurice de Schryver  
 Janet DeCesaris  
 Ioannis Deligiannis  
 Dziemianko Dziemianko  
 Angeliki Efthymiou  
 Asimakis Fliatouras  
 Thierry Fontenelle  
 Angeliki Fotopoulou  
 Zoe Gavriilidou  
 Alexander Geyken  
 Rufus Gouws  
 Sylviane Granger  
 Oddrun Grønnevik  
 Patrick Hanks  
 Ulrich Heid  
 Anna Iordanidou  
 Miloš Jakubiček  
 Jelena Kallas  
 Marianna Katsogiannou  
 Ilan Kernerman  
 Annette Klosa



Dimitra Koukouzika  
Simon Krek  
Tita Kyriakopoulou  
Lothar Lemnitzer  
Robert Lew  
Marie-Claude L'Homme  
Phillip Louw  
Carla Mareello  
Tina Margalitadze  
George Mikros  
Maria Mitsiaki  
Rosamund Moon  
Argyro Moustaki  
Magali Paquot  
Stellios Piperidis  
Nataschia Ralli  
Michael Rundell  
Tanneke Schoonheim  
Max Silbertzein  
Elsabe Taljard  
Carole Tiberius  
Lars Trap-Jensen  
Anna Vacalopoulou  
Geoffrey Williams  
George Xydopoulos  
Tanara Zingano Kuhn



## Table of Contents

# PAPERS ..... 9

## THE DICTIONARY-MAKING PROCESS

<b>THE MAKING OF THE DIRETES DICTIONARY: HOW TO DEVELOP AN E-DICTIONARY BASED ON AUTOMATIC INHERITANCE</b> .....	13
<i>Barrios M. A.</i>	

<b>DICTIONNAIRE DES FRANCOPHONES - A NEW PARADIGM IN FRANCOPHONE LEXICOGRAPHY</b> .....	23
<i>Dolar K., Steffens M., Gasparini N.</i>	

<b>REDUCE, REUSE, RECYCLE: ADAPTATION OF SCIENTIFIC DIALECT DATA FOR USE IN A LANGUAGE PORTAL FOR SCHOOLCHILDREN</b> .....	31
<i>Ježovnik J., Kenda-Jež K., Škofic J.</i>	

## RESEARCH ON DICTIONARY USE

<b>"GAME OF WORDS": PLAY THE GAME, CLEAN THE DATABASE</b> .....	41
<i>Arhar Holdt Š., Logar N., Pori E., Kosem I.</i>	

<b>UNDERSTANDING ENGLISH DICTIONARIES: THE EXPERIENCE FROM A MASSIVE OPEN ONLINE COURSE</b> .....	51
<i>McGillivray B., Nesi H., Rundell M., Süle K.</i>	

## LEXICOGRAPHY AND LANGUAGE TECHNOLOGIES

<b>TOWARDS AUTOMATIC LINKING OF LEXICOGRAPHIC DATA: THE CASE OF A HISTORICAL AND A MODERN DANISH DICTIONARY</b> .....	63
<i>Ahmadi S., Nimb S., McCrae J., Sørensen N.</i>	

<b>INTERLINKING SLOVENE LANGUAGE DATASETS</b> .....	73
<i>Bajčetić L., Declerck T.</i>	

<b>KARTU-VERBS: A SEMANTIC WEB BASE OF INFLECTED GEORGIAN VERB FORMS TO BYPASS GEORGIAN VERB LEMMATIZATION ISSUES</b> .....	81
<i>Ducassé M.</i>	

<b>MAKING DICTIONARIES VISIBLE, ACCESSIBLE, AND REUSABLE: THE CASE OF THE GREEK CONCEPTUAL DICTIONARY API</b> .....	91
<i>Giouli V., Sidiropoulos N.F.</i>	

<b>PRINCIPLED QUALITY ESTIMATION FOR DICTIONARY SENSE LINKING</b> .....	101
<i>Grosse J., Sauri R.</i>	

<b>DETERMINING DIFFERENCES OF GRANULARITY BETWEEN CROSS-DICTIONARY LINKED SENSES</b> .....	109
<i>Kouvara E., González M., Grosse J., Sauri R.</i>	

<b>A TYPOLOGY OF LEXICAL AMBIFORMS IN ESTONIAN</b> .....	119
<i>Vainik E., Paulsen G., Lohk A.</i>	

## LEXICOGRAPHY AND CORPUS LINGUISTICS

<b>BY THE WAY, DO DICTIONARIES DEAL WITH ONLINE COMMUNICATION? ON THE USE OF META-COMMUNICATIVE CONNECTORS IN CMC COMMUNICATION AND THEIR REPRESENTATION IN LEXICOGRAPHIC RESOURCES FOR GERMAN</b> .....	133
<i>Abel A.</i>	

<b>ΔΗΜΙΟΥΡΓΙΑ ΗΛΕΚΤΡΟΝΙΚΗΣ ΛΕΞΙΚΟΓΡΑΦΙΚΗΣ ΒΑΣΗΣ ΓΙΑ ΤΟ ΠΕΡΙΘΩΡΙΑΚΟ ΛΕΞΙΛΟΓΙΟ ΤΗΣ ΝΕ: ΑΡΧΙΚΟΣ ΣΧΕΔΙΑΣΜΟΣ</b> .....	141
<i>Χριστοπούλου Κ., Ξυδόπουλος Ι. Γ.</i>	

<b>«ΤΑ ΣΤΑΛΘΕΝΤΑ Η ΤΑ ΣΤΑΛΜΕΝΑ ΜΗΝΥΜΑΤΑ;» – ΑΠΟΛΙΘΩΜΑΤΑ ΤΩΝ ΑΡΧΑΙΩΝ ΜΕΤΟΧΩΝ ΣΤΑ ΣΥΓΧΡΟΝΑ ΛΕΞΙΚΑ ΚΑΙ ΣΤΑ ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ</b> .....	151
<i>Ιορδανίδου Α.</i>	

<b>SEMANTIC RELATIONS IN THE THESAURUS OF ENGLISH IDIOMS: A CORPUS-BASED STUDY</b> .....	157
<i>Giztova G., Ismagilova L.</i>	

<b>INTENSIFIERS/MODERATORS OF VERBAL MULTIWORD EXPRESSIONS IN MODERN GREEK</b> .....	163
<i>Mexa M., Markantonatou S.</i>	

<b>BUILDING A CONTROLLED LEXICON FOR AUTHORIZING AUTOMOTIVE TECHNICAL DOCUMENTS</b> .....	171
<i>Miyata R., Sugino H.</i>	

## BI- AND MULTILINGUAL LEXICOGRAPHY

<b>RECONCEPTUALIZING LEXICOGRAPHY: THE BROAD UNDERSTANDING</b> .....	183
<i>Leroyer P., Köhler Simonsen H.</i>	

<b>A MORPHO-SEMANTIC DIGITAL DIDACTIC DICTIONARY FOR LEARNERS OF LATIN AT EARLY STAGES</b> .....	193
<i>Márquez Cruz M., Fernández-Pampillón A.Mª.</i>	



<b>ΕΝΔΟΓΛΩΣΣΙΚΗ ΚΑΙ ΔΙΑΓΛΩΣΣΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΣΥΝΩΝΥΜΙΑΣ. ΣΥΓΚΡΙΤΙΚΗ ΜΕΛΕΤΗ ΛΟΓΟΤΕΧΝΙΚΩΝ ΜΕΤΑΦΡΑΣΕΩΝ ΜΕ ΔΙΓΛΩΣΣΑ ΚΑΙ ΜΟΝΟΓΛΩΣΣΑ ΛΕΞΙΚΑ.....</b>	<b>203</b>
--	------------

Povrogiάννη A.

<b>TOWARDS THE SUPERDICTIONARY: LAYERS, TOOLS AND UNIDIRECTIONAL MEANING RELATIONS.....</b>	<b>215</b>
---	------------

Tavast A., Koppel K., Langemets M., Kallas J.

## LEXICOGRAPHY FOR SPECIALISED LANGUAGES, TERMINOLOGY AND TERMINOGRAPHY

<b>LEMMA SELECTION AND MICROSTRUCTURE: DEFINITIONS AND SEMANTIC RELATIONS OF A DOMAIN-SPECIFIC E-DICTIONARY OF THE MATHEMATICAL FIELD OF GRAPH THEORY.....</b>	<b>227</b>
--	------------

Kruse T., Heid U.

<b>A THEMATIC DICTIONARY FOR DOCTOR-PATIENT COMMUNICATION: THE PRINCIPLES AND PROCESS OF COMPILATION.....</b>	<b>235</b>
---	------------

Kudashev I.S., Semenova O.V.

## LEXICOGRAPHY AND SEMANTIC THEORY

<b>ΤΟΠΩΝΥΜΙΑ ΤΗΣ ΕΛΛΗΝΙΚΗΣ ΚΑΙ Η ΣΧΕΣΗ ΤΟΥΣ ΜΕ ΤΗ ΝΕΟΕΛΛΗΝΙΚΗ ΓΛΩΣΣΙΚΗ ΕΙΚΟΝΑ ΤΟΥ ΚΟΣΜΟΥ.....</b>	<b>247</b>
---	------------

O.B. Bobrova

<b>DEFINITIONS OF THE OXFORD ENGLISH DICTIONARY AND EXPLANATORY COMBINATORIAL DICTIONARY OF I. MEL'ČUK.....</b>	<b>255</b>
---	------------

Margalitadze T.

<b>FRAME SEMANTICS IN THE SPECIALIZED DOMAIN OF FINANCE: BUILDING A TERMBASE TO AID TRANSLATION.....</b>	<b>263</b>
--	------------

Pilitsidou V., Giouli V.

## PHRASEOLOGY AND COLLOCATION

<b>LE TRAITEMENT DES PROVERBES DANS LES DICTIONNAIRES EXPLICATIFS ROUMAINS DU XIX<sup>e</sup> SIÈCLE.....</b>	<b>275</b>
---	------------

Aldea M.

<b>THE INTERACTION OF ARGUMENT STRUCTURES AND COMPLEX COLLOCATIONS: ROLE AND CHALLENGES IN LEARNER'S LEXICOGRAPHY.....</b>	<b>285</b>
--	------------

Giacomini L., DiMuccio-Failla P., Lanzi E.

<b>EVALUATION OF VERB MULTIWORD EXPRESSIONS DISCOVERY MEASUREMENTS IN LITERATURE CORPORA OF MODERN GREEK.....</b>	<b>295</b>
---	------------

Stamou V., Malli M., Takorou P., Xylogianni A., Markantonatou S.

## HISTORICAL AND SCHOLARLY LEXICOGRAPHY AND ETYMOLOGY

<b>CREATING A DTD TEMPLATE FOR GREEK DIALECTAL LEXICOGRAPHY: THE CASE OF THE HISTORICAL DICTIONARY OF THE CAPPADOCIAN DIALECT.....</b>	<b>305</b>
--	------------

Karasimos A., Manolessou I., Melissaropoulou D.

<b>JOHN PICKERING'S VOCABULARY (1816) RECONSIDERED: AMERICA'S EARLIEST PHILOLOGICAL EXPLORATION OF LEXICOGRAPHY.....</b>	<b>315</b>
--	------------

Miyoshi K.

<b>STUDYING LANGUAGE CHANGE THROUGH INDEXED AND INTERLINKED DICTIONARIES.....</b>	<b>321</b>
---	------------

Ore C.-E., Grønvik O.

## LEXICOLOGICAL ISSUES OF LEXICOGRAPHICAL RELEVANCE

<b>WHEN NEOLOGISMS DON'T REACH THE DICTIONARY: OCCASIONALISMS IN SPANISH.....</b>	<b>333</b>
---	------------

Bueno Ruiz P.J.

<b>ARABIC LOANWORDS IN ENGLISH: A LEXICOGRAPHICAL APPROACH.....</b>	<b>343</b>
---	------------

Fournier P., Latrache R.

<b>LOANBLENDS IN THE SPEECH OF GREEK HERITAGE SPEAKERS: A CORPUS-BASED LEXICOLOGICAL APPROACH.....</b>	<b>351</b>
--	------------

Gavriilidou Z., Mitits L.

## REPORTS ON LEXICOGRAPHICAL AND LEXICOLOGICAL PROJECTS

<b>INVENTORY OF NEW ROMANIAN LEXEMES AND MEANINGS ATTESTED ON THE INTERNET.....</b>	<b>363</b>
---	------------

Barbu A.M., Lupu I., Stoica-Dinu O., Teleoacă D.L., Toroipan T.

<b>LBC-DICTIONARY: A MULTILINGUAL CULTURAL HERITAGE DICTIONARY. DATA COLLECTION AND DATA PREPARATION.....</b>	<b>371</b>
---	------------

Farina A., Flinz C.

<b>TO DISCRIMINATE BETWEEN DISCRIMINATION AND INCLUSION: A LEXICOGRAPHER'S DILEMMA.....</b>	<b>381</b>
---	------------

Petersson S., Sköldbberg E.

<b>THE MORFFLEX DICTIONARY OF CZECH AS A SOURCE OF LINGUISTIC DATA.....</b>	<b>387</b>
---	------------

Štěpánková B., Mikulová M., Hajič J.

<b>ANNOUNCING THE DICTIONARY: FRONT MATTER IN THE THREE EDITIONS OF FURETIÈRE'S DICTIONNAIRE UNIVERSEL.....</b>	<b>393</b>
---	------------

Williams G., Galleron I., Stincone C.



**TERMINOLOGY AND TERMINOGRAPHY**

<b>TERM VARIATION IN TERMINOGRAPHIC RESOURCES: A REVIEW AND A PROPOSAL</b> .....	405
<i>Cabezas-García M., León-Araúz P.</i>	

<b>REVISITING POLYSEMY IN TERMINOLOGY</b> .....	415
<i>L'Homme M.-C.</i>	

**LEXICOGRAPHY FOR SPECIAL NEEDS**

<b>SIGN LANGUAGE CORPORA AND DICTIONARIES: A MULTIDIMENSIONAL CHALLENGE</b> .....	427
<i>Vacalopoulou A.</i>	

**POSTERS ..... 435****LEXICOGRAPHY AND LANGUAGE TECHNOLOGIES**

<b>LEARNING DICTIONARY SKILLS FROM GREEK EFL COURSEBOOKS: HOW LIKELY?</b> .....	439
<i>Dalpanagioti Th.</i>	

**LEXICOGRAPHY AND CORPUS LINGUISTICS**

<b>THE GREEK CHILDREN SPOKEN LANGUAGE CORPUS (GCSL CORPUS) / ΤΟ ΠΡΟΦΟΡΙΚΟ ΣΩΜΑ ΚΕΙΜΕΝΩΝ ΕΛΛΗΝΟΦΩΝΩΝ ΠΑΙΔΙΩΝ (ΠΣΚΕΠ): ΠΑΡΟΥΣΙΑΣΗ, ΕΦΑΡΜΟΓΕΣ ΚΑΙ ΠΡΟΟΠΤΙΚΕΣ</b> .....	449
<i>Motsiou E.</i>	

**LEXICOGRAPHY FOR SPECIALISED LANGUAGES, TERMINOLOGY AND TERMINOGRAPHY**

<b>AUDIO RECORDINGS IN A SPECIALIZED DICTIONARY: A BILINGUAL TRANSLATION AND PHRASE DICTIONARY OF MEDICAL TERMS</b> .....	457
<i>Sviķe S., Šķirmante K.</i>	

**HISTORICAL AND SCHOLARLY LEXICOGRAPHY AND ETYMOLOGY**

<b>PAPER QUOTATION SLIPS TO THE ELECTRONIC DICTIONARY OF THE 17TH- AND 18TH-CENTURY POLISH - DIGITAL INDEX AND ITS INTEGRATION WITH THE DICTIONARY</b> .....	465
<i>Bilińska-Brynk J., Rodek E.</i>	

<b>THE ELECTRONIC DICTIONARY OF THE 17TH- AND 18TH-CENTURY POLISH - TOWARDS THE OPEN FORMULA ASSET OF THE HISTORICAL VOCABULARY</b> .....	471
<i>Bronikowska R., Majdak M., Wiecezorek A., Żółtak M.</i>	

**REPORTS ON LEXICOGRAPHICAL AND LEXICOLOGICAL PROJECTS**

<b>THE DEVELOPMENT OF THE OPEN DICTIONARY OF CONTEMPORARY SERBIAN LANGUAGE USING CROWDSOURCING TECHNIQUES</b> .....	479
<i>Lazić Konjik I., Milenković A.</i>	

<b>THE NEW ONLINE ENGLISH-GEORGIAN MARITIME DICTIONARY PROJECT. CHALLENGES AND PERSPECTIVES.</b> .....	485
<i>Tenieshvili A.</i>	

<b>ISSUES IN LINKING A THESAURUS OF MACEDONIAN AND THRACIAN GASTRONOMY WITH THE LINGUAL SYSTEM</b> .....	493
<i>Toraki K., Markantonatou S., Vacalopoulou A., Minos P., Pavlidis G.</i>	

**SOFTWARE DEMONSTRATIONS ..... 499****LEXICOGRAPHY AND LANGUAGE TECHNOLOGIES**

<b>XD-AT: A CROSS-DICTIONARY ANNOTATION TOOL</b> .....	503
<i>González M., Buxton C., Saurí R.</i>	

<b>AUGMENTED WRITING AND LEXICOGRAPHY: A SYMBIOTIC RELATIONSHIP?</b> .....	509
<i>Köhler Simonsen H.</i>	

<b>IDEOMANIA AND GAMIFICATION ADD-ONS FOR APP DICTIONARIES</b> .....	515
<i>Caruso V., Monti J., Andrisani A., Beatrice B., Contento F., De Tommaso Z., Ferrara F., Menniti A.</i>	

**LEXICOGRAPHY AND CORPUS LINGUISTICS**

<b>SKEMA: A NEW TOOL FOR CORPUS-DRIVEN LEXICOGRAPHY</b> .....	523
<i>Baisa V., Tiberius C., Ježek E., Colman L., Marini C., Romani E.</i>	

<b>CROATPAS: A LEXICOGRAPHIC RESOURCE FOR CROATIAN VERBS AND ITS POTENTIAL FOR CROATIAN LANGUAGE TEACHING</b> .....	529
<i>Marini C., Ježek E.</i>	

**INDEX OF AUTHORS ..... 535**



## Foreword

The *Euralex XIX International Conference* is organized by the SynMorPhoSe laboratory of the Department of Greek Philology, Democritus University of Thrace, and was scheduled to be held from 8 to 12 September 2020, in Alexandroupolis, Greece. The motto of XIX Euralex is *Lexicography for Special Needs*, highlighting the demand for compiling dictionaries accessible to the general public, with a view to providing equal opportunities to all dictionary users.

Given the unprecedented circumstances caused by COVID-19, a joint decision was made by the Euralex board members and the congress organizing committee to reschedule the event for 7 to 11 September 2021 and to enable the authors to publish their papers in two separate volumes, the first scheduled to be released in November 2020 and the second in August 2021, just before the congress.

The proceedings offer a unique opportunity to lexicographers to submit their original work, present and discuss the most significant research findings, engage in a scientific dialogue that deepens our knowledge in lexicography or disseminate new ideas and lexicographic practices.

This first volume of the *Euralex XIX Proceedings* is dedicated, as a mark of esteem for her scientific career, to the memory of our dear colleague Tanneke Schoonheim, who passed away unexpectedly on 25 August to our great dismay and sorrow. Tanneke was one of the nicest, warmest, most talented and effective colleagues and a key figure of the executive board of Euralex. We will miss her deeply.

This volume includes 44 papers, 8 posters and 5 software demonstrations. All submissions have been blind-reviewed by two independent reviewers. In case of doubt, a third independent opinion was asked for. As in previous congresses, contributions were submitted on various topics of lexicography, including, but not limited to, the following fields:

- The Dictionary-Making Process
- Research on Dictionary Use
- Lexicography and Language Technologies
- Lexicography and Corpus Linguistics
- Bi- and Multilingual Lexicography
- Lexicography for Specialised Languages, Terminology and Terminography
- Lexicography of Lesser Used languages
- Phraseology and Collocation
- Historical Lexicography and Etymology
- Lexicological Issues of Lexicographical Relevance
- Reports on Lexicographical and Lexicological Projects
- Terminology and Terminography
- Lexicography for Special Needs

The organizing committee would like to thank all authors for submitting their papers to be included in this volume and all the esteemed colleagues who accepted to review the papers. We are also grateful to the colleagues who participated in the hard work of the EURALEX 2020 programme committee at the beginning of March 2020, just a few days before local lockdowns due to COVID-19.

As the chair of the congress, I would like to acknowledge the precious work of the members of the organizing committee who joined efforts with me to make this first volume of the *Euralex XIX Proceedings* possible: Elina Chadjipapa, Cryssa Dourou, Asimakis Fliatouras, Spyridon Kiosses, Lydia Mitits, Maria Mitsiaki and Stavroula Mavrommatidou.

**Zoe Gavriilidou**

Chair of *Euralex XIX International Conference*,

November 2020









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**The Dictionary-Making Process**







# The Making of the *Diretes* Dictionary: how to develop an e-dictionary based on automatic inheritance

Barrios M. A.

Complutense University, Spain

## Abstract

*DiRetEs* is a Spanish monolingual e-dictionary that contains around 100,000 collocations and semantic relations formalized by means of Lexical Functions (LFs). LFs are different formulas, each one appropriate for a different group of collocations or lexical-semantic relations. This dictionary is based on *BADELE.3000* database. The peculiarities of this database are: a) it was built based on a map of semantic labels, a sort of hyperons of the lemmas; and b) it was designed to implement two principles: the principle of lexical inheritance which claims that most of the words sharing a hyperonym (such as *emotion*) could be present in similar collocations (*to feel joy, sadness, envy, etc.*); and the principle of the domain of LFs which claims that the analysis of LFs domains (which means, the set of words this LF was created for) is useful to predict collocations. The combination of both principles in the design of the database allows the lexicographer to automatically obtain new sets of collocations described by means of LFs; up till now it was applied to only one database, *BADELE*, in only one language, that being Spanish. In this paper we will present the methodological problems in connection with the automatic inheritance we face right now: predicting collocations by semantic labels and rewriting the map of semantic labels.

**Keywords:** e-dictionary; lexical inheritance; Lexical Functions

## 1 Introduction: automatic inheritance and dictionaries

*DiRetEs* is a Spanish monolingual e-dictionary (Barrios 2019) based on *BADELE.3000* database (Barrios and Bernardos 2007). This database includes not only collocations but also paradigmatic lexical relations, such as synonyms, ontological relations, such as *part of*, and speech acts, such as formulaic routines. Theoretical guidelines of the project are based on the Meaning-Text Theory principles (Mel'čuk 1996, 2014; Polguère 1997, 2014; Apresjan 2000) and inspired by some lexicographers works (Atkins and Rundell 2008; Fontenelle 2008; Hanks 2009, 2013; Granger 2012), however the central point is the automatic inheritance of some lexical relations (Mel'čuk and Wanner 1996; Barrios 2010), a concept which we will attempt to explain in the following lines.

Combinatorial Dictionaries (Bosque 2004, 2006) and Explanatory Combinatorial Dictionaries (Mel'čuk et al 1995; Mel'čuk and Polguère 2008), are dictionaries specializing in lexical co-occurrence of words. Lexical inheritance was a very promising proposal of Mel'čuk and Wanner (1996) for these kinds of dictionaries: a way to describe productive lexical relation in a particular lexical field. The authors claim that most of the words sharing a hyperonym (such as 'emotion') could be present in similar collocations (*to feel joy, sadness, envy, etc.*). There is a Lexical Function (LF) attached to these collocations, which means that there is a formal way to express this productive lexical relation: in the Meaning-Text Theory (MTT) these collocations are attached to Oper, a LF useful for light verbs, such as *feel, do, make* or *have*. The Mel'čuk and Wanner proposal saw a complementary approach in Reuther (1996) and a lengthy explanation in Milićević (1997). All of them offered a powerful theoretical approach, however they do not apply it in a large dictionary. Domain inheritance was proposed by Barrios (2009, 2010) as a way to predict productive lexical relations attached to some LFs. It can be understood as a second step in the Mel'čuk and Wanner way. The analysis of LFs domains (which means, the set of words for which this LF was created) is useful to predict collocations: for instance, we can predict that many abstract nouns are related to some light-verbal collocations. If we create the list of entries of our dictionary and first of all classify them by lexical fields, we can see that, as in English and German, the Spanish verb *sentir* (to feel) is present in the emotions field. But we can also predict that *dar* (to give) is productive in the affection field (*dar un beso*, give a kiss; *dar un abrazo*, give a huge) and hit field (*dar una patada*, to kick; *dar un puñetazo*, to punch; *dar un codazo*, to nudge); and *tener* (to have) is present in the disease field (*tener fiebre*, to have a fever; *tener diabetes*, to have diabetes; *tener cáncer*, to have cancer), etc. Therefore the preliminary question is, not only how many lexical fields are predictable for the verb *to feel* and the LF Oper, but primarily how many other lexical fields are predictable in general for light verbs, and consequently, for the LF Oper. Similarly, as we will see, we can reflect on the other LFs which are not related to -light verbs collocations.

Automatic inheritance was proposed as a methodological task based on both principles by Barrios (2010) and implemented when designing *BADELE.3000*. Actually the design of this database was produced to prove that both the principle of lexical inheritance and the principle of domain of LFs were valid. It contained 3,300 nouns (the most frequent in the Spanish spoken in Spain) and 20,700 collocations formalized by means of LFs: around 9,000 lexical relations were obtained automatically and 11,700 were added manually (Barrios, 2010). Up till now the database has grown presently contains 19,845 words and multi-words and 101,988 lexical and semantic relations described by means of Lexical Functions; approximately a third of them, 32,948, were automatically obtained.

At present we are working on a new dictionary based on *BADELE.3000*, called *Diretes*. As far as we know, both



*BADELE.3000* and *Diretes* are the only dictionaries that implement the automatic inheritance, there is no other dictionary in any other language in the world that does this. There are some other dictionaries and tools developed in the MTT with LFs, such as *Le Réseau Lexicale du Française*<sup>1</sup> and *Dicoùbe*,<sup>2</sup> the specialized dictionaries *DicoInfo*, *DicoEnviro* and *JuriDico*,<sup>3</sup> the English and Russian *ETAP-4* dictionaries<sup>4</sup> and the Spanish dictionary of emotions *DICE*.<sup>5</sup> Most of these dictionaries were developed following the MTT proposals, and some of them bring some new perspective to the theoretical model: *Le Réseau Lexicale du Française* is a hand-crafted net where all kinds of lexical relations are presented in a cognitive way (Polguère 2014), developed by a knowledge based lexicographic editor (Nabil et al 2012); *DicoEnviro* allows the implementation of the FrameNet methodology (L'Homme 2016), and *ETAP-4* applies some Moscow School techniques to the automatic translation (Apresjan et al 2002). However, as mentioned before, none of them implement the automatic inheritance.

From our point of view, the automatic inheritance allows us not only to develop both manually and automatically an e-dictionary but also to verify the validity of both principles. From a practical perspective, implementing both principles allows the lexicographers to save time, to the point that while in 2010 *BADELE.3000* contained 20,700 collocations, *Dicoùbe* (the French dictionary coetaneous) contained around 24,000 collocations: considering that the Spanish database was developed for one only researcher in one year and the French version was developed for a team during some years, we conclude that the automatic inheritance is convenient for any lexicographic task related to collocations and lexical relations. The only condition necessary is that the database needs to be designed with the ability to produce the inheritance automatically (Barrios and Bernardos 2007).

In this paper we will present two methodological problems that we face in connection with the automatic inheritance: a) the necessity of an accurate prediction by hyperons; and b) the necessity of rewriting a map of semantic labels on which the dictionary is based. We will focus on one theoretical question associated with both problems: the concept of the semantic label with which we work. The *Diretes*'s project is on course and the first phase is scheduled to be concluded in one more year. Each phase is distributed around different lexical fields. In this phase we are working on the fields of food, clothes and professions. Lack of space does not allow us to explain in detail the tables of the database, however we will present three of them: the table of Lexical-Semantic Relations, the table of Semantic Predictions and the table of Semantic Labels.

This paper is organized as follows: Section 2 presents how we implement automatic inheritance in *Diretes*; Section 3 is consecrated to the problems we face right now: predicting collocations (3.1), predicting inheritance within the table of predictions (3.2), and the maps of semantic labels and its revision (3.3); Section 4 shows some results; and finally Section 5 summarizes our conclusions.

## 2 Implementing the automatic inheritance in *Diretes*

Automatic inheritance is present in a high proportion in *Diretes*. As previously mentioned, presently we have 32,948 lexical relations not only automatically obtained but also automatically formalized. We must emphasize that this inheritance is possible for collocations, not for paradigmatic or ontological relations. In order to visualize how the automatic inheritance in the dictionary is applied, we will present an extract from the table of the Lexical Relations.

Figure 1 shows several cases of different words associated to *pan* (bread). The first column of the table of Lexical Relations shows an internal registration number attached to each collocation of the dictionary.

The three following columns contain the most significant information. Indeed, the second one shows the LF associated with each one of the collocations. As we have explained before, the LF is a function proposed by the MTT for productive lexical and semantic relations; for instance, the second row below, underlined in red, contains the LF *CausFunc<sub>0</sub>*, which means “to cause something to exist”. The third column contains the lemmas and its grammatical information: respectively, word class and morphological features, for instance *pan*, *s.*, *masc.*, *sg.* (bread, noun, masc. sg.). It also shows the semantic label, which is a sort of hyperon, such as *producto para comer* (product to eat). The fourth column shows the value of the Lexical or Semantic relation: for the second row below, the relation between the LF *CausFunc<sub>0</sub>* and *pan* (bread) is expressed by *hacer* (to make), which is a verb automatically inherited. Then, adding the values of these columns we deduce that there is a collocation meaning ‘to cause the bread to exist’, that is *hacer pan* (to make bread).

The last three columns provide additional information and they are quite useful in the final process of each phase of the project, which is the revision process (it is necessary to revise each formalization for each collocation or semantic relation). The fifth column is related to the lexical automatic inheritance: *si* (yes) means that the collocation was automatically inherited; *no* means that it was manually added. The information in this column is obtained automatically. The said example underlined in red, *hacer pan* (to make bread), was automatically obtained, and also *elaborar* (to elaborate) and *cocinar pan* (cook bread), underlined in green. However, the collocation with the verb *cocer* (bake), underlined in blue, was manually added, as shown by the word *no*.

The sixth column offers the possibility of rejecting the inheritance (if some default or mistake is detected). Let us imagine that the first expression underlined in green, *cocinar pan* (to cook bread) sounds quite unfamiliar for the reviewer of the dictionary and the fifth column shows that it was automatically inherited (see the results underlined in yellow). What should be the next step? This person must check the collocation in the corpus we use: the dictionaries, some corpus of

<sup>1</sup> <https://lexical-systems.atilf.fr/spiderlex/>

<sup>2</sup> <http://olst.ling.umontreal.ca/dicouebe/index.php>

<sup>3</sup> [http://olst.ling.umontreal.ca/?page\\_id=335](http://olst.ling.umontreal.ca/?page_id=335)

<sup>4</sup> <http://cl.iitp.ru/>

<sup>5</sup> <http://www.dicesp.com/paginas>



*Sketch Engine* and other search engines. If the corpora prove that it is a frequent collocation, there is nothing to change as all the values of the column by default are marked with *no*, which means “not reject”.

For this concrete case, the data prove that *cocinar pan* is a frequent collocation, however if the reviewer happens to find any mistake, then he should select the *sí* option. All of the *sí* results will be omitted in the web page of the dictionary, however are present in the database. At the end of each phase of the project, we will analyse the set of errors detected in the automatic inheritance: it constitutes rich information for the research on cognitive knowledge of the lexicon. Indeed, it shows how the language distinguishes features of objects or concepts that we do not distinguish consciously (Bosque 2004; Barrios 2010): one example is provided by the false inherited collocation *#ponerse un bolso* (#to put a handbag on), which sounds quite odd in Spanish and English, versus *ponerse una mochila* (to put a backpack on) (more details and a possible explanation in Barrios 2013).

Id-FA	Id-Argumento	Id-Valor	Here	Rec	ELE
!685 A0Degrada	pan (s. m. sg.) 1 - Producto para comer	enmohecido (adj. c. c.) 1 - Sin a	No	No	B
!684 AntiBon	pan (s. m. sg.) 1 - Producto para comer	correoso (adj. c. c.) 1 - Sin asig	No	No	B
!758 AntiBon	pan de Calatrava (p. c. -) 1 - Postre	empalagoso (adj. c. c.) 1 - Sin a	Sí	No	S
!874 AntiBonFinFact0	panificadora (s. f. sg.) 1 - Pequeño electrodoméstico	averiar (v. -) 1 - Sin asignar	Sí	No	S
!872 AntiBonFinFact0	panificadora (s. f. sg.) 1 - Pequeño electrodoméstico	estropear (v. -) 1 - Sin asignar	Sí	No	S
!817 AntiBonFinFact0	panificadora (s. f. sg.) 1 - Pequeño electrodoméstico	romper (v. -) 1 - Sin asignar	No	No	B
!126 AntiMagn	panadería (s. f. sg.) 1 - Local comercial	pequeño (adj. c. c.) 1 - Rasgo fi	No	No	A
!135 AntiMagn-temp	panadería (s. f. sg.) 1 - Local comercial	moderno (adj. c. c.) 1 - Sin asig	No	No	A
!149 AntiReal1	panadería (s. f. sg.) 1 - Local comercial	incendiar (v. -) 1 - Sin asignar	No	No	C
!087 AntiReal2	panadería (s. f. sg.) 1 - Local comercial	robar (v. -) 1 - Sin asignar	Sí	No	S
!689 Bon	pan (s. m. sg.) 1 - Producto para comer	crujiente (adj. c. c.) 1 - Sin asig	No	No	S
!764 Bon	pan de Calatrava (p. c. -) 1 - Postre	delicioso (adj. c. c.) 1 - Sin asig	Sí	No	S
!769 Bon	pan de Calatrava (p. c. -) 1 - Postre	exquisito (adj. c. c.) 1 - Sin asig	Sí	No	S
!762 Bon	pan de Calatrava (p. c. -) 1 - Postre	sabroso (adj. c. c.) 1 - Sin asig	Sí	No	S
!139 CausAntiBonFact0	panadería (s. f. sg.) 1 - Local comercial	robar (v. -) 1 - Sin asignar	Sí	No	S
!869 CausDenuovoFact0	panificadora (s. f. sg.) 1 - Pequeño electrodoméstico	arreglar (v. -) 1 - Acción	Sí	No	S
!818 CausDenuovoFact0	panificadora (s. f. sg.) 1 - Pequeño electrodoméstico	reparar (v. -) 1 - Sin asignar	No	No	B
!429 CausFact0	panadería (s. f. sg.) 1 - Local comercial	abrir (v. -) 1 - Acción	Sí	No	S
!483 CausFact1	filete empanado (s. m. sg.) 1 - Carne	poner (v. -) 1 - Sin asignar	Sí	No	S
!952 CausFunc0	pan (s. m. sg.) 1 - Producto para comer	amasar (v. -) 1 - Acción	No	No	S
!826 CausFunc0	empanada (s. f. sg.) 1 - Sin asignar	cocer (v. -) 1 - Sin asignar	No	No	B
!953 CausFunc0	pan (s. m. sg.) 1 - Producto para comer	cocer (v. -) 1 - Sin asignar	No	No	S
!820 CausFunc0	empanada (s. f. sg.) 1 - Sin asignar	cocinar (v. -) 1 - Sin asignar	No	No	B
!954 CausFunc0	pan (s. m. sg.) 1 - Producto para comer	cocinar (v. -) 1 - Sin asignar	Sí	No	S
!373 CausFunc0	panadería (s. f. sg.) 1 - Local comercial	construir (v. -) 1 - Sin asignar	Sí	No	S
!821 CausFunc0	empanada (s. f. sg.) 1 - Sin asignar	elaborar (v. -) 1 - Sin asignar	No	No	B
!955 CausFunc0	pan (s. m. sg.) 1 - Producto para comer	elaborar (v. -) 1 - Sin asignar	Sí	No	S
!760 CausFunc0	pan de Calatrava (p. c. -) 1 - Postre	elaborar (v. -) 1 - Sin asignar	Sí	No	S
!817 CausFunc0	empanada (s. f. sg.) 1 - Sin asignar	hacer (v. -) 1 - Sin asignar	No	No	A
!956 CausFunc0	pan (s. m. sg.) 1 - Producto para comer	hacer (v. -) 1 - Sin asignar	Sí	No	S
!768 CausFunc0	pan de Calatrava (p. c. -) 1 - Postre	hacer (v. -) 1 - Sin asignar	Sí	No	S

Figure 1: Extract from the table of Lexical-Semantic Relations with some lexical relations associated to *pan* (bread)

The last column is manually added and shows the level of Spanish that is appropriate for a student to teach this lexical or semantic relation: A, B and C levels follow the *European Framework of Reference for Languages*. We have added three additional levels in this last column: E (that means for experts), which is the level adequate for terminology; V (that means vocabulary), a level for unfamiliar words for many native speakers which constitutes rich vocabulary present in literature and some books; and S, which means “*sin asignar*” (not assigned), which is a temporal mark (automatically present by default) prior to the selection of the level.

### 3 Problems arising when implementing automatic inheritance

#### 3.1 Predicting collocations

The first problem that arises when working with automatic inheritance is that all the predictions should be applied automatically before working manually: for instance, the value of CausFunc<sub>0</sub> for ‘prepared food’ is *to make*, for ‘music’ is *to compose*; and for ‘literature’ is *to write*. That means that all the nouns that could be labelled as ‘prepared food’ would combine with *to make*, such as *bread, salad, paella* or *soup*. Similarly *to compose* combines with *symphony, song*, and *melody*; and *to write* with *novel, poem* and *essay*. As we will see in section 3.3, *Diretes* has the same map of labels as *BADELE.3000*. In order to save time and effort all the predictions (the relation between CausFunc<sub>0</sub> and ‘prepare food’/to make, etc.) and the inheritances should be done before starting with the manual addition of some other collocations. These predictions are a result of the introspection of the lexicographer via whom data can at times be found in the combinatorial dictionaries (Bosque 2004, 2006). Unfortunately, not all of these predictions are necessarily accessible for all the components of the team working on the dictionary: some predictions demand experience and a strong knowledge



of the MTT model. We will extrapolate on this point.

There are two possible ways to work with the predictions, each one of them aligned with a different difficulty level. The easiest way to predict productive relations is quite similar to the methodology applied by Mel'čuk and Wanner (1996): suppose the team is working with words such as *potato*, *tomato* and *cucumber*, and they observe that *to plant* is a productive verb for these nouns. As we can predict this collocation for all the set of vegetables, we write *to plant* in the table of predictions and we describe the appropriated LF (CausFunc<sub>0</sub>, because, as we stated before, it means "to cause something to exist"). After the prediction, we apply the automatic inheritance and we obtain the fifty corresponding collocations formalized by means of this LF. We have avoided writing them manually. This methodology is useful for non MTT-experts and applicable to different lexical fields and different LFs. The second methodology, however, is more difficult. Applying the prediction by domains of LFs (Barrios 2010) as a first step, as we have said previously, demands not only a strong knowledge of the MTT model but also the ability to go from abstraction (the meaning of some LFs) to the lexicon (the potential domain for each one of these LFs). We will attempt to explain the process.

Firstly we should think about each one of the LFs and their potential meaning. Consider the case of CausFunc<sub>0</sub> and its meaning, 'to create'. We should calculate how many lexical fields could be the potential domain for this LF. In order to reach the answer we connect the extra-linguistic knowledge with the linguistic knowledge, and we conclude that if we can create objects, tools, food, leisure products, etc., at least one verb necessarily exists that expresses the meaning 'to create' for all the words naming these realities. In Spanish we describe 164 predictions for this LF, as Figure 2 shows; the relation between the second and fourth columns could be literally translated into English as *fruit/cultivate*, *animal cabin/build*, *building/raise*, *rule/dictate*, *theoretical principle/discover*, *energy/produce*, etc. The case of *pescado*, *capturar/pescar* (fish, catch) is underlined in red as in Spanish we can accept this is a particular case of 'creation': the word for the animal *pez* (fish) differs from the word for the food *pescado* (fish), similar to the English words *pig/pork*. That is the reason we associate CausFunc<sub>0</sub> to *pescar un pez* (to catch a fish), as it means "to create the food fish".

For some other lexical relations the LF CausFunc<sub>0</sub> is not adequate, such as the cases of *consensuar/ negociar una norma jurídica* (to agree on a legal rule) or *trazar una obra pública* (to plot a public work). When some people agree on a legal rule, these people do not create a new rule but the conditions by which this rule can be dictated. A similar situation involves the action of plotting a public work. We cannot use CausFunc<sub>0</sub> however, is there any other way to formalize these lexical relations? When there is no LF adequate for any productive relations, in the MTT model it is possible to provide a new way to formalize them: if any researcher should discover a new productive relation that could be understood as a LF, he can propose a new LF which will be classified as a non-standard LF. There is a non-standard LF fairly close to CausFunc<sub>0</sub>, called *EssayerCausFunc<sub>0</sub>*, which means "to try to cause something to exist" (Essayer was proposed by Polguère 2007). We translate the French verb *essayer* to the equivalent Spanish one, *intentar*; consequently, as the examples underlined in blue (Figure 2) show, we work with *IntentarCausFunc<sub>0</sub>*.

All the examples underlined in red and blue in the Figure 2 exemplify how the lexicographer should have not only a high level of MTT knowledge in order to predict the domain of standard LFs, but also familiarity with non-standard LFs:

Relaciones semánticas		Predicciones semánticas			
Id-PS	Id-ES	LF	Id-FA	Id-A	Heredado
628 Norma jurídica		CausFunc0	promulgar (v. -)	1 - Sin asignar	No
649 Norma jurídica		IntentarCausFunc0	negociar (v. -)	1 - Sin asignar	No
648 Norma jurídica		IntentarCausFunc0	consensuar (v. -)	1 - Sin asignar	No
627 Norma jurídica		CausFunc0	emitir (v. -)	1 - Sin asignar	No
625 Norma jurídica		CausFunc0	aprobar (v. -)	1 - Acción	No
677 Obra Pública		CausFunc0	construir (v. -)	1 - Sin asignar	No
686 Obra Pública		IntentarCausFunc0	trazar (v. -)	1 - Sin asignar	No
2614 Pequeño electrodoméstico		CausFunc0	montar (v. -)	1 - Sin asignar	Sí
717 Percepción		CausFunc0	causar (v. -)	1 - Sin asignar	Sí
718 Percepción		CausFunc0	desprender (v. -)	1 - Sin asignar	No
723 Pescado		CausFunc0	pescar (v. -)	1 - Sin asignar	No
722 Pescado		CausFunc0	capturar (v. -)	1 - Sin asignar	No
745 Pescado azul		CausFunc0	pescar (v. -)	1 - Sin asignar	Sí
744 Pescado azul		CausFunc0	capturar (v. -)	1 - Sin asignar	Sí
767 Pescado blanco		CausFunc0	pescar (v. -)	1 - Sin asignar	Sí
766 Pescado blanco		CausFunc0	capturar (v. -)	1 - Sin asignar	Sí
1278 Pieza de bisutería o joyería		CausFunc0	diseñar (v. -)	1 - Sin asignar	No
24004 Postre		CausFunc0	preparar (v. -)	1 - Sin asignar	Sí
23194 Postre		CausFunc0	elaborar (v. -)	1 - Sin asignar	Sí
24003 Postre		CausFunc0	hacer (v. -)	1 - Sin asignar	Sí
831 Proceso		CausFunc0	causar (v. -)	1 - Sin asignar	Sí
835 Proceso humano		CausFunc0	causar (v. -)	1 - Sin asignar	Sí
838 Proceso médico		CausFunc0	causar (v. -)	1 - Sin asignar	Sí
1292 Producto cinematográfico		CausFunc0	dirigir (v. -)	1 - Sin asignar	No
1294 Producto cinematográfico		CausFunc0	producir (v. -)	1 - Sin asignar	No
1293 Producto cinematográfico		CausFunc0	hacer (v. -)	1 - Sin asignar	No
851 Producto energético		CausFunc0	producir (v. -)	1 - Sin asignar	No
1241 Producto para comer		CausFunc0	cocinar (v. -)	1 - Sin asignar	No
1243 Producto para comer		CausFunc0	hacer (v. -)	1 - Sin asignar	No
1244 Producto para comer		CausFunc0	preparar (v. -)	1 - Sin asignar	No
1242 Producto para comer		CausFunc0	elaborar (v. -)	1 - Sin asignar	No

Figure 2: Extract from the table Semantic Predictions, some data prior to the inheritance of CausFunc<sub>0</sub> ('to create')



The last rows of Figure 2 (underlined in green) show the collocations predicted for CausFunc<sub>0</sub> and *producto para comer* (product to eat): *cocinar*, *hacer*, *preparar* and *elaborar el pan* (to cook, to make, to prepare and to produce the bread). Now we can return to Figure 1 and check that *bread* has been labelled as *producto para comer* (product to eat), so the collocations underlined in red and green in Figure 1, *hacer*, *elaborar* and *cocinar el pan* (to make, to produce and to cook the bread) are also present in Figure 2, because they were automatically obtained from the Table of Semantic Predictions; while *cocer* (bake) (underlined in blue in the Figure 1) is not, because it was manually added directly to the Table of Lexical-Semantic Relations.

In *Diretes*, only for the LF CausFunc<sub>0</sub>, we have 90 different semantic labels and 156 inheritable collocations (see the number underlined in pink, Figure 2), all of them predicted by the lexicographer's introspection. After the inheritance, we obtained 2,447 collocations related to CausFunc<sub>0</sub>. If we consider that the total number of collocations for this LF is 4,901, we observe that almost 50 percent was automatically obtained. Once again the data show that the automatic inheritance saves time and effort. However, on the other hand, this small example proves that any project applying the automatic inheritance demands lexicographers with a strong knowledge of the MTT model and of the Lexicology and Semantics of the natural language object of the dictionary. We will comment on the examples underlined in yellow in the next section.

### 3.2 Predicting the automatic inheritance within the table of Semantic Predictions

There are some Semantic Predictions that can be inherited within the table of Semantic Prediction, as the last column in Figure 2 shows. That column contains two examples underlined in orange with the value *no*, and four examples underlined in yellow with the value *yes*. If we look at the preceding columns, we see that the value *no* is attached to *pescado* (fish, in the second column underlined in red) and the value *yes* is attached to *pescado azul* (blue fish) and *pescado blanco* (white fish), underlined in yellow. That means that the verbs *capturar*, *pescar* (fish/catch) were added manually for 'fish', and automatically inherited by 'white fish' and 'blue fish' within the table of semantic predictions. Thus we can produce not only inherited collocations but also inherited predictions.

At this point, we should say that we attempt to collect mostly linguistic information and that we also attempt to differentiate between linguistic items and ontological information. Subsequently, the question that arises is: is the relation between *fish* and *blue fish* linguistic or extra-linguistic? Is it any piece of information of real life or is it an expression we should work with?

In the MTT the relation between concepts such as 'fish' and 'blue fish' is close to the LF Gener, which means 'generic concept'. This LF is conceptually close to a hyperonym but it is not a hyperonym, because it does not form explicit semantic relations (such as the hypernym does) but a lexical relation, such as the one between *republic* and *state* (we can say *republican state*), or *liquid* and *substance* (we say *liquid substance*), or *process* and *regeneration* (we say *process of regeneration*) (examples taken from Mel'čuk 2015: 194). Then, in order to know if a word such as *fish* and its relation with *salmon* is a candidate for Gener, we attempt to build an expression for both words, such as *the fish salmon*. As it does work, we could formalize this relation such as (1) shows:

(1) Gener (salmon) = fish

As the word *blue* is a predicate that combines with *fish* it cannot be a value of Gener. As far as we know, within the MTT model, the LF Gener is only explored at the lemma level, that means describing words such as *salmon*, *sardine*, *hake* or *sea bass*, and its relation with *fish*, as the French Dictionary *Le réseau lexicale (LRL)* does (see the French entries *saumon*, *sardine*, *carpe*, *loup de mer*, etc., which contain formalizations like the one proposed in (1).

In the next Section we will analyse with more detail this LF and its relation with the concept of a semantic label. We will also attempt to answer the mentioned question, is the distinction between blue fish and white fish linguistic or extra-linguistic?

In *Diretes* we do not work with the LF Gener but with a concept close to this LF called *semantic label* (Milicevik 1997; Polguère 2003, 2011). A semantic label is a descriptive tool, equivalent to a hyperon and to the genus in the Aristotelic terminology. Milicevik points out that semantic labels are useful and well known in Artificial Intelligence, but there is no theoretical linguistic approach in this area except for technical applications. The semantic label of a word is usually the central meaning of this word (Milicevik 1997: 36-37), and can be taken from the definitions of good dictionaries (Polguère 2011). Milicevik (1997: 38-39) points out that there are three conditions for any semantic label: a) it takes up a central position in the meaning of the word (such as the meaning 'emotion' in *joy*); b) it reflects sufficiently enough the co-occurrence of this word (*to feel an emotion*, *to suffer an emotion*; *to enjoy an emotion*, etc.); c) it is useful to label a group of words (such as 'emotion' and *joy*, *anger*, *fear*, *envy*, etc.).

One particularity of the French *LRL* that Polguère develops is that the semantic labels are not directly present: they work with the LF Gener and with a sort of short paraphrase which expresses a central meaning of the word. Polguère (2011) explains that while WordNets works with the concept of synsets, which combines meaning and grammatical information, the hierarchy of semantic labels he proposes should be attached only to the meaning. Curiously enough, these paraphrases are expressed necessarily attached to a grammatical role, such as shown by some *LRL* examples: *admirateur* (admirer) and *admiratif* (admired) share the meaning 'who shows a feeling' (from this point we will translate the French *LRL* paraphrases into English). Compare this paraphrase with the one of *admirable*, 'which has a particular feature'; *admiration*, 'feeling', and *admirer* (to admire), 'to feel a feeling'. These examples and the other lemmas of the *LRL* demonstrate that the paraphrases expressing the semantic label have been redacted according to the syntactic function of the word described.

As we will explain in Section 3.3, we work with a slightly different concept of Milicevik's semantic label.



### 3.3 Revising the table of semantic labels

As mentioned in the previous Section, regarding the semantic labels, the French dictionary *LRL* shows some differences with *Diretes*. The *LRL* works with the relation between the word *fish*, *Gener* and *salmon* (see (1), and with paraphrases such as ‘relatif à un animal’ (related to an animal) (see the entry *saumon* in the *LRL*). *Diretes*, as Figure 2 shows (come back to the examples underlined in yellow), contains not only ‘fish’ as a semantic label, but also ‘blue fish’ and ‘white fish’. What is the reason for this? The answer is that we attempt to obtain a higher granularity in our description in order to exploit as much as possible the automatic inheritance.

Figure 3 presents an extract from our table of semantic labels. The first column corresponds to what we call *raíz* (root) (underlined in green), which is the first distinction between words attached to entities (labelled as ‘*ser*’, being) and words attached to predicates or abstract nouns (labelled as ‘*concepto*’, concept). We have a count of nine levels in our table of semantic labels (see the column underlined in pink). This table was present in *BADELE.3000* however we are adding some new semantic labels, although presently no great changes affect its structure. The first levels respond to conceptual distinctions, and the last ones contain semantic labels defined by linguistic features: we will try to explain this distinction via the example of the words naming different types of food and sweets.

From the lowest level to the highest, the following table shows how we classify different types of labels. In the original database we counted on the label ‘*dulces y postres*’ (sweets, created for *cake*, *ice-cream*, etc.) and the label ‘*platos preparados*’ (prepared dishes, for *paella*, *croquetas* etc.). Both labels are labelled as ‘*alimento preparado*’ (prepared food), which in turn is labelled as ‘*producto de consumo*’ (consumed product), which in turn is labelled as ‘*producto*’ (product) (see the examples underlined in red in Figure 3).

Raíz	Nivel 1	Nivel 2	Nivel 3	Nivel 4	Nivel 5	Nivel 6	Nivel 7	Nivel 8	Nivel 9
Ser	Ente	Producto	Producto artístico	Tipo de obra de teatro, novela o pe					
Ser	Ente	Producto	Producto artístico	Tipo de obra literaria					
Ser	Ente	Producto	Producto de consumo	Aderezo					
Ser	Ente	Producto	Producto de consumo	Alimento	Ahumados				
Ser	Ente	Producto	Producto de consumo	Alimento	Alimento vegetal	Cereal			
Ser	Ente	Producto	Producto de consumo	Alimento	Alimento vegetal	Legumbre			
Ser	Ente	Producto	Producto de consumo	Alimento	Alimento vegetal	Verdura			
Ser	Ente	Producto	Producto de consumo	Alimento	Carne				
Ser	Ente	Producto	Producto de consumo	Alimento	Embutido				
Ser	Ente	Producto	Producto de consumo	Alimento	Fiambre				
Ser	Ente	Producto	Producto de consumo	Alimento	Fruto	Baya			
Ser	Ente	Producto	Producto de consumo	Alimento	Fruto	Fruta			
Ser	Ente	Producto	Producto de consumo	Alimento	Fruto	Fruto seco			
Ser	Ente	Producto	Producto de consumo	Alimento	Marisco				
Ser	Ente	Producto	Producto de consumo	Alimento	Pasta				
Ser	Ente	Producto	Producto de consumo	Alimento	Pescado	Pescado azul			
Ser	Ente	Producto	Producto de consumo	Alimento	Pescado	Pescado blanco			
Ser	Ente	Producto	Producto de consumo	Alimento	Pulpos y calamares				
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce	Dulce cocido		
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce	Dulce congelado		
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce	Dulce frito		
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce	Dulce horneado		
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce			
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Dulces y postres	Dulce			
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Plato preparado				
Ser	Ente	Producto	Producto de consumo	Alimento preparado	Refrigerio				
Ser	Ente	Producto	Producto de consumo	Bebida	Bebida alcohólica				
Ser	Ente	Producto	Producto de consumo	Bebida	Infusión				
Ser	Ente	Producto	Producto de consumo	Producto para la alimentación					
Ser	Ente	Producto	Producto de consumo	Producto que genera adicción					
Ser	Ente	Producto	Producto de consumo	Salsa y condimentos	Condimento	Especia			
Ser	Ente	Producto	Producto de consumo	Salsa y condimentos	Condimento	Hierba			
Ser	Ente	Producto	Producto de consumo	Salsa y condimentos	Producto de aderezo				
Ser	Ente	Producto	Producto de consumo	Salsa y condimentos	Salsa				
Ser	Ente	Producto	Producto de innovación						
Ser	Ente	Producto	Producto de la actividad del homi						
Ser	Ente	Producto	Producto de limpieza						

Figure 3: Extract from the table of Semantic Labels: preliminary data that explains the inheritance

‘Producto’ (‘product’) (see the third column in Figure 3) is an ontological label, and it can be attached to any object in the world which was produced by any person. Even if we think about the verb *producir* (to produce) as a candidate for collocations with the nouns naming these realities, it does not necessarily work on the linguistic level: some nouns that could be labelled as ‘product’ combine with this verb (*producir pizza*, to produce pizza; *producir mesas*, to produce tables; *producir petróleo*, to produce petrol) and some others do not (*#producir una infusión*, to produce an infusion; *#producir una escultura*, to produce a sculpture; *#producir agua sucia*, #to produce dirty water). This apparently incoherent behaviour points out however a coherent rule: any concept (such as that the Spanish concept ‘*producir*’, to produce) can take on a different role to its equivalent word. As the Spanish verb *producir* is attached to the context of a business production, when someone makes a tea at home, even if this person is producing a tea in same way, do not use the verb *producir*. Something similar happens within the fourth column: the label ‘consume product’ can be attached to the verb *consumir* (consume), however in Spanish, not all the words that could be labelled as ‘*producto de consumo*’ (consumable product) necessarily combine with this verb.

However, within the fifth column, the label ‘*alimento preparado*’ (prepared food, underlined in red in Figure 3) was created for words such as *paella*, *croqueta*, etc. All of them combine with the verb *preparar* (to prepare). Similarly, there are different collocations for the following semantic labels.

Let us come back now to the case of ‘fish’ (underlined in blue in Figure 3). We check in our corpus and see that *salmon* combines with *graso* (fatty) (we say *el salmon es graso*, the salmon is fatty). This combination is attached to the LF Pred (which means “to be”) and we also observe that it is productive not only for *salmon* but for any blue fish. We conclude



that there is at least one collocation liable to be inherited (*graso*, fatty plus some nouns of fish), and we create a new semantic label for this set of nouns, which is ‘blue fish’. The remaining nouns of fish will be labelled as ‘white fish’ (for them the combination with *graso* (fatty) is unusual in Spanish, and consequently, it will not be inherited).

As a result, our rule is quite simple: if there is at least one collocation productive for a group of words, we create a semantic label for them. This methodology allows us: a) to implement the automatic inheritance; and b) painting a map of semantic labels (some of them, as mentioned in section 2, unknown for our linguistic conscience) which is different for any ontology as it is partially based on concepts and mostly on linguistic behaviour.

The reason for this mixed organization is that we require our database to be useful not only for dictionaries but also for terminology and for ontologies. As is well known, working with terms implies working with concepts, because terms use to be monosemic, and at this level meanings overlap with concepts. Then, from level 1 to level 3 of our table of semantic labels, we work mostly with semantic labels attached to concepts. From level 5 to 8 we find semantic labels defined mostly based on linguistic features. In level 4, presently, we find semantic labels mixed (some of them are attached to concepts, some are based on linguistic behaviour). Finally level 9 is preserved for the future work on terminology.

Many of the semantic labels we work with were present in *BADELE.3000* but, the more we work on a particular lexical field, the more detailed is the semantic description of the words described. That implies that at times we discover new semantic labels and we add them to the table of the semantic labels of the e-dictionary. That was the case of *sweets*: in Spanish we use the word *dulce* (sweet) for *pasteles* (pie), *bizcochos* (cake), *galletas* (cookies), *natillas* (custard), etc. Some of them are baked, some of them cooked, some of them fried and some others are made without heat, but *a priori* we did not distinguish them because we do not use any Spanish expression equivalent, for instance, to the English expression *baked sweet*. We could create a semantic label such as ‘dulce hecho con calor’ (‘sweet made with heat’) but this potential semantic label raises two problems: on one hand, paraphrases such as this one are less intuitive than any word or expression in Spanish; on the other hand, we cannot apply the inheritance of the three verbs, *cocer* (to cook), *hornear* (to bake) and *freír* (to fry), to all the nouns of sweets we make with heat. There is a simple solution: we can divide the nouns of sweets made with heat into three groups, each one of them for each verb. Then, from the original semantic label ‘*dulces y postres*’ (sweets) we obtain four different semantic labels: ‘*dulces horneados*’ (baked sweets, such as cake), which combine with *hornear* (bake); ‘*dulces cocidos*’ (cooked sweets, such as custard) which combines with *cocer* (to cook); ‘*dulces fritos*’ (fried sweets, such as churros), which combines with *freír* (to fry); and *dulces congelados*’ (frozen sweets, such as ice-cream) which combines with *congelar* (to freeze) and *derretir* (to melt). Note that some of our paraphrases sound quite unusual in Spanish, such as *dulces cocidos*, however they are explicative enough and useful for the inheritance.

We can conclude that our concept of semantic label presents some differences to Milićević’s concept. The semantic label we work with demands three conditions: a) it is useful for a group of words (such as ‘blue fish’ for *salmon*, *sardine*, *tuna*, and ‘baked sweet’ for *pie*, *cake*, *cookies*); b) the label is a meaning (such as ‘fish’ or ‘sweet’) or a restricted meaning (such as ‘blue fish’, or ‘cooked sweet’), which implies that it is not necessarily part of the definition (note that ‘fish’ is part of the meaning of *salmon* but ‘blue fish’ is not); c) the label should reflect at least one co-occurrence (such as *the salmon is fatty*, *to bake the cookies*, *to cook the custard*).

## 4 Results

In *Diretes*, presently, we have made 1,614 predictions, 819 were predicted by introspection and 795 were inherited from some other predictions, which means that almost half predictions were automatically obtained from some manually added predictions. A total of 233 semantic labels were involved in these predictions. All of these semantic labels were used when labelling 7,774 words and multi-words, which is the number of entries of *Diretes* labelled up until the present (12,069 words are labelled temporally as *sin asignar*, not allocated yet, most of them verbs, adjectives and adverbs). A total of 101,988 lexical and semantic relations were described by means of Lexical Functions, and 32,948 of them were automatically obtained and formalized.

After applying automatic inheritance, we manually add the rest of the lexical-semantic relations. Figure 4 shows some of the 139 lexical relations we formalized around the word *pan* (bread). We use mostly standard LFs, some of them adjectival (see the examples underlined in blue). The first one is AntiBon, which means “bad”, and is applied to relations such as *pan sobado/resobado* (rubbing bread), *pan de ayer* (lit. bread from yesterday, which is a not fresh bread in our culture). A second adjectival LF is A<sub>0</sub>Degrad, which means “damaged”, and is applied to relations such as *pan correoso* (lit. flexible bread), *pan enmohecido* (moldy bread), *pan seco* (old dry bread), *pan duro* (hard bread).

There are some verbal LFs (underlined in red), such as CausFunc<sub>0</sub>, which means “create”, applied to *elaborar pan* (to produce the bread), *hacer pan* (to make the bread), etc. A second verbal LF is Degrad, which means “degenerate”, applied to *fermentarse el pan* (to ferment the bread) and to *revenirse el pan* (to go off the bread). The third one is IncepReal<sub>1</sub> that means “to start doing what is expected to be done with this object”: *probar el pan* (to try the bread), *catar el pan* (to taste the bread).

The set of LFs underlined in green, however, does not correspond to standard LFs. As mentioned in the first section, the set of standard LFs is a powerful tool for the description of lexical-semantic relations, however this set is not complete. We can create non-standard LFs if it is necessary; consider that these LFs can be empirically found (Mel’cuk 1996: 45), and that they are a sort of candidate for new standard LFs.

In our dictionary there is a problem relating to the richness of semantic and ontological relations and the lack of standard LFs: the actual set of standard LFs does not allow reflection on the relation between words such as *pan* (bread) and *panaderia* (bakery), *empanada* (patty), *panificadora* (bread maker), *empanar* (to bread), *panadero* (baker), *barra de pan* (baguette), etc. It is necessary therefore to create non-standard Lexical Functions for them.



In order to propose a new non-standard LF we should consider two particularities of the standard LFs: there is a significant diversity between the values of any standard LF and there is necessarily a large number of cases for any standard LF (Polguère 2007: 52-53). The author claims that there are LFs which have been proven to satisfy the preliminary conditions, called breadth and diversity, respectively, but only for one language. This set of LFs are then called *local standard LFs*, and we should write them in the local language in which they exist. Polguère summarizes some proposals of Èrastov in 1968 based on the lexicographic task, which saw the LFs Cap, Culm and Prox recognized as standard LFs and added to the MTT model. Polguère (2007) proposes De nouveau as a new non-Standard LF meaning “again”, and claims that we need to develop dictionaries in many other languages before proposing it as a universal and standard LF.

In *Diretes* we work with some non-standard LFs that are candidates to be labelled as local standard LFs. Figure 4 shows some of them, ARTIFEX, FACERE CUM, FACTUS CUM and LOCAL (underlined in green):

Relaciones Semánticas									
Id-RS	Id-FA	Id-Argumento	-Y	Id-Valor	Here	Rec	ELE		
369684	A0Degrad	pan (s. m. sg.) 1 - Producto para comer		correoso (adj. c. c.) 1 - Sin asignar	No	No	B		
369685	A0Degrad	pan (s. m. sg.) 1 - Producto para comer		enmohecido (adj. c. c.) 1 - Sin asignar	No	No	B		
369683	A0Degrad	pan (s. m. sg.) 1 - Producto para comer		seco (adj. c. c.) 1 - Propiedad física	No	No	A		
369682	A0Degrad	pan (s. m. sg.) 1 - Producto para comer		duro (adj. c. c.) 1 - Rasgo físico	No	No	A		
369686	AntiBon	pan (s. m. sg.) 1 - Producto para comer		sobado (adj. c. c.) 1 - Sin asignar	No	No	B		
369687	AntiBon	pan (s. m. sg.) 1 - Producto para comer		resobado (adj. c. c.) 1 - Sin asignar	No	No	B		
369792	AntiBon	pan (s. m. sg.) 1 - Producto para comer		de ayer (loc. adv. -) 1 - Sin asignar	No	No	A		
532567	ARTIFEX	pan (s. m. sg.) 1 - Producto para comer		panadero (s. m. sg.) 1 - Tendero	No	No	S		
369689	Bon	pan (s. m. sg.) 1 - Producto para comer		crujiente (adj. c. c.) 1 - Sin asignar	No	No	S		
369789	Bon	pan (s. m. sg.) 1 - Producto para comer		reciente (adj. c. c.) 1 - Sin asignar	No	No	B		
18956	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		hacer (v. -) 1 - Sin asignar	Sí	No	S		
18957	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		hornear (v. -) 1 - Sin asignar	No	No	S		
18955	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		elaborar (v. -) 1 - Sin asignar	Sí	No	S		
18954	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		cocinar (v. -) 1 - Sin asignar	Sí	No	S		
18953	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		cocer (v. -) 1 - Sin asignar	No	No	S		
18952	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		amasar (v. -) 1 - Acción	No	No	S		
18958	CausFunc0	pan (s. m. sg.) 1 - Producto para comer		preparar (v. -) 1 - Sin asignar	Sí	No	S		
21288	Degrad	pan (s. m. sg.) 1 - Producto para comer		fermentar (v. -) 1 - Sin asignar	No	No	S		
21289	Degrad	pan (s. m. sg.) 1 - Producto para comer		revenir (v. -) 1 - Sin asignar	No	No	S		
532557	FACERE CUM	pan rallado (p. c. m. sg.) 1 - Alimento		empanar (v. -) 1 - Acción	No	No	S		
22270	Fact1	pan (s. m. sg.) 1 - Producto para comer		gustar (v. -) 1 - Sensación	Sí	No	S		
15911	Fact1	pan (s. m. sg.) 1 - Producto para comer		alimentar (v. -) 1 - Acción	No	No	S		
532562	FACTUS CUM	pan (s. m. sg.) 1 - Producto para comer		tosta (s. f. sg.) 1 - Sin asignar	No	No	S		
532561	FACTUS CUM	pan (s. m. sg.) 1 - Producto para comer		sándwich (s. f. sg.) 1 - Refrigerio	No	No	S		
532560	FACTUS CUM	pan (s. m. sg.) 1 - Producto para comer		medianoche (s. f. sg.) 1 - Sin asignar	No	No	S		
532559	FACTUS CUM	pan (s. m. sg.) 1 - Producto para comer		tostada (s. f. sg.) 1 - Sin asignar	No	No	S		
532556	FACTUS CUM	pan (s. m. sg.) 1 - Producto para comer		bocadillo (s. m. sg.) 1 - Refrigerio	No	No	S		
27052	IncepReal1	pan (s. m. sg.) 1 - Producto para comer		probar (v. -) 1 - Sin asignar	Sí	No	S		
27051	IncepReal1	pan (s. m. sg.) 1 - Producto para comer		catar (v. -) 1 - Sin asignar	Sí	No	S		
532558	LOCUS	pan (s. m. sg.) 1 - Producto para comer		panificadora (s. f. sg.) 2 - Local	No	No	S		
532554	LOCUS	pan (s. m. sg.) 1 - Producto para comer		panadería (s. f. sg.) 1 - Local comercial	No	No	S		
369709	Parte de	pan (s. m. sg.) 1 - Producto para comer		regajo (s. m. sg.) 1 - Sin asignar	No	No	C		
369701	Parte de	pan (s. m. sg.) 1 - Producto para comer		migaja (s. f. sg.) 1 - Sin asignar	No	No	C		
369702	Parte de	pan (s. m. sg.) 1 - Producto para comer		meaja (s. f. sg.) 1 - Sin asignar	No	No	C		
369703	Parte de	pan (s. m. sg.) 1 - Producto para comer		migajón (s. m. sg.) 1 - Sin asignar	No	No	C		
369705	Parte de	pan (s. m. sg.) 1 - Producto para comer		canterito (s. m. sg.) 1 - Sin asignar	No	No	C		

Figure 4: Extract from the table of Lexical-Semantic Relations: 38 from a total of 139 lexical-semantic *bread's* relations

We write these non-standard LFs in Latin and in capitals (we do not write them in Spanish) because this formal convention helps us in our daily task. Some of them are attached to productive Spanish morphological rules. That is the case of ARTIFEX: it is the name of the person who works professionally with something; see the case of *pan* (bread) and *panadero* (baker) (first example underlined in green in Figure 4).

The set of nouns of professionals and workers in Spanish is quite broad (we have 291 nouns in our database) and there are some productive suffixes attached to this field, such as *-ero* (*banquero*, banker; *barbero*, barber), *-ista* (*periodista*, journalist; *dentista*, dentist), *-or* (*conductor*, driver; *constructor*, builder), etc.

The non-standard LF ARTIFEX is useful for the nouns we labelled as ‘professionals’. Frequently we find this lexical relation between two words (*pan*, *panadero*; *periódico*, *periodista*; *barba*, *barbero*; *diente*, *dentista*) attached to a morphological link between these two words, naming one of them a professional and the other one an object this person works with.

What we call FACERE CUM and FACTUS CUM (see the second and third set of examples underlined in green) means respectively “to do with” and “made with”. The first one is useful for verbs, such as *empanar con pan rallado* (bread with breadcrumbs); the second one is useful for relations between nouns, such as *sandwich*, *bocadillo* (a type of sandwich), *tostada* (toast), *medianoche* (bread roll), etc.

Finally, what we call LOCUS (see the fourth set of examples underlined in green) is a noun place related, in this case, to *bread*: *panadería* (bakery) and *panificadora* (a sort of semi-industrial bakery where the daily bread is made, the fresh bread consumed in Spain).



## 5 Conclusions

We have summarized some of the problems we face when applying the automatic inheritance in the Spanish e-dictionary *Diretes*. The point from which we start in our methodology is the map of semantic labels we work with. It was designed originally for *BADELE.3000*, which is the data base our e-dictionary is based on, and is now growing by new semantic labels.

The semantic label we are working with is a word or an expression that: a) is useful for a group of words (such as ‘baked sweet’ for *pie, cake, cookies*); b) reflects at least one co-occurrence (such as *to bake the cookies*); c) is a meaning (such as ‘sweet’) or a restricted meaning (such as ‘baked sweet’), but it is not necessarily part of the definition.

Based on this concept of semantic labels, we are able to predict the relations that can be inherited, and automatically obtain collocations formalized by means of Lexical Functions, such as *to bake*, the LF CausFunc<sub>0</sub> (which means “create”) and all the nouns labelled as ‘baked sweet’, *cookies, cake, pie*, etc. After the inheritance of all the lexical-semantic relations predicted, we manually add the non-predicted relations (which are taken from the data of some corpora we use) and the corresponding LFs. The predictions demand a strong knowledge of the MTT model, which implies the lexicographer is expert not only in standard LFs but also in non-standard LFs.

Summarizing the novelty of this project is the implementation of the automatic inheritance (both lexical and domain) and the particular concept of semantic label on which it is based: the condition upon which a new semantic label is created is that there is at least one particular collocation that distinguishes this set of words from some other sets of words.

## 6 References

- Apresjan, J. (2000). *Systematic Lexicography*. Oxford: Oxford University Press.
- Apresjan, J., Boguslavsky, I., Iomdin, L., Tsinman L. (2002). Lexical Functions in NLP: Possible Uses. Computational Linguistics for the New Millennium: Divergence or Synergy? In M. Klenner, H. Visser (eds.). *Festschrift in Honour of Peter Hellwig on the occasion of his 60th Birthday*. Frankfurt: Peter Lang, pp. 55–72.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Barrios Rodríguez, M.A. (2009). Domain, domain features of lexical functions and generation of values by analogy according to the MTT approach. Proceedings of Fourth International Conference on Meaning-Text Theory. <http://olst.ling.umontreal.ca/pdf/ProceedingsMTT09.pdf> [30/05/2020].
- Barrios Rodríguez, M.A. (2010). *El dominio de las funciones léxicas en el marco de la Teoría Sentido-Texto. Estudios de Lingüística del español* (ELiEs), vol 30. <http://elies.rediris.es/elies30/index30.html> [30/05/2020].
- Barrios Rodríguez, M.A. (2019). ¿Aún queda alguien para quien no exista un diccionario? *Diretes*, un diccionario electrónico apto para máquinas. In M.C. Cazorla, M.A. García Aranda, & P. Nuño (Eds.), *Lo que hablan las palabras. Estudios de lexicología, lexicografía y gramática en honor de Manuel Alvar Ezquerro*. Lugo: Axac.
- Barrios, M. A., Bernardos, S. (2007). BaDELE3000: An implementation of the lexical inheritance principle. In Reuther, Tillman; Wanner, Leo. (eds.) *Wiener Slawistischer Almanach. Sonderband 69*, pp. 68-77.
- Bosque, I. (2004). *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: S. M.
- Bosque, I. (2006). *Diccionario combinatorio práctico del español contemporáneo*. Madrid: S. M.
- Hanks, Patrick (2013). *Lexical analysis. Norms and Exploitations*. Cambridge, Massachusetts: the MIT Press.
- Fontenelle, T. (2008). “Using a Bilingual Dictionary to Create Semantic Networks”. In Fontenelle (ed.), *Practical Lexicography. A reader*. Oxford: Oxford University Press, 169-189.
- Gader, N., Lux-Pogodalla, V. and A. Polguère (2012). Hand-Crafting a Lexical Network with a Knowledge-based graph editor. *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III)*, pp. 109–126.
- Granger, S. (2012). “Introduction. Electronic Lexicography: from challenge to opportunity”. In Granger, Sylviane & Paquot, Magali (eds.). *Electronic Lexicography*. Oxford: Oxford University Press, 1-11.
- Hanks, P. (2009). The impact of corpora on dictionaries. In Baker, Paul (ed.). *Contemporary Corpus Linguistics*, London: Continuum, 114-236.
- Hanks, P. (2013). *Lexical Analysis*, Cambridge: The MIT Press.
- L’Homme, M. C. (2016). Terminologie de l’environnement et Semantique des cadres. *Congrès Mondial de Linguistique Française - CMLF 2016*, SHS Web of Conferences 2 05010 (2016)
- Mel’čuk, I. (1996). “Lexical functions: A tool for the description of lexical relations in a lexicon”. In Wanner, L. *Lexical functions in lexicography and natural language processing*. Amsterdam/ Philadelphia: John Benjamins, 209-278.
- Mel’čuk, Igor (2014): *Semantics. from meaning to text*. Vol 3. Amsterdam/Philadelphia: John Benjamins.
- Mel’čuk, I., Clas. A., Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-le-neuve: Duculot.
- Mel’čuk, I., Wanner, L. (1996). Lexical Functions and Lexical Inheritance. In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins Publishing Company, pp. 209-278.
- Mel’čuk, I., Polguère, A. (2008). *Lexique actif du français - l’apprentissage du vocabulaire fondé sur 20.000 dérivations sémantiques et collocations du français*. Bruxelles : De Boeck & Larcier.
- Milićević, J. (1997). Étiquettes sémantiques dans un dictionnaire formalisé du type Dictionnaire Explicatif et Combinatoire. Montreal : Université de Montreal.
- Polguère, A. (1997). Étiquetage sémantique des lexies dans la base de données dico. *Traitement automatique des langues*, 44(2), pp. 39–68.
- Polguère, A. (1997). Meaning-Text Semantic Networks as Formal Language. In L. Wanner (ed.). *Recent Trends in Meaning-Text Theory*. Amsterdam: John Benjamins Publishing Company, pp. 1–24.
- Polguère, A. (2007). Lexical Function Standardness. In L. Wanner, *Selected Lexical and Grammatical Issues in the*



- Meaning-Text Theory*. Amsterdam/Philadelphia: John Benjamins, pp. 43-96.
- Polguère, A. (2011). Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de Lexicologie*, 98, pp.197–211.
- Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, Vol. 27 No. 4, pp. 396–418.
- Reuther, T. (1996) On Dictionary Entries for Support Verbs: The Cases of Russian VESTI, PROVODIT' and PROIZVODIT'. In Wanner, L. (ed), *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: Benjamins, pp. 181-208.

### **Acknowledgements**

The project “Diretes: Diccionario Reticular Español. Diccionario analógico y relacional con acceso en red desde el sentido y desde la forma” has been founded by the grant FFI2017-83293 of the Spanish Ministry of Science and Economy. I would like to express my gratitude to Deborah Paton for all her suggestions on the English revision of the paper and to the anonymous reviewers for their helpful comments. Any remaining errors are our own.



# Dictionnaire des francophones - A New Paradigm in Francophone Lexicography

Dolar K.<sup>1</sup>, Steffens M.<sup>2</sup>, Gasparini N.<sup>3</sup>

<sup>1</sup> CREE, Inalco

<sup>2</sup> Department of Languages, Literature and Communication, Utrecht University

<sup>3</sup> Institut International Pour La Francophonie (2IF)

## Abstract

*Dictionnaire des francophones* (DDF) is a general francophone dictionary, the result of an institutional-collaborative project, the goal of which is to provide a new online resource. It aims to cover all varieties of the French lexicon from a descriptive point of view and to highlight the plurality of linguistic norms while endeavouring to treat different linguistic varieties equally. The paper focuses on the dictionary-making process and lexicography technologies used in the project. Some particularly innovative aspects of the DDF are discussed, such as the institutional support and the scientific background in which the project is grounded; the hybrid nature of the dictionary, combining imported resources in a relational database, enriched by a complex speaker-based collaborative input; inclusivity of linguistic variation and the modes of its representation. Taking into account these characteristics as well as some other features of the dictionary lead us to the conclusion that the DDF is a unique object in comparison to existing traditional and collaborative resources, providing a new paradigm in francophone lexicography.

**Keywords:** *Dictionnaire des francophones*; professional dictionaries; collaborative lexicography; general dictionary; francophone dictionary; linguistic variation of the French language; plurality of norms

## 1 Introduction

The lexicon, like any part of language, varies across contexts. The diatopic and diastratic varieties of French have been the subject of a long tradition of lexicographic description and analysis. In contrast to the prescriptive ideology of the second half of the 20th century (for example *Refrancisons-nous* by Frère Jean-Ferdinand in Québec in 1951, *Chasse aux belgicismes* by Hanse et al. 1971, etc.), a lot of recent research, often focusing on lexical variation, describes and values linguistic varieties (for example Rézeau 2001; Mercier & Verreault 2002; Thibault 2004 and 2008; Glessgen & Thibault 2005; Bernet & Rézeau 2010; Francard et al. 2015). Research in this field has also led to the production of quality linguistic atlases and glossaries (ALW; Dulong & Bergeron 1980). These resources, mainly focusing on lexical variation, are most often grounded in the field of contrastive lexicography, aiming to describe a well-defined subset of words used in a particular region or by a specific community of speakers. To date, there is no general dictionary of French integrating all its varieties and meeting scientific standards.

The paper focuses on the dictionary-making process and lexicography technologies used in a new online resource pursuing this aim, *Dictionnaire des francophones* (DDF). Its public launch is planned in late 2020 but might be postponed due to the health crisis. The DDF is a general francophone dictionary, the result of an institutional-collaborative project which covers all varieties of the French lexicon from a descriptive point of view. It integrates endogenous norms and highlights the plurality of linguistic norms. The DDF is a hybrid object combining several existing dictionaries and collaborative input (under a free license and open access). Three innovative aspects of the DDF will be highlighted in the present paper: its specific institutional and scientific background, its hybrid structure as well as its inclusivity with regard to linguistic variation.

## 2 The institutional and scientific context of the DDF

To define the role that the DDF wishes to play in the field of current French and francophone lexicography, it is necessary to present the institutional context of its conception and its descriptive scientific ambition in more detail.

### 2.1 Institutional background

The DDF is an ongoing lexicographic project led by the *Délégation générale à la langue française et aux langues de France* (DGLFLF), the department of the French Ministry of Culture in charge of language protection and planning, and the *Institut international pour la Francophonie* (2IF, a part of Université Jean Moulin Lyon 3). The conception and designing process was led by the authors of this paper and reviewed by an international scientific committee, chaired by Bernard Cerquiglini.

Benefiting from the financial and technical support of various francophone organizations, the DDF is mostly under an open data license, setting an example of open cultural data, as a part of the proactive policy from the French Ministry of Culture as well as of the European Plan S (<https://www.coalition-s.org/>), aiming to make the research data accessible to the public. The DDF is one of the few institutional lexicographic projects (another example is *LEO*, a multilingual dictionary with collaborative input, designed by Technical University of Munich), that consciously promote the collaborative input of speakers in the description of their own language. To date, no other lexicographic project joins all



of these aspects, which puts the DDF in a unique position.

## 2.2 A francophone dictionary with a plurality of norms

The body of research on variation in the French language across time, space and society (Gadet 2003; Völker 2009), led by lexicologists, lexicographers, dialectologists or sociolinguists, highlights the specificities of different communities of speakers. Although, considering the entire lexicon (i.e. what varies but also what is common to all French speakers) has proven to be useful for both historical linguistics (Chambon 2006; Greub 2002) and synchronic linguistics (Baronian & Martineau 2009; Courbon 2012): it seems that a rigorous definition of units is necessary both for understanding the mechanisms of language evolution over time and the description of its current functioning on phonetic, morphological, syntactic, lexical and pragmatic levels. Nevertheless, the description of the relation between varieties and common lexemes is often not well developed (Poisson 2002; Violette 2006; Guérin 2008).

The contrastive approach favoured so far was driven by a real descriptive ambition which valued the varieties of French language and enabled the collection of precious linguistic data. However, this approach itself isolates and confines a lot of French-speaking communities' uses at the margins. Regionalisms and sociolects in general dictionaries are treated the same way, marked as a deviation in regard to the provided linguistic norm. The fact that regional lexemes are integrated in the same lexical networks as "standard lexemes" is not visible. If the existing lexicographic description of regionalisms helps to exhibit the living heterogeneity of the French language (Bavoux 2008: 17), it fails to give an entire and faithful picture of the actual use to the public at large and de facto maintains what Robillard calls the "platypus syndrome" (Robillard 2008: 325, our translation): language is presented as a juxtaposition of badly assembled parts.

Thus, when speakers commonly use regional or socially marked lexemes, sometimes without any awareness of the deviation from the norm conveyed by dictionaries, the referential resources may give them a devalued or folkloric image of their own linguistic practice. In Quebec, such considerations led to the creation of *Usito*, an online reference dictionary defining the endogenous North American French standard. The recent transition of this resource to free online access is an important step to a larger dissemination of scientific descriptions of francophone varieties.

Among the existing collaborative resources, only *Wiktionnaire*, the French part of *Wiktionary*, offers the theoretical possibility of covering the entire French-speaking field, but the chosen mode of collaboration, based on consensus (new contributions can modify and overwrite the existing ones), tends to privilege the variety of the majority of the actual contributors, in this case Metropolitan French. This predilection is especially visible in the definitions: for example, if the Quebec pronunciation of *tofu* [tofy] is mentioned in *Wiktionnaire*, although without specific labelling, it is the form *soja* (and not *soya* used in Quebec) which is mentioned in the definition (Vincent 2016 and 2017). The description of diatopic and diastratic varieties in such resources might be promising, but it still lacks homogeneity, precision and reliability. We need a general dictionary of French, one that meets scientific standards and is fully grounded in the pluricentric French-speaking world (Lüdi 2012), where all regional varieties are included and equally valued. The DDF aims to fill this gap with an open crowdsourced approach.

## 3 DDF – A hybrid object

Generally speaking, the scientific resources face a major problem in trying to keep their content up to date. The process of integrating new data provided by the scientific community is long and complex and the input from speakers is most often very poor or practically non-existent. From that perspective, the data input in the DDF, as well as its structure, is particular: it networks existing resources, which can be enriched through a collaborative speaker-based input<sup>1</sup>. We will briefly present the data model, the structure and the resources of the DDF.

### 3.1 Data model of the DDF

Combining several existing resources and opening the dictionary to the speakers necessarily implies a reflection on its form. All dictionary articles have an underlying formal structure (see Atkins and Rundell 2008; Renders 2015), the choice of which conditions and constrains the conveyed information (Mazziotta 2016). The development of native electronic dictionaries (such as *Usito*) may lead to conceptualizing dictionaries differently, for example by adding graphs to simple tree structures directly imposed by the medium of print (Heiden 2004; Měchura 2016). This graph structure allows multiple and personalized access, without traditional constraints (Steinlin et al. 2004), by designing an adapted modular consultation interface and its corresponding mobile applications.

The DDF does not function as a simple resource portal, gathering and displaying search results from different separate resources, but consists of a structured database integrating and networking different sources of data and content. The first step was to define its structure in order to be able to adequately tag different imported resources and enable a collaborative input. As in many other digital dictionaries, the Ontolex Lemon model (McCrae et al. 2017), a standard model in lexicography and terminology, and its lexicographic module Lexicog, seemed to be the best choice for the creation of the RDF database (Resource Description Framework, W3C 2004a, Měchura 2016). Some changes have been applied to the model in order to obtain a more fine-grained labelling of language varieties and to point out semantic relations between lexemes. For instance, the property *Place* (location) was added to each written form and definition since one of the main objectives of the DDF is to highlight the diatopic variation of French. The model used in the DDF is open and reusable (see full model in Steffens et al. 2020).

In this RDF database, every entry is based on the written form of the lexeme related by the *LexicalSense* property to

<sup>1</sup> A similar method was used in the early years of *Wiktionnaire*, where the 8th edition of *Dictionnaire de l'Académie française* (1932-1935) and *Dictionnaire de Littré* (1883) were used as a substrate.



linked data. Each entry is attributed to an author. The DDF opted for a classic principle in data modelling: each entry in the database corresponds to a form, a meaning, an example and a set of labels related to the form by its meaning. If one form has two or more definitions, different entries can be related through the *SenseRelation* property indicating semantic relations between definitions. The same set of relations is used to organize semantic relations between different lexical items. An entry could also have no definition, only relations to other entries, as flecational forms for instance. The specificity of the linked data used for the DDF is that each information has a uniform resource identifier (URI) and all data is integrated in a large network connected with other networks that contain lexicographical or conceptual data. The structure is not hierarchical but a bit more verbose than in some other databases. Another way to access the data, beyond the public interface, is through the SPARQL request language. SPARQL is derived from SQL language and allows complex queries in the database, for example a list of words with a specific sequence of letters, a synonym, at least two examples and a geographical indication. Results of SPARQL queries could be displayed as tables or maps, another innovative aspect of the DDF.

### 3.2 Existing dictionaries included in the DDF

In the DDF database, essentially thanks to the support of *Agence universitaire de la francophonie* (AUF), these existing resources are aligned following the same data model:

- *Inventaire des particularités lexicales du français en Afrique noire* (Équipe IFA 1988) gives lexical equivalents between French spoken in Benin, Burkina Faso, Cameroon, Central African Republic, Ivory Coast, Mali, Niger, Democratic Republic of the Congo, Rwanda, Senegal, Chad and Togo on the one hand and French spoken in France on the other. The resource was digitized and tagged; the content is richly described, but sometimes outdated, and an update is necessary.
- *Dictionnaire des belgicisms* (Francard et al. 2015), as its name indicates, is a dictionary collecting the particularisms of Belgian French. It has been digitized and tagged before being imported in the DDF.
- *Base de données lexicographiques panfrancophone* (<http://www.bdlp.org/>) is an online database which gathers in one place several scientific dictionaries with lexical items used in different varieties of French around the world. Since this resource was already digitized, it only needed to be tagged.
- *Wiktionnaire* is a part of *Wiktionary*, a multilingual collaborative dictionary, hosted by Wikimedia Foundation that also hosts *Wikipedia*, placed under a free license; only the part describing French is included in the DDF. This resource is not a traditional dictionary but its quality is more than sufficient to be included in the DDF since it is rich in neologisms and quotes from various sources, though it remains insufficient in description of usages, as already mentioned above. Since it is being constantly enriched, appropriately tagged updated copies will be uploaded regularly to the DDF.

The scientific committee of the project has the mission to gather a maximum of lexicographical resources describing different diatopic varieties of the French language. Other resources will be added to the DDF database in the near future. One should note that the source of all data gathered in the DDF database is always clearly stated. The content of these resources cannot be deleted or modified: the contributors can enrich and update the data only by adding new information.

### 3.3 Collaborative input

Following the long tradition of sociolinguistic inquiries, the description of linguistic variation and of the common lexicon based on a crowdsourced approach has proven to be highly pertinent. The relevance of asking speakers to identify and map linguistic uses has been present since the beginning of the 20th century through linguistic geography studies (Swiggers 1999; Lauwers et al. 2002; Leemann et al. 2016). With the development of the Internet, lexicography has been liberalized and democratized: a computer or a telephone connected to the Internet is enough to take part in the dictionary-making process. Bilingual or monolingual language dictionaries, but also encyclopaedias (such as *Wikipedia*) and linguistic and cultural resources in general, have thus become widely accessible to the public at large who can consult them, but also contribute in different ways. Data on lexical units selected by researchers are traditionally collected by means of closed-ended questions (for instance, “What do you call a pastry containing chocolate: *chocolatine* or *pain au chocolat*?”) or by means of images, usually with a choice of set replies (see the project *Français de nos régions*, Avanzi et al. 2016). Since collaborative resources offer open unstructured fields that allow speakers to share unexpected data on lexical units chosen by themselves, new or very specific words or uses as well as less known aspects of the language, such as the variation of norms, can be documented (Which words are acceptable in which regions? Which words are criticized? By whom?). Online collaborative lexicography has been developing substantially over the past twenty years and it now appears to be essential for lexicography in general. Within the framework of this publication we adopt the following definition of *collaborative lexicography*: activity which integrates the contributions of a community and creates through the Internet a virtual space in which the contributors participate, collaborate and support each other in writing of dictionary articles and the dictionary-making process (Dolar 2017a; Cotter & Damaso 2007; Meyer & Gurevych 2012; Granger & Paquot 2012). Collaborative lexicography forms a vast and diverse field of linguistic description: the technical methods of data collection, the types of data collected as well as their representation vary greatly from one collaborative resource to another. The collaborative resources available online today are constantly changing and range from blogs and forums (such as the *Babel Project*) to more structured resources focusing on spoken French (for instance *Blazz*, *Le Dictionnaire de la Zone*, *La Parlure* or *Urbandico*) and *Wiktionnaire*, whose form and content is close to professional dictionaries. The advantage of these resources is that they include data provided directly by speakers, but their major drawback is that they do not meet the official scientific standards, since they are not organized into a homogeneous structure that would make using the resource more efficient, and they are not user-centric or designed in accordance with



modern standards.

While most of the existing collaborative resources are based on a single mode of data collection, the objective of the DDF is to offer the possibility of contributing in different ways and thus to obtain a maximum amount of relevant information, gathering content, semantic relations, examples, pronunciations (in the future maybe also audio files)<sup>2</sup> and comments. The methods and modes of contribution to the DDF are based on best practices of existing collaborative dictionaries (see Dolar 2017b and Steffens 2017). As already pointed out, in the DDF, data can only be added (additive type of contribution), but not modified or merged (aggregative type). The DDF allows several modes of contribution; one can

- add a new lexical entry via the provided contribution form (see figure 1 below), including its written form, place of use, definition and example as mandatory information (other non-mandatory information can be provided in the appropriate fields),
- add a new definition to a form already present in the database, thus creating a new lexical entry,
- add an example or other types of information to existing lexical entries – such as grammatical categories, usage labels and semantic relations.

The purpose of the contribution form with strictly defined lexicographic fields is to obtain structured lexicographic information which is directly integrated into the dictionary structure.

Figure 1: Contribution form (structured data)

Discussion on forums and validation of content are also included. These two further forms of contribution are designed as follows:

- As the DDF pays special attention to discussions and negotiations that may arise around certain formal or historical aspects and norm-related topics, it offers the possibility of posting comments about written form and pronunciation, etymology and usage, including notes on linguistic policies and norms, etc. The comments do not have a predefined form (open debate, non-structured data) and follow a forum-like flow, close to a chat or a conversation, which enables the users to share information that was not included in the contribution form. If this kind of information is already provided in the imported resources, it will also appear in the discussion forums. In this case, the first message of the conversation is the indication of source and it cannot be edited, but contributors can write other comments, replies to comments and vote for other users' contributions. A preview of the comment with the most votes is displayed on the main page of the lexical entry.
- The contributors are thus also encouraged to validate existing contributions and comments. Three options are available in the DDF: [✓ Je valide] validates existing contributions and comments and brings them higher on the list of results, [! Je signale] draws attention to a problem. Inappropriate contributions that contravene French laws (racism, incitement to violence, etc.) can be deleted by the administrators ([- À supprimer] initiates the process of deletion). The possibility of votes involves the users in the process of sorting out the contributions. Information is then not deleted but ordered according to the number of votes, ideally prioritizing the most relevant one.

Since it is assumed that the contributors are beginners without any particular lexicographic skills, they are guided through the contribution process. Contributing is divided into micro-tasks and the DDF offers pedagogical and technical support during the process. For instance, short definitions accompany linguistic terms, didactical inserts explain each type of linguistic information and advice and tips are given dynamically. Other dynamic elements will be implemented, namely a human-machine dialogue, where questions will adapt according to the given answers. This tool will help contributors to participate without introducing any technical or scientific vocabulary.

Given the structure of the dictionary and data collected from the contributors, the DDF is actually a doubly hybrid resource, combining structured data and talk/forum discussions on the one hand, and professional resources and collaborative input on the other. To ensure quality and transparency, the source of each information always remains visible and clearly stated. Contributions in the DDF are submitted to a proofreading process carried out a posteriori by peers.

#### 4 DDF – An inclusive dictionary

<sup>2</sup> The DDF aims to be a multimodal object and recorded pronunciations will be added in collaboration with the Lingua Libre project (<https://lingualibre.org>). The possibility of adding other pronunciations (geographically tagged) is being implemented.



The main interest of the DDF is its inclusiveness both in terms of nomenclature, since all varieties of French can be present and adequately represented without restriction, and in terms of accessibility to all users, including learners. Despite the growing trend among educational resource developers to integrate online tools, the great potential of collaborative lexicography and crowdsourcing initiatives is insufficiently exploited in teaching programs and methods (Sabou et al. 2012; Steffens 2016). Among other users of the DDF, French language learners are able to discover the specificities of regional varieties of French in order to achieve a better understanding of the interactions between varieties, for instance by identifying the lexical causes of communication failures (the use of regional lexicons).

The collaborative approach, based on voluntary contributors acting as witnesses of linguistic usage and feeding the database in a dynamic and continuous manner, seems to be an effective mean of covering all varieties of the language, including those which are not represented in the corpora of traditional lexicographers. However, the project faces some major challenges: ensuring the scientific quality of the collected data by accompanying the contributors through the process, making it accessible to everyone (see 3.3), but also designing a visualization mode of data that allows the variation to be represented in its integrity, without eclipsing or isolating it.

The visual representation of lexicographic data plays a key role in the accessibility and comprehension for the public at large. To optimize the readability of the data, six criteria were used: 1) intuitiveness of the interface, 2) clarity and comprehensibility of the information, 3) univocity of colour codes used in the interface, 4) simple and easy-to-access functions, 5) accuracy of the data, 6) representation of the required types of data. Furthermore, to cope with the “difficulty of combining portability and small device size with a comfortably large display” (Lew 2010: 299), the DDF interface was developed primarily for mobile phones.

#### 4.1 Geography and diatopic variation

The geographic subsets of data are clearly delimited by a colour code (see figure 2) in order to highlight regional lexical networks and to avoid confusion (Vincent 2011 and 2016). This systematic way of presenting data helps to give an accurate picture of the distribution of the described units (Are they common to speakers of all varieties of French, everywhere in the world? Is their use limited to France alone or another particular French-speaking region?). First and most visible is thus the diatopic variation: data display follows geographical location – data is specified at city level and is displayed at region/country level.

The order of display of different meanings in the result list is based on several criteria. The most salient one is the geographical adequation between the reader and the data. For example, if the user is located in Dakar, the search results that are tagged as geographically nearest will appear higher in the result list. The user can specify a city and the interface infers in which region and country it is situated since linked data give access to Geonames (<http://www.geonames.org/>), a database of locations defined with relations to each other. This is set by default but it is also possible to personalize the search criteria by indicating a preferred semantic domain, and more options of personalization are planned for the next versions of the DDF. The order of definitions is further based on votes and on semantic relations between definitions. The latter is still being implemented, but ideally in the case of a definition Y with a specified relation to X (*Y by hyponymy of X*), Y would be displayed after X.

**Terme recherché : savoir**

[ Définition ]

Belgique	Monde	Rwanda	Sénégal, Côte d'Ivoire, Burkina Faso
verbe Pouvoir — <b>Note</b> : Dans le sens « avoir la capacité de ». Il peut être à l'indicatif, dans une phrase positive.	verbe Avoir dans la mémoire.	verbe, transitif Pouvoir, être capable de.	verbe, transitif Connaitre.

**Discussion sur l'étymologie**

Du latin populaire *\*sapere*], en latin classique *sapere*, « avoir de la sagesse », avec influence de *sapiens* « sage », d'où « être perspicace », « comprendre », puis « savoir », et élimination du classique *scire* « savoir ». Très ancien français : *savoir* (Serments de Strasbourg), puis *savoir*, et enfin *savoir*. Pendant très longtemps, du moyen français jusqu'au XVIII<sup>e</sup> siècle, le mot s'écrivait *savoir* par fausse régression au latin classique *scire* (« savoir »). Il fallut attendre 1740 pour que l'Académie française enregistrât, en la troisième édition de son dictionnaire, le mot sous sa graphie actuelle.

Figure 2: Search results highlighting the diatopic variation

#### 4.2 Usage labels

As the main objective of the DDF is to represent and document all varieties of French including their different registers and uses in specific social contexts, the project is based on an inclusive and descriptive perspective, far from any



prescriptive goals. In order to describe precisely the conditions of use of lexical items, sociolinguistic labels were implemented in the DDF. The aim of this labelling is to document, rather than to legitimize, particular uses: the labels are based on facts directly observable by the contributors (Who says that? In what context?).

Each meaning of a given form can thus be linked to different diastatic and diatopic labels. The inventory of labels aims to reflect and include different lexicographical traditions (Hausmann 1989). The main and most common lexicographic labels, integrated to controlled vocabularies, are present, but some minor editorial choices have been applied (for example, replacing *populaire*, pejorative and outdated, by *très familier*). The labels also aim to be user-friendly and accessible to contributors – during the contribution process they appear as a closed-ended list, accompanied by short definitions.

Both the inclusive approach and descriptive labelling are essential for teaching French, in particular to non-French-speaking learners, for whom it is necessary to provide information on actual usage in various French-speaking varieties by giving them information on the context in which an expression could – or should – be used. The goal is not to impose a certain use but rather to reflect the diversity and thus to allow both Francophones and learners of French as a foreign or second language to interact with a wide range of examples and uses and to adapt their linguistic practice to the given circumstances.

## 5 Conclusion

The social impact of having a collaborative resource integrating different varieties of French from a holistic perspective can be seen at different levels. A complete description of different varieties of the language, of their sociolinguistic relations, of the objective norm (how one really speaks) and of the endogenous local norms (what is considered acceptable in a given linguistic community) has many benefits, not only because it averts situations of deep linguistic insecurity, but also because it preserves the Francophone linguistic heritage worldwide by describing and promoting its diversity. From a social perspective, the free, online, collaborative and dynamic DDF creates a space for people, from the very young to the elderly, to share their lexical usages and their linguistic and cultural knowledge and competences.

The DDF is a unique project, presenting many innovations in comparison to existing traditional and collaborative resources. The modular platform of the DDF offers a new model for accompanied collaborative lexicography, seeking to exploit all potentials of the Internet by making a wide range of linguistic but also cultural data accessible to everyone. The platform gives access to an up-to-date, constantly renewed image of French varieties spoken in different parts of the world. In the present paper we outlined three main innovative features of DDF: the institutional support and the scientific background in which the project is grounded (plurality of linguistic norms and equal treatment of linguistic varieties of French); the hybrid nature of the dictionary, which combines imported resources in a relational database, enriched by a complex speaker-based collaborative input; inclusivity of linguistic variation and the modes of its representation.

Other innovative aspects of the DDF should also be mentioned. The DDF has great potential as a teaching and learning resource. Since there is a real need of useful tools for teaching communication in various French-speaking regions (Steffens & Baiwir 2020), games, mobile applications and other didactical materials are in preparation in collaboration with several francophone organizations. One should also point out the user experience in regard to both, reading (ergonomics of the pages, numerous displaying options) and the contribution process (technical support and pedagogical tools). The DDF will also provide some basic sociolinguistic information about the contributors (via the log used for the proofreading process) and readers (via the metric tool Matomo). This type of data will not be fully publicly accessible for GDPR compliance, but available for the scientific community.

The position of this new object in relation to differential lexicography, online dictionaries and collaborative resources cannot yet be fully described and defined. However, due to the scientific objectives of this institutional-collaborative project (creating a general dictionary, integrating variational aspect, plurality of linguistic norms and the common lexicon), its accessibility and features, the structure of the database and its complex input, it is safe to suggest that the DDF offers a new paradigm in francophone lexicography.

## 6 References

- ALW = *Atlas linguistique de la Wallonie* (1953- ). Liège: Presses universitaires de Liège.
- Atkins, S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.
- Avanzi, M., Barbet, C., Glikman, J. & Peuvergne, J. (2016). Présentation d'une enquête pour l'étude des régionalismes du français. In *Actes du 5<sup>e</sup> Congrès mondial de linguistique française (Tours, 4-8 July 2016)*, SHS Web of Conferences, 27. Accessed at [https://www.shs-conferences.org/articles/shsconf/pdf/2016/05/shsconf\\_cmlf2016\\_03001.pdf](https://www.shs-conferences.org/articles/shsconf/pdf/2016/05/shsconf_cmlf2016_03001.pdf) [30/05/2020]
- Baronian, L. & Martineau, F. (2009). *Le français d'un continent à l'autre. Mélanges offerts à Yves Charles Morin*. Québec: Presses de l'Université de Laval.
- Bavoux, C. (2008). *Le français des dictionnaires. L'autre versant de la lexicographie française*. Bruxelles: De Boeck Supérieur.
- BDLP = *Base de données lexicographiques panfrancophone*. Accessed at <http://www.bdlp.org> [30/05/2020]
- Bernet, C. & Rézeau, P. (2010). *Dictionnaire des expressions quotidiennes - On va le dire comme ça*. Paris: Poche.
- Blazz. Accessed at <http://www.blazz.fr/> [30/05/2020]
- Chambon, J.-P. (2006). Lexicographie et philologie: réflexions sur les glossaires d'éditions de textes (français médiéval et préclassique, ancien occitan). In *Revue de linguistique romane*, 70, pp. 123-141.
- Cotter, C. & Damaso, J. (2007). Online Dictionaries as Emergent Archives of Contemporary Usage and Collaborative Codification. In *QMOPAL - Queen Mary's Occasional Papers Advancing Linguistics*. Accessed at <http://linguistics.sllf.qmul.ac.uk/linguistics/media/sllf-migration/departement-of-linguistics/09-QMOPAL-Cotter-Damas>



- [o.pdf](#) [30/05/2020]
- Courbon, B. (2012). Représenter la diversité linguistique dans un dictionnaire monolingue: de la “traduction interne” à l’intégration sémantique. In M. Heinz (ed.) *Dictionnaires et traduction*. Berlin: Frank und Timme, pp. 153-196.
- Dictionnaire de la Zone*. Accessed at <http://www.dictionnairedelazone.fr/> [30/05/2020]
- Dolar, K. (2017a). Les dictionnaires collaboratifs en tant qu’objets discursifs, linguistiques et sociaux. PhD thesis. Université Paris Nanterre, Paris, France.
- Dolar, K. (2017b). Les dictionnaires collaboratifs non institutionnels dans l’espace francophone: éléments de typologie et bilan. In *Repères – Dorif*, 14. Accessed at [http://www.dorif.it/ezine/ezine\\_articles.php?art\\_id=380](http://www.dorif.it/ezine/ezine_articles.php?art_id=380) [30/05/2020]
- Dulong, G. & Bergeron, G. (1980). *Parler populaires du Québec et de ses régions voisines: Atlas linguistique de l’Est du Canada*. Québec: OLF.
- Francard, M. (2011). L’intégration des régionalismes dans les dictionnaires de référence du français. Le cas des belgicismes. In *Français et Société*, special issue, pp. 13-25.
- Francard, M., Geron, G., Wilmet, R. & Wirth, A. (2015). *Dictionnaire des belgicismes*. Bruxelles: De Boeck.
- Gadet, F. (2003). La variation: le français dans l’espace social, régional et international. In M. Yaguello (ed.) *Le grand livre de la langue française*. Paris: Le Seuil, pp. 91-152.
- Glessgen, M.-D. & Thibault, A. (2005) (eds). *La lexicographie différentielle du français et le Dictionnaire des régionalismes de France*. Actes du Colloque en l’honneur de Pierre Rézeau, Strasbourg: Presses Universitaires de Strasbourg, pp. III-XVII.
- Granger, S. & Paquot, M. (2012) (eds.). *Electronic Lexicography*. Oxford: Oxford University Press.
- Greub, Y. (2002). Les régionalismes lexicaux du moyen français et la formation des français régionaux, d’après l’exemple d’un corpus de farces (1450-1550). PhD thesis. Université de Neuchâtel, Neuchâtel, Switzerland.
- Guérin, É. (2008). Le “français standard”: une variété située? In *Actes du Congrès Mondial de Linguistique Française - CMLF’08*. Paris: Institut de Linguistique Française, pp. 2303-2312. Accessed at <https://www.linguistiquefrancaise.org/articles/cmlf/pdf/2008/01/cmlf08250.pdf> [30/05/2020]
- Hausmann, F. J. (1989). Die Markierung im allgemeinen einsprachigen Wörterbuch: eine Übersicht. In F. J. Hausmann *et al.* (eds.) *Wörterbücher/Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/Encyclopédie internationale de lexicographie*, I. Berlin/New York: Walter de Gruyter, pp. 649-657.
- Heiden, S. (2004). Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex. In *JADT 2004: 7es Journées internationales d’Analyse statistique des Données Textuelles*, pp. 577-588.
- IFA = Équipe IFA (1983). *Inventaire des particularités lexicales du français en Afrique noire*, AUPELF. Paris: Edicef.
- La Parlure*. Accessed at <http://www.laparlure.com> [30/05/2020]
- Lauwers, P., Simoni-Aurembou, M.-R. & Swiggers, P. (2002). *Géographie linguistique et biologie du langage: autour de Jules Gilliéron*. Leuven: Peeters.
- Leemann, A., Kolly, M.-J., Purves, R., Britain, D. & Glaser, E. (2016). Crowdsourcing Language Change with Smartphone Applications. In *PLoS One*, 11(1). Accessed at <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0143060> [30/05/2020]
- Lüdi, G. (2012). French as a pluricentric language. In M. Clyne (ed.) *Pluricentric Languages. Differing Norms in Different Nation*. Berlin: De Gruyter, pp. 149-178.
- Mazziotta, N. (2016). Représenter la connaissance en linguistique. Observations sur l’édition de matériaux et sur l’analyse syntaxique. Habilitation à diriger des recherches (Habilitation of Conducting Research). Université Paris Nanterre, Paris, France.
- McCrae, J., Bosque-Gil, J., Garcia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*, pp. 587-597. Accessed at <https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf> [30/05/2020]
- Měchura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing*, pp. 97-104. Accessed at: <http://www.lexiconista.com/raslan2016.pdf> [30/05/2020]
- Mercier, L. & Verreault, C. (2002). Opposer français “standard” et français québécois pour mieux se comprendre entre francophones? Le cas du *Dictionnaire québécois français*. In *Le Français moderne*, 70(1), pp. 87-108.
- Meyer, C. & Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*, Oxford: Oxford University Press, pp. 259-291.
- Poisson, E. (2002). Français en usage au Québec et dictionnaires. In C. Verreault, L. Mercier & T. Lavoie (eds.) *Le français, une langue à apprivoiser: textes des conférences prononcées au Musée de la civilisation (Québec, 2000-2001) dans le cadre de l’exposition Une grande langue: le français dans tous ses états*. Québec: Presses de l’Université Laval, pp. 93-111.
- Projet Babel*. Accessed at <http://projetbabel.org/index.php> [30/05/2020]
- Renders, P. (2015). L’informatisation du Französisches Etymologisches Wörterbuch. Modélisation d’un discours étymologique. Strasbourg: ELiPhi-Editions de linguistique et de philologie.
- Rézeau, P. (2001). *Dictionnaire des régionalismes de France*. Bruxelles: De Boeck/Duculot.
- Robillard, D. de (2008). Revendiquer une lexicographie francophone altéritaire constructiviste pour ne plus saler du sucre. In C. Bavoux (ed.) *Le français des dictionnaires*. Bruxelles: De Boeck Université, pp. 321-335.
- Sabou, M., Bontcheva, K. & Scharl, A. (2012). Crowdsourcing research opportunities: lessons from natural language processing. In *i-KNOW ’12 Proceedings of the 12th International Conference on Knowledge Management and*



*Knowledge Technologies.*

- Steffens, M. (2017). Lexicographie collaborative, variation et norme: le projet 10-nous. In *Repères – Dorif*, 14. Accessed at [https://www.dorif.it/ezine/ezine\\_articles.php?art\\_id=393](https://www.dorif.it/ezine/ezine_articles.php?art_id=393) [30/05/2020]
- Steffens, M. (2018). Figement, langage et didactique du français. In *Langues et linguistique*, 37, pp. 120-134. Accessed at <http://www.lli.ulaval.ca/recherche/revues/revue-langues-et-linguistique/index-des-numeros-et-articles/vol-37-2018/> [30/05/2020]
- Steffens, M. & Baiwir, E. (in print). Intégrer la variation diatopique à l'enseignement du français: le rôle des outils numériques, *ÉLA*.
- Steffens, M., Dolar, K., & Gasparini, N. (2020). Structuration de données pour un dictionnaire collaboratif hybride. In: *Terminologie & Ontologie: Théories et Applications. Actes de la conférence TOTh 2019*, pp. 413-426.
- Steinlin, J., Kahane, S., Polguère, A. & El Ghali, A. (2004). De l'article lexicographique à la modélisation objet du dictionnaire et des liens lexicaux. In *Actes de EURALEX'2004*, pp. 177-186.
- Swiggers, P. (1999). La géographie linguistique de Jules Gilliéron: aux racines du changement linguistique. In *Cahiers Ferdinand de Saussure*, 51, pp. 113-132.
- Thibault, A. (2004). Dictionnaire suisse romand: Particularités lexicales du français contemporain. Geneva: Éditions ZOÉ.
- Thibault, A. (2008) (ed.). Richesses du français et géographie linguistique: Recherches lexicographiques sur les variétés du français en France et hors de France (Tome 2). Bruxelles: Duculot.
- Urbandico*. Accessed at <http://www.urbandico.com> [30/05/2020]
- Urban Dictionary*. Accessed at <https://www.urbandictionary.com> [30/05/2020]
- Usito*. Accessed at <https://usito.usherbrooke.ca/> [30/05/2020]
- Vincent, N. (2011). Combien faut-il de dictionnaires pour décrire le français? In O. Bertrand & I. Schaffner (eds.) *Variétés, variations et formes du français*. Palaiseau: Éditions de l'École polytechnique, pp. 389-404.
- Vincent, N. (2016). La prise en compte de plusieurs variétés nationales dans un dictionnaire du français: exercice de lexicographie pratique. In C. Molinari & D. Gavinelli (eds.) *Espaces réels et imaginaires au Québec et en Acadie: enjeux culturels, linguistiques et géographiques*. Milano: Led.
- Vincent, N. (2017). Présence et légitimité des variétés nationales dans les dictionnaires gratuits en ligne. In *Repères – Dorif*, 14. Accessed at [https://www.dorif.it/ezine/ezine\\_articles.php?art\\_id=379](https://www.dorif.it/ezine/ezine_articles.php?art_id=379) [30/05/2020]
- Violette, I. (2006). Pour une problématique de la francophonie et de l'espace francophone: réflexions sur une réalité construite à travers ses contradictions. In *Francophonies d'Amérique*, 21, pp. 13-30.
- Völker, H. (2009). La linguistique variationnelle et la perspective intralinguistique. In *Revue de Linguistique Romane*, 73, pp. 27-76.
- W3C (2004). Accessed at <http://www.w3.org/2004/OWL> [30/05/2020]
- Wiktionnaire*. Accessed at [https://fr.wiktionary.org/wiki/Wiktionnaire:Page\\_d%E2%80%99accueil](https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d%E2%80%99accueil) [30/05/2020]



# Reduce, Reuse, Recycle: Adaptation of Scientific Dialect Data for Use in a Language Portal for Schoolchildren

Ježovnik J., Kenda-Jež K., Škofic J.

*Research Centre of the Slovenian Academy of Sciences and Arts, Slovenia*

## Abstract

The children's language portal *Franček*, currently under development, will consist of eight modules providing pupils and secondary-school students with a variety of lexical information. The entries in the dialect module consist of onomasiological and semasiological sections, and an optional commentary. The dialect module was derived from dialect data contained in the two already-published volumes of *Slovenian Linguistic Atlas* (SLA), a dialect atlas aimed primarily at qualified readers. As the original presentation was deemed too complex for direct use in education, indices of morphological analyses were used instead. They were first transformed into a custom XML format, following which descriptive data from SLA was used to mark some dialect forms for exclusion, to assign frequency labels to others, and to add commentaries. Finally, secondary entries, which form the basis of semasiological sections, were generated. Beside links to original dialect maps and entries in SLA, the dialect module will also include a recording interface through which pupils will be encouraged to record and submit dialect lexemes from their own dialects.

**Keywords:** Slovene dialects; Slovenian; *Franček*; lexicography; children's dictionary; dialect data

## 1 Introduction

*Franček* ([www.francek.si](http://www.francek.si), currently in beta version) is a language portal aimed at Slovenian schoolchildren – the name is a diminutive of *Fran* ([fran.si](http://fran.si)), the central Slovenian dictionary portal –, currently under development within the scope of a European Social Fund project *Portal Franček, Language Counselling Site for Teachers of Slovenian and School Dictionary of Slovenian* (2017–2021), led by the Research Centre of the Slovenian Academy of Sciences and Arts, Fran Ramovš Institute of the Slovenian Language. The language portal will provide lexical information to pupils and students in elementary and secondary schools. It is structured around a central headword list, to which eight modules, i.e. databases with different types of lexical information, are linked. The modules provide information on the words' meanings, synonyms, morphological characteristics, pronunciation, pragmatics, dialectal variation, etymology, and historical usage. The content of the modules as seen on the web portal is directly visualized from an underlying database and not manually constructed per se, and based on the users' age group, certain modules are either visualized or omitted. The dialect module is based on dialect data gathered in the field as part of the compilation of the *Slovenian Linguistic Atlas* (SLA), the most comprehensive dialect atlas of Slovenian featuring a detailed analysis of spatial lexical variation of basic vocabulary and cultural heritage lexicon presented through the use of geolinguistic methods. Despite its heterogeneity and a few apparent drawbacks (see below), the dialect material gathered for SLA is to this day the richest and most extensive resource on Slovenian spatial linguistic variation.

## 2 Spatial Linguistic Variation in Slovenian

Slovenian is a dialectally highly diverse language, featuring more than 40 distinct dialect varieties spoken by approximately 2 million speakers on an area of roughly 25.000 km<sup>2</sup> (excluding the Slovene-speaking diaspora in non-neighbouring countries). Geographical, historical, societal, and political circumstances have brought about the emergence of different types of speech communities. These differ not only in use of dialect varieties, sometimes so distant from one another to impede mutual intelligibility, but also by the status of their respective dialects in the language repertoire of a community, their use in different (public and private) domains, and their role in building a local (and national) identity. Next to dialects with high prestige and a relatively stable structure (cf. Kenda-Jež, Bitenc 2015) there are also those more subjected to dialect levelling and merging into regional dialects (Lundberg 2013: 69–96). Many Slovenian-minority speech communities in Italy, Austria, and Hungary are characterized by weak intergenerational language transmission (Steenwijk 2003: 221–223; Pronk 2009: 4; Zorko 2009: 15); this is also true for suburbanized areas near large cities with higher levels of migration (Škofic 1998).

Given such diversity, complicated patterns of dialect use and prestige, and frequent diglossic relations with the standard language, depending on the dialect, finding a suitable way of presenting dialect data to children is at the same time very important and not without its difficulties. In addition to providing teaching support in accordance with school curricula, our main goal was to increase awareness of spatial language variation and the status of dialects in local communities.

The stereotypically negative attitude towards the use of dialects within educational settings has only in the second half of the 20<sup>th</sup> century been replaced with a gradual understanding of the relations between local nonstandard varieties and the spoken standard language (Gruden, in print) as used in formal domains. Due to differing linguistic situations in different microsettings it is reasonable to adapt curricula to specific circumstances of particular localities; this depends, to a large



extent, on the ableness, or rather, differing relative difficulty of integration of local linguistic variety uses into a wider context of spatial variation of the Slovene language.

### 3 Slovenian Linguistic Atlas as a Resource

The main sources of information for the dialect module were the two published volumes of the most comprehensive dialect atlas of Slovenian, the *Slovenian Linguistic Atlas* (SLA) (Škofic et al. 2011; Škofic, Šekli et al. 2016), which focus on the semantic fields of *humans* (specifically: the body, illnesses, and family) and *farms*, respectively.

The concept for SLA was first formulated in the 1930s by Fran Ramovš. Its author developed the concept according to the principles of contemporary European linguistic geography adhering to the French model established by Gillieron (ALF), which favoured the format of a “geographically arranged dictionary” (Ramovš 1934) with point-text maps. SLA was first envisioned as a preliminary survey of dialect variation; thus, neither the exact layout of the network of 312 research points nor its density were conclusively finalized at that time. After a decade of intense fieldwork research by Tine Logar (1946–1958), which also served to assess the validity of the proposed classification of Slovenian dialects, the project switched from short-term to long-term. Owing mostly to the complexity of the dialect matter, the French model preferring a single fieldwork researcher as a means of achieving unity of registration was deemed inappropriate and a number of researchers were assigned to the project, including university students. The related questionnaire (cf. Benedik 1999) remained the main tool for research of phonetic and phonological features of Slovene dialects well into the 1990s.

Each entry in SLA covers a single question of the afore-mentioned questionnaire. It consists of a dialect map, a list of all registered dialect lexemes as originally transcribed, arranged by research points, and a comprehensive commentary. The latter is divided into several subsections, most relevant of which for the purposes of this paper is the so-called morphological analysis. It serves to consolidate dialect variations (due to phonetic differences) of lexemes to their uniform standardized forms and segment them by morphemes to thus provide information on their word-formational characteristics as well as their origin (inherited or borrowed).<sup>1</sup> The morphological analyses also form the basis for cartographic representations. In contrast to the original concept, the dialect lexis is not presented by superimposing the transcribed data directly on the map but by various types of qualitative point-symbol maps, i.e. lexical, word-formational, and semantic maps. The symbols are chosen on the basis of morphological analyses in accordance with the methodology of the *Slavic Linguistic Atlas* (OLA 36–38, 55–59) and sometimes supplemented with isoglosses. Maps with abstract presentation of cross-lexemic and intralexemic relations require a skilled reader who is able to relate graphic representations with the relevant data from the commentary. As such, they are not well-suited for direct use in an educational setting.

Owing mostly to the longitude of the project, the underlying dialect database is very heterogeneous, which further complicates attempts at simplified, yet still accurate presentations. Through the years, the methodology has undergone a number of transformations, motivated by new findings on one hand and by theoretical and methodological developments on the other:

- At the beginning, a specially designed questionnaire consisted of 646 lexical and 170 grammatical questions; both the selection of the lexis and the selection of phonetic and morphological characteristics to be surveyed were based on the knowledge of dialectal differentiation available at the time. In 1961, the questionnaire was amended and restructured to enable a more structured approach. Following restructuring, the lexical part of the questionnaire currently consists of 802 onomasiological and 25 semasiological questions.
- The model of the preferred dialect speaker/informant has changed as well.<sup>2</sup> Based on early fieldwork experience, a two-fold model for surveying different age and socio-educational groups evolved, preferring: (1) children up to the age of 14 and speakers between the ages 46 and 70; and (2) the rural population and high-school or university-educated informants, mostly teachers (Kolarič 1954: 185–186; Logar 1958/59: 129). Until the year 1958, 42 children up to the age of 15 and 28 teenagers between the ages 16 and 20 participated in the research (vs. only 35 informants between the ages 71–95). After 1958, the described model was gradually replaced with one favouring speakers of the oldest generation. The distribution of informants by gender is balanced. The core group of informants was born between 1890 and 1930; the average interviewee age has risen by approximately 30 years between 1946 and 2000 (35 vs. 66, Kenda-Jež 2002: 155–160).
- The majority of the interviews by Tine Logar were conducted in the years 1946–1965. The intensity of fieldwork somewhat lessened due to a period of intensive fieldwork for OLA (1966–1975). Fieldwork again increased in the 1980s and again in the 2010s, just before print publications of the first two volumes of SLA.
- From 1975 onward, the dialect data is transcribed according to a standard Slovenian phonetic transcription, a derivative of the standard used in (OLA). Prior to this standardization, several formats of transcription had been used over the years. Because dialect data has been recorded through a longer period of time and because a retroactive harmonization especially of the older data would require significant rechecking in the field, it was decided to instead publish the dialect material as originally transcribed to avoid confusion or mistakes in the process of transliteration. Due to tradition and the already-mentioned connections to OLA, IPA has not been introduced into Slovenian dialectology on a systemic scale.

<sup>1</sup> Information on borrowed lexicon, i.e. which language (and which of its temporal or spatial varieties the lexeme was borrowed from), is, to list but one example, useful for assessing language contact in the past and present, aiding cultural, historical, ethnological, sociological etc. studies.

<sup>2</sup> The model is not consistent with the generalized notion of the “ideal speaker” of classic Anglo-Saxon dialectological studies based (NORM = non-mobile older rural male; Chambers, Trudgill 2002: 29).



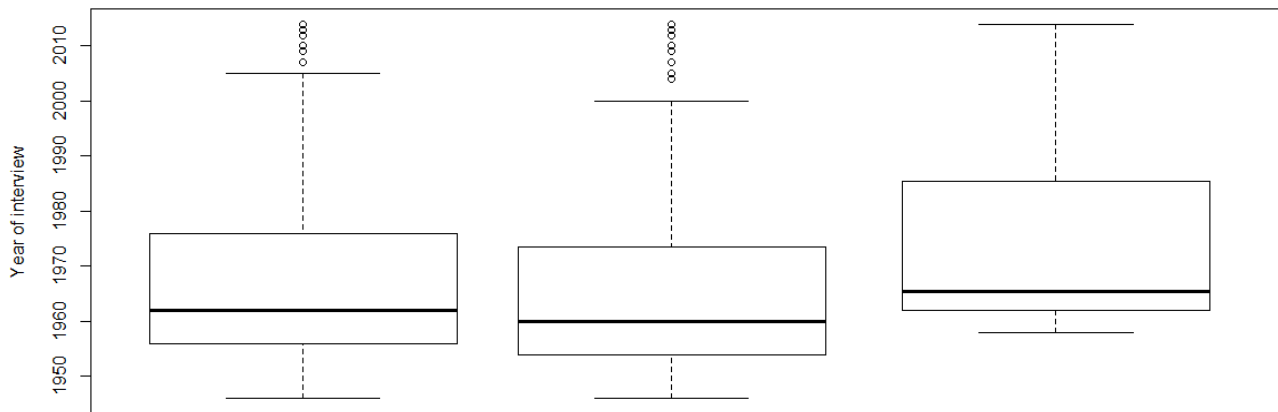


Figure 1: Boxplot presentations of the median years of interviews – left to right: all interviews (n=520, Me=1962); interviews conducted within Republic of Slovenia (n=436, Me=1960); interviews conducted in neighbouring countries in communities where Slovene is spoken as a minority language (n=84, Me=1965.5).<sup>3</sup>

#### 4 The *Franček* Dialect Module

The module providing information on word usage in Slovene dialects is intended for pupils older than 10 years. Entries in the dialect module have a three-part structure:

- an onomasiological section (“Which words are used to describe this concept and in which dialects?”),
- an optional commentary,
- a semasiological section (“Which concepts does this word (also) denote and in which dialects?”).

**1** Kako se reče **babica** v narečjih?

---

V narečjih se reče najpogosteje **stara mati**, **stara mama**, ponekod **babica**, **baba**, **babej**, oziroma **babi**, **oma**, **omama**, v primorski narečni skupini tudi **nona**.

🚩 Besede **nona**, **oma**, **omama** in **baka** so prevzete iz sosednjih jezikov. V koroških narečjih se reče krajše tudi **bica** namesto **babica**.

**2** Kaj še pomeni **babica** v narečjih?

---

Beseda **babica** lahko ponekod pomeni **tašča**.

📍 Kje govorijo tako? Poglej na [zemljevidu](#).

**3** 🗣️ Te zanima podrobnejši opis narečne rabe te besede? Poglej v [Slovenski lingvistični atlas](#) na Franu.

🎧 Ali tudi pri vas rečete **babica**? Ali kako drugače? [Posnemi se in nam pošlji](#) svoj posnetek.

Figure 2: Entry *babica* “grandmother” featuring an onomasiological section (1) with commentary (marked with a flag) and a semasiological section (2) with links (3) below.<sup>4</sup>

Each entry consists of at least one onomasiological or semasiological section, though combinations of the two types under the same entry are possible (as in the example in Figure 2). If an entry has multiple meanings, these are treated in separate iterations of sections.

The onomasiological section is introduced by a title question “How do we say (headword) in dialects?”, followed by a typified list of normalized dialect lexemes equipped with so-called frequency labels or, in the case of geographically specific lexemes, with a corresponding listing from a simplified list of 40 (sub)dialects and 7 dialect groups. Headwords of onomasiological sections are the commonest standard-Slovene lexemes denoting concepts in question that are also used as keywords in the SLA questionnaire.

The optional commentary conveys various information of interest on the words’ origins, their status as loanwords, on

<sup>3</sup> The data was analysed in R by Kaja Hacin Beyazoglu.

<sup>4</sup> Graphic presentations from the portal *Franček* are not yet final and subject to potential change.



phonetic, accentual, and morphological variation, etc.

The semasiological section is introduced by the question “What can (headword) also mean in dialects?” If original entries in SLA provide answers to the question: “Which words do you use to describe this concept?”, then it follows logically that dialect lexemes inversely beg the question: “Which concept are you expressing by using this word?” Headwords of semasiological sections are standardized forms of dialect lexemes.

Links to original SLA maps and SLA entries on the portal Fran are provided in the footer (first and second lines under 3 in Figure 2). The final line of the footer encourages pupils and students to record and submit lexemes typical of their own dialect varieties (cf. Section 4.4).

Construction of the dialect module occurred in three basic stages, which we dubbed Reduce, Reuse, and Recycle.

#### 4.1 Reduce

In the first stage, the amount of information provided by the original database was reduced. As the original dialect maps were meant as scientific presentations of dialect data, the originals were deemed too complex for school use (cf. Section 3). Because SLA was conceived as a primarily printed publication, some crucial data was edited or added to the maps manually and couldn't be faithfully replicated from the underlying database in an automated fashion without significant manual input. It was therefore decided to forgo them in favour of textual descriptions.<sup>5</sup> The entries nevertheless include links to original cartographic material for advanced pupils or as teaching material.

Likewise, it was agreed that original lists of phonologically transcribed dialect data would be too difficult for users to understand and would make linking to the main headword list an impossible task. Instead, lists of standardized forms were extracted from the indices of morphological analyses, converted to a custom XML format by key of question codes from the SLA questionnaire, and normalized. This process yielded 238 entries containing a total of 5108 dialect lexemes which formed a basis for the onomasiological sections.<sup>6</sup> Out of these, six so-called semantic entries were eliminated because they differed from the rest conceptually: instead of listing dialect expressions denoting their respective semantemes, these SLA entries were meant to probe for semantic variation of select lexemes in dialects; out of these six, five headwords are also included as normal entries in SLA in any case.

##### B

**baba** 1/110 (V610), 1/125 (V639)

**babej** 1/110 (V610), 1/127 (V612)

**babi** 1/110 (V610)

**babica** 1/110 (V610), 1/127 (V612)

[...]

**mama** 1/104 (V605), 1/110 (V610), 1/119 (V624), 1/127 (V612) ▶ **mama lastra** 1/119 (V624); **mama (ta) stara** 1/110 (V610); **od mame brat** 1/113 (V616); **od moža mama** 1/127 (V612); **stara mama** 1/127 (V612); **(ta) pisana mama** 1/119 (V624); **(ta) stara mama** 1/110 (V610)

**mamej** 1/104 (V605)

**mamica** 1/110 (V610)

**mamika** 1/104 (V605)

[...]

**stara** ▶ **njegova stara** 1/138 (V637); **(ta) stara** 1/110 (V610), 1/125 (V639); **ta stara** 1/127 (V612)

**stara mati** 1/110 (V610)

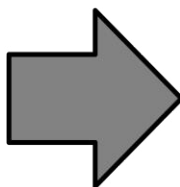
**staraka** 1/110 (V610)

**staratelj** 1/123 (V793)

**starci** 1/102 (V646)

**starček** 1/109 (V609)

**starčka** 1/110 (V610)



```
<geslo geslo-id="000029">
  <iztočnica>stara mati</iztočnica>
  <dolga_iztočnica>stara mati</dolga_iztočnica>
  <SLA_sklop>
    <karta_komentar_SLA>1/110</karta_komentar_SLA>
    <vpr_SLA>V610</vpr_SLA>
  </SLA_sklop>
  <narečni_sklop>
    <narečna_oblika>
      <oblika>(ta) stara</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>(ta) stara mama</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>(ta) stara mati</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>baba</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>babej</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>babi</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>babica</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>baka</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>mama (ta) stara</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>mama</oblika>
    </narečna_oblika>
    <narečna_oblika>
      <oblika>mamica</oblika>
    </narečna_oblika>
  </narečni_sklop>
</geslo>
```

[...]

Figure 3: Extract from the indices of morphological analyses (left, with question codes underlined) and the entry *babica* “grandmother” (originally *stara mati* – multi-word headwords were changed to their usual single-word synonyms to enable linking to the main headword list) in custom XML code post-transformation (right, with dialect forms underlined).

<sup>5</sup> A separate project is being pursued that would enable such a reworking and provide the public with an interactive and enriched presentation of dialect data, including interactive maps (Škofic 2013). A limited beta version is published at [http://gismo.zrc-sazu.si/flexviewers/Test\\_Vicic/SLA1/SLA\\_kmetija/](http://gismo.zrc-sazu.si/flexviewers/Test_Vicic/SLA1/SLA_kmetija/).

<sup>6</sup> Three entries published separately in Jakop 2012 in the format of SLA entries were added to the database manually.



Afterwards, problematic forms, included in SLA for scientific accuracy and marked accordingly, were manually marked for exclusion. This included irrelevant forms (“wrong answers”, e.g. nomen loci *gnojšče* “place where manure is kept” under the question pertaining to “manure” beside the answer *gnoj* “manure” in the same research point); unclear forms (e.g. hapax legomenon *gjam* “manure”, limited to one research point); dubious forms (e.g. in modern standard Slovene the lexeme *truplo* (used as a question in the SLA questionnaire) denotes “cadaver”, but it also used to carry the meaning of “body” in contemporary standard Slovene at the time of the earliest interviews; consequently, and also due to dialect lexical variation, it is not always clear which of the two meanings the informants provided answers for).

## 4.2 Reuse

Information from original commentaries and linguistic maps was reused to assign frequency labels to remaining dialect lexemes for each entry. The frequency labels range from *everywhere* for universally used lexemes through *most frequently* for commonly used lexemes, *in some parts* for lexemes used in more than one dialect groups and finally *rarely* for lexemes used in more than one dialect not in the same dialect group. Geographically specific lexemes limited to only one dialect group or only one dialect were labelled with the corresponding listing from a simplified list of 40 (sub)dialects and 7 dialect groups.

Based on the same data regarding frequency and distribution, particularly infrequent lexemes were marked for exclusion from onomasiological sections as well. This was necessary to achieve greater clarity, i.e. to avoid overly bloated entries, as almost half of the entries contained more than 30, and some as many as 80–90 different dialect lexemes.

The same information was finally reused to compose commentaries where deemed necessary, sometimes including some of the previously excluded forms as curiosities. Commentaries typically pertain to notable (i.e. recognized as locally typical by a wider populace) phonetic or accentual variation with lexically identical items, e.g. bilabial pronunciation of /l/ (e.g. *vas* as opposed to *las* “hair”) or typical stress placement (e.g. *óko*, *babica* as opposed to the more frequent *okó* “eye”, *bábica* “grandmother”), notable semantic peculiarities (e.g. *hči* meaning also “female child” beside the more frequent “daughter”), origin of especially uncommon lexemes or loanwords from neighbouring languages (e.g. *lilahen*, *vilahen*, both “bedsheet”, borrowed from Middle High German, or *baka*, *oma*, *nona*, all “grandmother”, borrowed from Croatian, German, and Italian/Friulian, respectively) with the particular aim of increasing awareness about language contact and interaction, etc. The total count of all included dialect lexemes was brought down to 1706, and a total of 126 commentaries were added.

## 4.3 Recycle

As the last step before linking the dialect module with the main headword list, semasiological sections were added. As already mentioned, if onomasiological sections provide answers to the question: “Which words do you use to describe this concept?”, it follows logically that dialect lexemes inversely beg the question: “Which concept are you expressing by using this word?” This was achieved by automatically generating new entries from dialect lexemes (or adding subsections to existing ones in the cases of separate (non-parent) homonymous entries) and the process yielded 1508 additional so-called secondary entries. In secondary entries, the dialect lexemes of “primary” entries become headwords and the headwords of their parent primary entries are carried over to be used as definitions (cf. Figure 4); the frequency labels are also copied. While such recycling might seem redundant within the scope of the dialect module itself, it comes to relevance when accessing the information through an outside source, i.e. the headword list, which might include headwords not included in the module as primary entries, but nevertheless present as dialect lexemes.

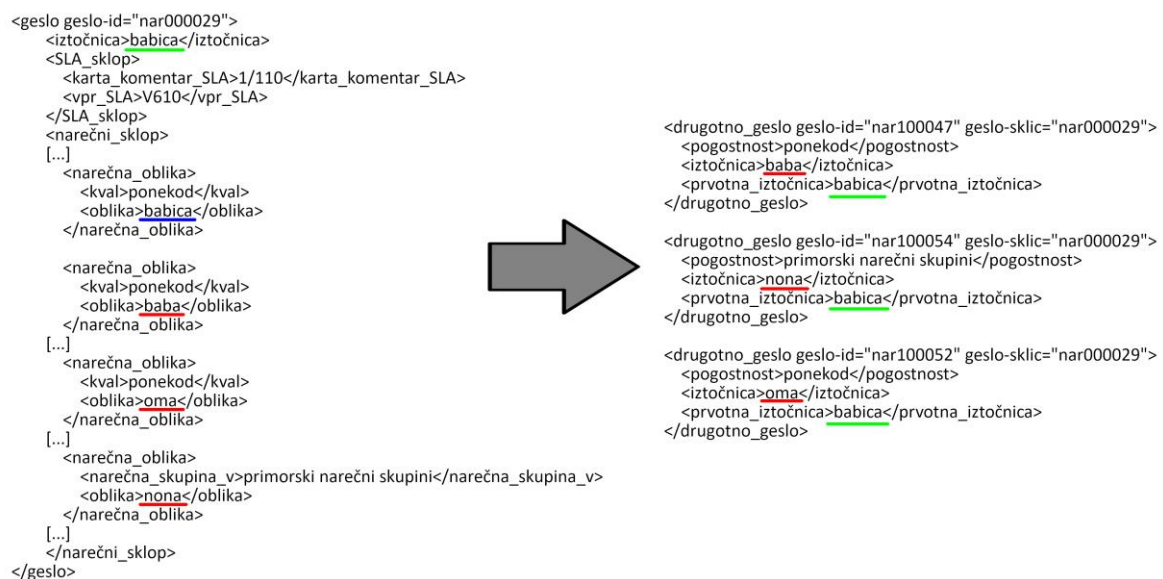


Figure 4: “Primary” entry *babica* “grandmother” (right) and its derived secondary entries (left). Dialect forms serve as headwords for secondary entries (underlined in red) except when they are homonyms of their parent headwords (the case of *babica*, underlined in blue on the left); the parent headwords serve as definitions in secondary entries (underlined in green).



#### 4.4 Renew

As previously mentioned, fieldwork gathering of dialect data for the purposes of compiling SLA began in 1946 and the majority of data was gathered in the 1960s (cf. Figure 1). As one anonymous reviewer of the extended abstract justly notes, these facts bring into question the currency of the presented dialect data. To partly mitigate this drawback, the dialect module will also provide the option for children to record and submit their own suggestions. In entries with onomasiological sections, the portal's visitors are prompted to record the dialect lexeme used in their local environments and submit it; in entries with semasiological sections, users are instead prompted to provide meanings for headwords, if they recognize them from their local environments. The submitted entries will be approved by a moderator and published for other users to listen to. Headwords with no entries in the dialect module will feature only the described prompt in order to, in time, fill the gaps in the module. Beside expanding and updating the current database with more recent dialect data – possibly beyond the scope of the portal itself –, this feature also aims to encourage interactivity in the school setting and children's participation.

#### 5 Conclusion

Given the dialectal diversity of Slovenian, complicated patterns of dialect use and prestige, and diglossic relations with the standard language, depending on the dialect, finding a suitable way of presenting dialect data to children is very important and difficult at the same time. Despite providing dialect data that is in some cases potentially outdated for reasons outside the developers' control, the dialect module of the *Franček* children's language portal will hopefully represent a major step in that direction. Further assessment in collaboration with 19 elementary and secondary schools is pending at the time of writing; it will guide additional fine-tuning of the presentation, and the module is expected to expand with the release of further volumes of SLA.

#### 6 References

- ALF = Gilliéron, J. (1902–1910). *Atlas linguistique de la France* 1–13, Paris: Champion.
- Benedik, F. (1999). *Vodnik po zbirki narečnega gradiva za Slovenski lingvistični atlas (SLA)*. Ljubljana: ZRC SAZU, Založba ZRC.
- Chambers, J. K., Trudgill, P. (2002). *Dialectology*. Cambridge: University Press.
- Gruden A. [in print]. Poimovanje socialne zvrstnosti pri govornih slovenščine in vpliv šolanja. In M. Bitenc, M. Stabej (eds.) *Sociolingvistične iskricke*, Ljubljana: Znanstvena založba Filozofske fakultete.
- Jakop, T. (2012). Izrazi za spolovila v gradivu za Slovenski lingvistični atlas in pri Ivanu Koštiću. In *Jezikoslovni zapiski*, 18 (2), pp. 37–55.
- Kenda-Jež, K., Bitenc M. (2015). Language Variation in Slovene: A case study of two geographically mobile speakers. In Torgensen et al. (eds) *Language Variation – European Perspectives V: Selected papers from the Seventh International Conference on Language Variation in Europe (ICLaVE 7), Trondheim, June 2013*. Amsterdam – Philadelphia: John Benjamins, pp. 31–42.
- Kenda-Jež, K. (2002). Model idealnega govornika v slovenskih dialektoloških raziskavah. In M. Jesenšek et al. (eds.) *Med dialektologijo in zgodovino slovenskega jezika: Ob življenjskem in strokovnem jubileju prof. dr. Martine Orožen*. Maribor: Slavistično društvo, pp. 150–165 (Zora 18).
- Kolarič, R. (1954). Die slowenische Mundartforschung. In *Orbis: Bulletin International de Documentation Linguistique* (Louvain) 3 (1), pp. 182–188.
- Logar, T. (1958/59). Iz priprave za lingvistični atlas. In *Jezik in slovstvo*, 4, pp. 129–135.
- Lundberg, G. (2013). *Dialect leveling in Haloze, Slovenia*. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta.
- OLA = *Общеславянский лингвистический атлас – Вступительный выпуск: Общие принципы. Справочные материалы* (1994). Moskva: Nauka.
- Pronk, T. (2009). *The Slovene Dialect of Egg and Potschach in the Gailtal, Austria*. Amsterdam – New York, NY: Rodopi.
- Ramovš, F. (1934). *SLA – Slovenski lingvistični atlas*. Manuscript, stored by the Library of SAZU.
- Steenwijk, H. (2003). Resian as a minority language. In M. Janse, S. Tol (eds.), *Language Death and Language Maintenance: Theoretical, practical and descriptive approaches*, Amsterdam – Philadelphia: John Benjamins, pp. 215–226.
- Škofic, J. (1998). Govor celjskega predmestja Gaberje. In *Jezikoslovni zapiski*, 4, pp. 89–98.
- Škofic, J. (2013). Priprava interaktivnega Slovenskega lingvističnega atasa. In *Jezikoslovni zapiski* 19 (2), pp. 95–111.
- Škofic, J. (ed.), Gostenčnik, J., Horvat, M., Jakop, T., Kenda-Jež, K., Kostelec, P., Nartnik, V., Petek, U., Smole, V., Šekli, M., Zuljan Kumar, D. (2011). *Slovenski lingvistični atlas 1. Človek (telo, bolezni, družina)*. Ljubljana: Založba ZRC, ZRC SAZU. Accessed at: <https://fran.si/204/sla-slovenski-lingvisticni-atlas> [28/05/2020]
- Škofic, J. (ed.), Šekli, M. (ed.), Gostenčnik, J., Hazler, V., Horvat, M., Jakop, T., Ježovnik, J., Kenda-Jež, K., Nartnik, V., Smole, V., Zuljan Kumar, D. (2016). *Slovenski lingvistični atlas 2. Kmetija*. Ljubljana: Založba ZRC, ZRC SAZU. Accessed at: <https://fran.si/204/sla-slovenski-lingvisticni-atlas> [28/05/2020]
- Zorko, Z. (2009). Prekmursko goričko podnarečje v Porabju na Madžarskem. In Irena Novak Popov (ed.), *Slovenski mikrokozmosi – medletni in medkulturni odnosi*, Ljubljana: Slavistično društvo Slovenije (Zbornik Slavističnega društva Slovenije 20), pp. 13–27.



### Acknowledgements

This paper was financed as part of the project *Portal Franček, Language Counselling Site for Teachers of Slovenian and School Dictionary of Slovenian* by the Ministry of Culture of the Republic of Slovenia and by the European Social Fund (contract No. C3340-17-20800), and the research programmes P6-0038 and P5-0408 financed by the Slovenian Research Agency.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Research on Dictionary Use**







# “Game of Words”: Play the Game, Clean the Database

Arhar Holdt Š., Logar N., Pori E., Kosem I.

*University of Ljubljana*

## Abstract

The paper presents the “Game of Words” (in Slovene: Igra besed), a mobile application purposed for a gamified improvement of two automatically compiled dictionaries for Slovene: the Collocations Dictionary of Modern Slovene and the Thesaurus of Modern Slovene. We provide a brief history of the game, and introduce its two modules that utilize collocation and synonym data respectively. A significant part of the paper is dedicated to the presentation of all the steps of the preparation of both datasets; this included addressing challenges brought by automatically extracted data from the corpus, and filtering out sensitive content considering the potential users. Crowdsourcing aspects of the game are discussed, especially in terms of the lessons learned in the development process, and how one needs to strike a good balance between the lexicographic intentions, numerous possibilities of using language information, and the enjoyment and motivation of playing the game. The paper concludes by outlining future plans, including further developments of the game both on the level of game modules and languages offered, in the framework of European projects and initiatives.

**Keywords:** Game of Words; GWAP; collocations; synonyms; crowdsourcing; gamification; responsive dictionary

## 1 Introduction

The recently published Collocations Dictionary of Modern Slovene and the Thesaurus of Modern Slovene<sup>1</sup> are innovative from the perspective of the dictionary-making process, introducing a concept of a “responsive dictionary” – a dictionary that is compiled entirely through automatic extraction methods, as soon as possible made available to the community both as a lexical database and as an online language resource, and after that continuously and transparently lexicographically improved, also with the help of user-provided feedback. For example, through the dictionary interface, users can vote positively or negatively on the automatically extracted data, and in the case of the Thesaurus also suggest additional synonyms to be included in the dictionary database. The Collocations Dictionary and the Thesaurus were already presented in literature (Krek et al. 2017; Arhar Holdt et al. 2018; Kosem et al. 2018a). Together with the methodology for user involvement, the need for getting users motivated to participate in resource improvement and enhancement was highlighted as one of the most crucial elements of the newly proposed workflows.

An important aspect of user involvement in the responsive dictionary development is that the involvement is direct and the task is explicit, i.e. the users are aware of the purpose of the task and the aim of their participation. However, we have to keep in mind that the involvement/feedback via the dictionary interface is secondary to dictionary consultation. This can significantly affect user motivation and the time they are willing to dedicate to providing feedback or suggestions. Consequently, we have decided to look into the possibilities offered by gamification, specifically games with a purpose, where the main purpose for the users is enjoyment, while the task remains in the background so that the users are often unaware they are providing information useful for linguistic/lexicographic purposes. As a result, we have developed a language game called Igra besed (Game of Words) that challenges players on their knowledge of Slovene collocations and synonyms, while supporting the improvement of the previously mentioned lexical resources: the Collocations Dictionary of Modern Slovene and the Thesaurus of Modern Slovene.

In the paper, we present the idea behind the mechanics of the game and the implementation of all its modules, together with the description of the data-preparation process. Namely, the data to be included in the game had to be filtered in order to avoid sensitive issues, such as derogatory and potentially offensive lexica. Then, the collocation and synonym module are presented in more detail, followed by the discussion of the crowdsourcing perspectives and shortcomings of the game. We conclude by offering final remarks and presenting future plans, including the development of next versions of the game in a wider European context, i.e. collaboration with current European projects and actions.

## 2 Gamification and Language Resources

Gamification is closely related with the notion of Games with a Purpose (GWAP), a crowdsourcing mechanism for (typically benevolent, i.e. voluntary, not paid) implicit crowdsourcing. In the categorisation of crowdsourcing approaches, “implicit” means that the purpose of the task is secondary or even partially hidden to the participants, as opposed to “explicit” crowdsourcing where the task is the primary purpose of participation (Lyding et al. 2018). In the case of

<sup>1</sup> Both dictionaries are freely available online, (also) through an English user interface: the Collocation dictionary at <https://viri.cjvt.si/kolokacije/eng/> and the Thesaurus at <https://viri.cjvt.si/sopomenke/eng/>.



GWAPs, participants' primary goal and the source of motivation is to play a game and by doing so, they perform a specific underlying, pre-designed task.

When it comes to creating lexical infrastructure, some successful GWAPs were designed to annotate language data, for example Phrase Detectives (Poesio et al. 2013), JeuxDeMots (Lafourcade 2007), and ZombiLingo (Guillaume 2016). Nonetheless, the use of gamification, and crowdsourcing in general, in lexicography is still very limited. While the benefits of crowdsourcing have been thoroughly established (Lew 2013; Abel & Meyer 2013; Benjamin et al. 2015; Fišer & Čibej 2017), the implementation lags behind. For example, one popular way of dictionaries promoting their activities as crowdsourcing (or citizen science) is enabling user feedback via online forms or emails. However, this is rarely crowdsourcing as it is based on individual rather than crowd contributions, plus the methodology of including suggestions in the lexicographic workflow is not necessarily transparent. Rather than turning to crowdsourcing for the sake of keeping up with the new trend, we propose a gradual inclusion of user-involving approaches, where new ideas and their implementation are thoroughly evaluated by the users and can be continuously improved. The evaluation of crowdsourcing techniques available through the dictionary interface of the Collocations Dictionary and the Thesaurus were presented by Pori et al. (2020) and Arhar Holdt (2020).

The beginnings of Igra besed (Game of Words) go back to 2014 when the first version was published as an online game.<sup>2</sup> The game was part of the project funded by the Slovenian Ministry of Culture, with the aim being to devise innovative ways to promote the Slovenian language, and its use. The first version of the game was not conceived as a Game with a Purpose, we simply wanted to make a game that would be fun and didactic at the same time. But due to the lack of suitable and free language resources we were forced into using automatically extracted data from the Sketch Engine, and relatedly, devising the game in a way where the potential noise in the data would not affect the playing experience. This was also the reason why the first version had only one playing mode: players had to type in three possible collocates of the word, which were then scored according to the ranking on the list. There were two formats of playing: practice and duel (participants were able to challenge another player to a duel, and the latter could accept or reject this challenge). As far as working with automatic data was concerned, the underlying assumption, which was later confirmed by the analysis of user data, was that the users will not intentionally type in wrong information (in this case collocates).

The transition of Game of Words to a GWAP was mainly driven by three developments. Firstly, the proposal for a new dictionary of Modern Slovene published at the time (Gorjanc et al. 2015), which was a response to the lack of lexicographic resources describing modern Slovene, described in detail how crowdsourcing methods could be implemented in lexicographic workflow to speed up the dictionary-making process. Moreover, also as the answer to the lack of resources on modern Slovene, was the introduction of responsive dictionaries - using the approach "publish good (automatically extracted) data now, clean later" -, of which the crowdsourcing component was a key part. Secondly, crowdsourcing experiments with collocations we have conducted have indicated that explicit crowdsourcing was not the most suitable method for dealing with this particular type of lexical information. And thirdly, a detailed analysis of Game of Words logs has pointed out a potential of the game for not only validation of collocational information, but also for other more complex tasks such as determining the definite or indefinite form of the adjective before noun.

In version 2, Game of Words was thus significantly upgraded, in terms of content, playing modes, and medium. Much more collocational data was included, not only in terms of number of lemmas and collocations but also in terms of syntactic structures. Also, synonym data and synonym playing module were added. Importantly, based on the feedback of the users of the first version, the game has moved from the online to the mobile medium, i.e. was developed as a mobile app, available both for Android and iOS devices. The development of the version 2 was made within two different projects funded by the Slovene Ministry of Culture; "The promotion of a language mobile app" funded the development of the mobile app and the upgrade of the collocation module, and "The promotion of the Thesaurus of Modern Slovene" funded the addition of the synonym module.

### 3 Data Preparation

#### 3.1 Collocational Data

The basis for the collocation module of the game was the Collocations Dictionary of Modern Slovene, comprising 35,989 headwords and 7,338,801 collocations. Collocations were automatically extracted from the 1.2-billion-word Gigafida corpus of Slovene (Logar Berginc et al. 2012), using the Sketch Engine API, and also additionally filtered at the post-processing stage (for more see Kosem et al. 2018a). The Collocations Dictionary offered a rich resource of potential data for the game, allowing us to expand on the number of syntactic structures offered, something that was also requested by numerous players of the first version of the game. Based on the findings of the evaluation of automatically extracted collocational data, which was conducted within the KOLOS project<sup>3</sup> (Kosem et al. 2018b), we selected five syntactic structures that exhibited the highest percentage of good collocation candidates: adjective + noun (*osnovna šola* 'primary school'), noun + noun in genitive (*rezervacija sobe* 'room reservation'), verb + noun in accusative (*prevesti tekst* 'to translate text'), adverb + verb (*ironično komentirati* 'comment ironically'), and adverb + adjective (*zelo lep* 'very beautiful'). Since headwords (nouns, verbs, adjectives, adverbs) occupied different positions in the structures, this meant nine different syntactic structures altogether (for "noun + noun in genitive", only the version with the headword in the first position was taken). The total number of collocations in these nine syntactic structures was 2,723,551.

<sup>2</sup> The game is still available online at <https://www.igra-besed.si/>, however, only through a Slovene interface.

<sup>3</sup> KOLOS is the acronym for the national research project "Collocations as a Basis for Language Description: Semantic and Temporal Perspectives", funded by the Slovenian Research Agency (J6-8255).



The second step involved reducing the number of headwords and collocations according to statistical, morphosyntactic, and semantic criteria. This was needed in order to address certain problems that could affect the playing experience. The statistical filter we added was a minimum of 10 collocates per syntactic structure; this was mainly needed because of the new playing mode Choose (see Section 4.1) which required at least nine collocates. It should be stressed that this filter was implemented after the database had been filtered according to morphosyntactic and semantic criteria.

Morphosyntactic filters were related to either known problems with corpus annotation, or problems in collocation form due to lack of suitable resources. Thus, we removed all collocations containing collocates that were not in the Slovene morphological lexicon Sloleks (Dobrovoljc et al. 2018), which was used for assigning the right form to the collocates according to the case required by the syntactic structure. This filter was not applied in structures adverb + verb and adverb + adjective as lemma forms were always used in them. Also, we removed all collocations containing collocates beginning with a capital letter (there were no such headwords) as the evaluation showed that a large proportion of them is noise or are in incorrect form. In fact, many of these collocates were already removed in an earlier step as they were not in Sloleks. Another problematic group that was removed were 73 homonymous headwords as they were found as one lemma, so the automatically extracted data contained collocations for all, e.g. noun headword *tema* contained data for *têma* ('dark') and *téma* ('topic'). Finally, we removed 1,370 verbs as headwords in the relation verb + noun in accusative, as these verbs never or very rarely occurred with an object, meaning that the majority of their collocates were errors.

The semantic filter used was in the form of a stoplist that contained words (featuring as either a headword or a collocate) with a negative connotation in at least one of its meanings. The list was based on existing resources such as dictionaries, and privately compiled lists by researchers or journalists.<sup>4</sup> Words on the list included insults, pejorative expressions, vulgar words, etc. but also words that could cause discomfort like verbs *ubiti* ('to kill'), *uničevati* ('destroy'), *groziti* ('threaten'). While it could be argued that many of these words are not really problematic, we gave priority to the fact that the game could also be used in educational settings with young(er) users.

In the final step, rather than filtering the results, we conducted an additional step of post-processing. Namely, we added reflexive pronouns "si" or "se" to 1,358 reflexive verbs, or when the verb was reflexive in one of its meanings, indicated the possible use of reflexive pronoun in brackets, e.g. *umivati (se)* ('to wash (oneself)'). This was needed because listing a verb without the reflexive pronoun could elicit incorrect collocates from the players.

The final dataset for the collocation module contained 23,303 headwords (9,132 nouns, 8,423 adjectives, 3,953 verbs, and 1,795 adverbs) and 2,448,994 collocations. For comparison, the first version of the game contained 10,578 headwords (5,237 nouns and 5,341 adjectives) and 2,928,177 collocations.<sup>5</sup>

### 3.2 Synonym Data

The basis for the synonym module of the game was the Thesaurus of Modern Slovene. In its current version, the Thesaurus comprises 105,473 (single- or multi-word) headwords. Synonyms for these headwords were obtained automatically from The Oxford®-DZS Comprehensive English-Slovenian Dictionary (Šorli et al. 2006) and the Gigafida reference corpus of written Slovene (Logar Berginc et al. 2012). The synonyms -- or more precisely, 'synonym candidates', as the data has not yet been lexicographically checked -- are separated into two groups: "core" and "near". According to their assigned score of relatedness to the headword, "core" synonyms are believed to be the most relevant, and "near" synonyms only optionally useful (see Krek et al. 2017 for a more detailed methodology on the scoring and ranking of synonym candidates). From the moment the thesaurus was published, the user community also had the chance to provide additional suggestions for synonyms of any given headword (for more information on user involvement techniques see Arhar Holdt et al. 2018).

For the game, we wanted to use headwords that: (a) have enough synonyms for enjoyable gameplay; (b) are likely to be reliable considering the automated procedures that were used for the preparation of the Thesaurus; (3) are non-problematic for pedagogical use. We also wanted the game to progress in difficulty, as explained in Section 4.2. In the following paragraphs, we describe the decisions made to achieve the listed goals.

First, we arranged the headwords by the frequency of their corresponding synonym candidates. In the frequency count, we included only the automatically acquired core and near synonyms, not also the user-provided synonyms. The result was an ordered list ranging from the headword *hud* ('wild') with 110 synonym candidates to *gnati na vso moč* ('pushing with everything one has'), which is an example among 43,088 headwords with only one synonym candidate. Next, we filtered the list. To begin with, we filtered out 18,165 headwords consisting of three or more words, e.g. *ukvarjati se z;* *postaviti na glavo;* *po drugi strani* ('to attend to; to turn upside down; on the other hand'). We have furthermore eliminated from the frequency count all the synonym candidates with 3 or more words. This filtering step was conducted because it had been determined (Čibej & Arhar Holdt 2019) that due to the methodology features, multi-word synonym candidates include a higher portion of irrelevant material. Additionally, we filtered from the list 117 headwords with only one or two letters, e.g. *da;* *po;* *za* ('that; after; for'), as these comprise solely abbreviations and grammatical words, less suitable for the game. We have also filtered out any headwords that were on the previously described stoplist of derogatory and vulgar words (see Section 3.1), e.g. *pizda;* *peder* ('cunt; queer'). The final filtering condition was that among the remaining synonym candidates, at least five had to be in the core category. In this way, when a player entered three of these synonyms, the game could offer the remaining core synonyms as didactical suggestions (Section 4.2.3).

After filtering, we separated the list into single- and two-word headwords. For the first and main step of the manual selection, we focused on the 5,085 single-word headwords with at least 10 synonyms. In the manual check-up, we

<sup>4</sup> The list is continuously updated for future versions of the game, and other purposes.

<sup>5</sup> The reason why the first version had such a high number of collocations per headword lies in the fact that we included complete lists of collocates without any frequency threshold, so even collocates with a frequency of 1 made it to the list.



eliminated from the list 529 additional headwords that could be potentially problematic for pedagogical use. This step was entirely subjective, its primary goal was to ensure the safe use of the game in the classrooms. If the headword raised any doubt, it was marked for removal. It was interesting to notice that in the case of the synonym module, we were prone to eliminate not only headwords that were vulgar or sensitive, but also headwords that were probing the player for vulgar or sensitive synonyms. Typical examples of that were words alluding to sexual activities, such as *drgniti*; *položiti*; *poriniti* ('to rub; to lay; to push'). Another group that was module-specific were non-derogatory words describing unwanted human features, as these headwords might encourage students to enter (as "synonyms") names of their classmates or similar. Some examples: *lizunka*; *parazit*; *čudak* ('kiss-ass; parasite; weirdo').

From the list of 453 two-word headwords with at least 10 synonyms, 200 were manually selected for the game, all of them verbs with a reflexive pronoun, e.g. *umakniti se*; *obrniti se* ('to remove oneself; to turn oneself'). Compared to other examples of two-word headwords, these demonstrated the most reliable synonym candidates. After this step, the list comprised 4,756 headwords. To reach the desired 5,000, we manually selected the remaining 244 headwords from the single-word headwords with 9 synonyms, following the same criteria for selection as described above. Some examples of headwords that made it to the dataset in this final step were e.g. *dragocen*; *nalepiti*; *čarobno* ('valuable; to stick; magically'). Finally, the 5,000 headwords were separated into 500 sets by 10 according to the number of their synonyms. The sets were manually checked and rearranged to ensure that words from the same word-families or/and with the same meaning were not included in the same set, e.g. *odločen*; *odločno* ('decisive; decisively') where the first headword remained in set 003, while the second was moved to set 005. Arranged headwords were provided to the developers together with core and near synonyms that are used for scoring the player-provided entries (Section 4.2.2).

## 4 Game Modules

In this section, we present both collocation and synonym modules in more detail. It is noteworthy that even though the main focus in developing the mobile app was on gamification and language data, the visualization part of the end product was almost equally important. We explicitly wanted a clear and non-confusing appearance of the application; a design that would not distract the players and would enable them to focus on the content, rather than colors, shapes, or movements. Some initial players' reactions to the visualization of the Game of Words suggest we succeeded in this attempt, yet further user evaluations will be needed to confirm (or discard) this - for now - satisfying response.

### 4.1 Collocation Module

#### 4.1.1 Playing Modes

The collocation module, launched in September 2019, has introduced significant changes to the game compared to version 1. Namely, the old, online version of the game had only one grammatical structure to be completed with collocates (*adjective + noun*), only one game mode to be played (*typing*), and was also quite basic and straightforward regarding scoring of the results. In the mobile version, three different modes of playing are available: TYPE, CHOOSE, and DRAG.

In the Type mode, the format of which was not changed from the online version, players have to complete collocations by typing in three collocates of the given headword, e.g. three adjectives that typically precede a given noun (as shown in Figure 1). One game room of this mode included three headwords.

In the Choose mode, players are presented with three groups of three collocates and have to choose the most typical collocate in the group. Then, for bonus points, they need to arrange their selection according to the (perceived) typicality. Collocations used in the game are selected from three ranges in the list of collocates, one from each range – top 30%, 30-55%, 55-100%. In that way we ensure that for example three collocates next to each other in the list are not selected, and that it is easier to detect the most typical ones. One game room of the Choose mode included three headwords.

In the Drag mode, the players need to drag the collocates to one of the three options: headword A, headword B, and Bin. They are provided with nine randomly ordered words, consisting of three collocates from the list of one headword, three from the list of the other headword, and three distractors (at the moment, taken from completely different headwords and grammatical relations). The headwords compared are picked at random, but belong to the same word class and the same position in a given syntactic structure. Only one headword pair per game room was offered.





Figure 1: Game modes in the collocation module: *Type*; *Choose*; and *Drag*.

All modes are available in the Competitive format, which automatically creates game rooms at regular time intervals. Thus, the players, after picking the mode they want to play, enter the game room running at the time (or wait until the next one starts), play the same words/collocations and compete against each other. The second format used was called Thematic and gave us more control in picking group headwords on a certain common topic (e.g. winter holidays, Christmas, 50-year anniversary of Moon landing). The Thematic format, which was primarily devised to facilitate the promotional activities of the game, was open in a specified time span, each player could play the topic batch only once, and the top players received practical prizes.

#### 4.1.2 Scoring

The numerical score is one of two key feedback pieces of information that language games provide. The non-numerical feedback can be of different nature: the correct answer (if a wrong one is given), a suggestion of another possible response, etc. Due to the project financial restrictions, the collocation module of the Game of Words only provides the numerical score.

When discussing the scoring options, we first and foremost wanted to simplify the scoring system used in the first version as it was too detailed and difficult to comprehend. Furthermore, we paid attention to two issues: *when* in the gaming cycle should the player be presented with the score, and *what would be the best way* to distribute the points. As to *when*, the score is now shown at the end of each headword (pair) - while this extends the total time in the Type and Choose modes, it was thought important that the players have time to inspect their results on individual headwords rather than having a long scrollable list of results at the end.

In terms of point distribution, the scoring system was founded on the collocation salience data; however, the number of collocates per headword and the rating of each collocate were taken into account as well. For the Choose mode, three-point groups were formed based on the collocate ranges. For the Type mode, five scoring groups are used. For the Drag mode, we also used three groups, adding the small “reward bonus” (when a collocate from the list of a particular headword was thrown in the bin) to the correct and incorrect ones. The scoring was devised in a way that enabled comparability across different modes, given that, in addition to having leaderboards for each game mode, we also had the common leaderboard. In order to make the scoring system easy to comprehend, points are also translated into a five-star scale.

## 4.2 Synonym Module

### 4.2.1 Playing Modes

The synonym module of the game was launched in January 2020. The module introduces a solo playing format in which players have to enter three synonyms for a given word. The module is oriented towards didactical purposes and thus offers some learning-oriented features (described below). The didactical angle was included on the request of language teachers who were participating in the project as user evaluators of the Thesaurus of Modern Slovene. In their feedback, it was highlighted that dictionary-based gamification for vocabulary acquisition in Slovene would be extremely valuable, as existing teaching resources for his topic were very scarce and limited in scope.

The game includes 5,000 words, separated into 500 levels with a set of 10 words. The levels progress in difficulty established by the number of existing synonym candidates in the Thesaurus of Modern Slovene. There are two playing modes available, the Game mode (the players progress from one level to another), and the Practice mode (the players can choose the level). For example, level 1 includes headwords like *odličen* (‘excellent’), *divji* (‘wild’) and *uničiti* (‘to destroy’), which all have more than 50 synonyms in the Thesaurus. A more difficult level 300 includes headwords like *posušen* (‘dried’), *poskus* (‘a trial’) and *osvoboditi* (‘to liberate’) that have around 15 synonyms in the Thesaurus. The adjective *odločen* (‘decisive’) presented in Figure 2 appears at level 3 (or 003) as indicated on the top of the screen.



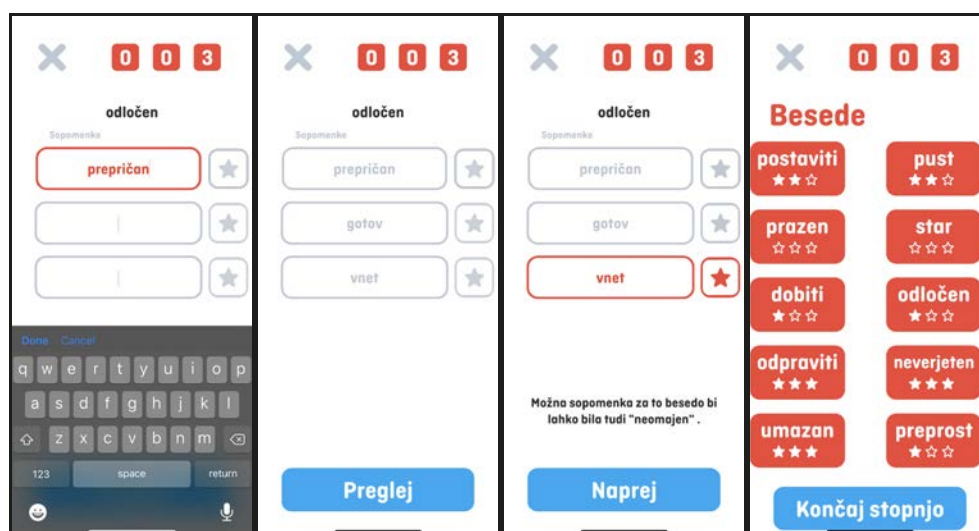


Figure 2: Synonym module: entering synonyms for *odločen* ('decisive') as part of Level 003.

#### 4.2.2 Scoring

The scoring system for the module is very straightforward. After every game, each player-provided synonym that is also found in the Thesaurus database is attributed a star. In this way, players can collect between zero and three stars per game. After a set of 10 games, the stars are transformed into game points, where each star is worth 100 points. The points are then added to the overall score and the player progresses in the Hall of Fame accordingly. For example, player-provided synonyms for *odločen* ('decisive') in Figure 1 are *prepričan* ('positive'), *gotov* ('certain') and *vnet* ('ardent'). For each of them, the player received one star. It is obvious that as long as the Thesaurus database is under development, such scoring is not entirely precise and can cause some frustration among the players. For this reason, we have included a disclaimer explaining that with every upgrade more user-suggested synonyms will be included in the scoring.

#### 4.2.3 Didactic Value

From the didactical point of view, the game facilitates and encourages the use of ICT in the classroom. Both modules of the game are aligned with the curriculum for Slovene as a school subject. The focus of the game is on working with empirical data relevant for specific thematic parts of school curriculum, which enables teachers and students easier transfer of knowledge into practice. In particular, under the guidance of the teacher, the game helps develop metalinguistic competence by teaching the students the concepts of synonyms, collocations, parts of speech (noun, adjective, verb etc.); develop linguistic competence through increased vocabulary and knowledge of syntax; learn to identify synonyms and their semantic/stylistic differences; learn to identify collocations as multiword units; learn to evaluate pros and cons of different types of language resources and to anticipate and identify errors in automatically prepared language resources; learn the importance of openly accessible language data in the digital era and the possibilities of including language community in the creation of openly available language resources.

In comparison to the collocation module, the synonym module has developed special features that support the process of teaching Slovene. For this, the Practice mode, where the player can jump to any given level and play without being scored on the joint leaderboard, is particularly useful and was in fact developed with the pedagogical purpose in mind. Using this mode, teachers have the possibility to find the levels optimally suited for their specific teaching purposes and focus on those in the classroom. Another important feature of the module is that it provides learning material. After the player-provided entries are evaluated, the game shows a possible synonym from the database that was not entered, thus helping the player enrich their vocabulary in Slovene. For example, on the last screen in Figure 2, the game suggested: *Možna sopomenka za to besedo bi bila tudi "neomajen"*. (A synonym for this word might also be 'unwavering'.) At the moment, the suggestions are acquired automatically from the database. In the future, we plan to manually check the suggestions as well as complement them with selected corpus examples to demonstrate their use in context.

## 5 Crowdsourcing Perspectives of the Game of Words

In this section, we discuss crowdsourcing aspects of the Game of Words and its playing modules in more detail. The whole crowdsourcing workflow consists of three stages, namely data preparation, annotation (when the game is played), and data analysis or results implementation. We have already described in detail how both datasets were prepared, but it is important to add that each headword, collocate, and collocation, as well as each headword and its synonyms had to be indexed before being uploaded into the database of the game. The same IDs are then part of the exported data, as this is the only way to ensure valid and quick analysis of the results. One shortcoming in terms of gamifying collocational and synonym data, and any type of linguistic data for that matter, is that sensitive and vulgar content has to be left out, especially if the game also serves pedagogical purposes.

The quality and reliability of data annotation is largely dependent on the playing mode. For example, the Type mode,



which is the most mobile unfriendly and was initially questioned by our designers, is the most reliable mode for crowdsourcing as the players need to enter their answers, whereas in the Choose and Drag modes, they simply (have to) choose between three given options. This is particularly problematic in the Choose mode where there is a chance, albeit a small one, that all three collocates offered are bad ones. The mobile unfriendliness of Type did not seem to bother the players in the synonym module, but it is true there they were not presented with a choice. In the collocation module, however, it was the Drag mode that proved the most popular mode among the players.

The differences in reliability of different modes made us think of what that means for evaluating annotator agreement. For example, how many player decisions are necessary and what needs to be the level of player agreement for a collocation/synonym that we can consider it to be good? We have looked at the data from the first version in attempt to get an answer to this, and we agreed that the acceptable number of annotations would be around 20. Clearly, the number of accepted answers coming from the Type mode could be much lower than at Choose and Drag, which is why we started to devise a scoring system in which a Type “votes” would have a higher annotation value than the Choose and Drag ones.

Another issue closely related to game modes is the crowdsourcing tasks one can do with them. The Type and Choose mode, for example, are much more suitable for the validation of good collocations or synonyms than the identification of bad ones. One can for example consider the never entered or chosen collocates/synonyms, especially those at the top of the list, as potentially bad, but this could still mean a great deal of manual analysis. The Drag mode, however, with the Bin option is also suitable for cleaning the bad collocation candidates.

One crucial matter that is vital for gamification, and which we have perhaps neglected a little bit when preparing both modules, is keeping good and regular control over the data that is being annotated. With that we mean that you cannot import a large dataset into the game, leave it for several months and hope that as much data as possible gets annotated. Let us take collocations, for example. From the lexicographic perspective, everything revolves around headwords, so the best possible method would be to crowdsource all collocations in all the syntactic structures of one headword first, and then move to the next headword. But this approach is not game-friendly as the players would get bored easily. Moreover, in our case, the game rooms for the collocation module are created randomly, and since the dataset is very large, this means that the likelihood of the same headword and collocates being offered more than once are rather low. In fact, we realized that it was the Thematic mode that showed the most potential for crowdsourcing since it was the easiest way to control collocational data, and the best way to motivate large groups of players. The synonym data, on the other hand, is much better controlled, with the only problem being that the order of levels is fixed; considering that the number of players completing a level decreases with each level, lower levels will get annotated more often than higher ones.

As the example of the gaming vs. lexicographic perspective above shows, crowdsourcing purposes (and the data they are related to) and optimal playability often contradict each other. An example of this was our experience when designing different games, as many features that would make the game more attractive, e.g. the first and last letter of the collocate shown, could not be used as the data had to be cleaned first. The entire process of game development became one great balancing act between the lexicographic intentions, numerous possibilities of using language information, and the enjoyment and motivation of playing the game.

## 6 Conclusion and Future Work

The gamification of lexicographic data in Slovenia is still in its infancy; however, the experience gained during the development of the Game of Words is invaluable for our community. The initial evaluations and analyses have shown promising results, but have at the same time pointed out several mistakes in our approach, which we aim to rectify in future versions of the game.

For those interested in developing games of this type, it should be stressed that the development needs to involve a very interdisciplinary team, i.e. not only linguists/lexicographers and computational linguists (for data preparation), but also mathematicians (for scoring etc.), graphic designers (for app design) and developers (for app programming). Crucially, a great deal of continuous proactiveness (i.e. dissemination) after the launch of such an app is required due to a plethora of different types of apps, not only linguistic ones, available on the market.

There is undoubtedly a lot of room for improvement of the game, both on the side of playability and crowdsourcing procedure. For instance, in addition to the already mentioned problem of low control over data annotation, we have found it difficult to get easy access to user logs - as the developers finished their work and moved to other projects, it has been often hard to get a person to export the data in the desired format. This of course means that the ‘responsiveness’ of our analyses, and relatedly dictionaries, has not been as quick as we would have liked.

The future of the Game of Words has become much brighter recently, as the game has attracted the attention of the European Lexicographic Infrastructure (ELEXIS), a Horizon 2020 project, which has one of the activities focused on the development of techniques and tools for crowdsourcing lexicographic data. This resulted in the development of the next version of the game, which will bring the game to other languages (beginning with English, Estonian, and Dutch) and address many crowdsourcing-related shortcomings mentioned in this paper. It will introduce a solo format for collocations, admin tools for easier uploading/downloading of the data and games, and dynamic data selection (e.g. non-annotated collocates will be given priority over already annotated ones). At the same time, the game attracted interest from the researchers involved in the EnetCollect COST Action aimed at connecting crowdsourcing and language learning (Lyding et al. 2018), and there are already plans to develop a module for marking corpus examples that could be potentially problematic for didactic purposes, e.g. due to the presence of sensitive issues or vulgar/derogatory vocabulary (Dekker et al. 2019) - a module that could serve the purposes of both language teachers/learners and lexicographers. Therefore, by widening the community working on the development and dissemination of the game, we can hope that the potential of crowdsourcing in lexicography and language learning can finally be fully exploited.



## 7 References

- Abel, A., Meyer, C. (2013). The dynamics outside the paper: user contributions to online dictionaries. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 179-194.
- Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Gantar, A., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., Robnik Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. Čibej, Jaka et al. (eds.) *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts. 401-410.
- Arhar Holdt, Š. (2020). How Users Responded to a Responsive Dictionary: The Case of the Thesaurus of Modern Slovene. In *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2): in print.
- Benjamin, M. (2015). *Crowdsourcing microdata for cost-effective and reliable lexicography*. No. CONF, pp. 213-221. Accessed at: <https://infoscience.epfl.ch/record/215062> [07/05/2020].
- Čibej, J., Arhar Holdt, Š. (2019). Repel the syntuders! A crowdsourcing cleanup of the thesaurus of modern Slovene. In I. Kosem, S. Krek (eds.) *eLexicography in the 21st century: proceedings of eLex 2019 Conference, 1-3 October 2019, Sintra, Portugal*. Brno: Lexical Computing, 338-356. Accessed at: [https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019\\_Proceedings.pdf](https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019_Proceedings.pdf).
- Dekker, P., Zingano Kuhn, T., Šandrih, B., Zviel-Girshin, R., Arhar Holdt, Š., Schoonheim, T. (2019). Corpus filtering via crowdsourcing for developing a learner's dictionary. In I. Kosem, T. Zingano Kuhn (eds.) *eLexicography in the 21st century (eLex 2019): smart lexicography: book of abstracts*. Brno: Lexical Computing, 84-85. [https://elex.link/elex2019/wp-content/uploads/2019/10/eLex\\_2019-Book\\_of\\_abstracts.pdf](https://elex.link/elex2019/wp-content/uploads/2019/10/eLex_2019-Book_of_abstracts.pdf).
- Dobrovoljc, K., Krek, S., Erjavec, T. (2018). The Sloleks Morphological Lexicon and its Future Development. In V. Gorjanc, P. Gantar, I. Kosem, S. Krek (eds.) *Dictionary of Modern Slovene: problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts, 42-63.
- Fišer, D. and Čibej, J. The potential of crowdsourcing in modern lexicography + Crowdsourcing workflows in lexicography. (2017). In V. Gorjanc et al. (eds.). *Dictionary of Modern Slovene: problems and solutions*, (Book series Prevodoslovje in uporabno jezikoslovje). 1st ed., e-ed. Ljubljana: Ljubljana University Press, Faculty of Arts. Accessed at: [http://www.ff.uni-lj.si/sites/default/files/Dokumenti/Knjige/e-books/dictionary\\_of\\_modern\\_slo.pdf](http://www.ff.uni-lj.si/sites/default/files/Dokumenti/Knjige/e-books/dictionary_of_modern_slo.pdf)
- Gorjanc, V., Gantar, P., Kosem, I., Krek, S. (eds.) (2015). *Dictionary of Modern Slovene: problems and solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts, 2017.
- Guillaume, B., Fort, K., Lefebvre, N. (2016). Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. *Proceedings of the International Conference on Computational Linguistics (COLING)*. Accessed at: <https://hal.inria.fr/hal-01378980/> [07/05/2020].
- Kosem, I., Krek, S., Gantar, P., Arhar Holdt, Š.; Čibej, J., Laskowski, C. (2018a). Collocations dictionary of modern Slovene. In J. Čibej et al. (eds.) *Proceedings of the 18th EURALEX International Congress: lexicography in global contexts*. Ljubljana: Ljubljana University Press, Faculty of Arts, 989-997.
- Kosem, I., Gantar, P., Krek, S., Čibej, J., Arhar Holdt, Š. (2018b). The Good, the Bad and the Noisy? An Analysis of Inter-Annotator Agreement on Collocation Candidates in Different Grammatical Relations. In J. Čibej et al. (eds.) *The XVIII EURALEX International Congress: Lexicography in Global Contexts Book of Abstracts*. Ljubljana: Ljubljana University Press, Faculty of Arts, 71-72.
- Krek, S., Laskowski, C., Robnik Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.), *Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017*, Leiden, Netherlands.
- Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. SNLP'07: 7th International Symposium on Natural Language Processing, Dec 2007. Pattaya, Chonburi, Thailand, 7.
- Lew, R. (2013). User-generated content (UGC) in online English dictionaries. In A. Abel and A. Klosa (eds.) *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess (OPAL – Online publizierte Arbeiten zur Linguistik)*. Mannheim: Institut für Deutsche Sprache, 9-30.
- Logar, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. and Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Lyding, V., Nicolas, L., Bédi, B., Fort, K. (2018). Introducing the European Network for Combining Language Learning and Crowdsourcing Techniques (enetCollect). Future-proof CALL: Language Learning as Exploration and Encounters – Short Papers from EUROCALL 2018, 176.
- Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. and Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems* 3, 3:1–3:44.
- Pori, E., Kosem, I., Čibej, J., Arhar Holdt, Š. (2020). User study: The attitude of dictionary users towards automatically extracted collocation data. In I. Kosem and P. Gantar (eds.) *Slovenščina 2.0*, in print.
- Šorli, M., Grabnar, K., Krek, S., Košir, T. (2006). Oxford-DZS comprehensive English-Slovenian dictionary. In *Proceedings of the XII EURALEX International Congress*. Edizioni dell'Orso: Università di Torino: Academia della Crusca, 631–637.



### Acknowledgements

The research presented in this paper was conducted within two projects titled “The promotion of a language mobile app” and “The Thesaurus of Modern Slovene: By the Community for the Community”, which were financially supported by the Ministry of Culture of the Republic of Slovenia (2018–2019). The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene). The authors also acknowledge the project Collocations as a Basis for Language Description: Semantic and Temporal Perspectives (J6-8255) was financially supported by the Slovenian Research Agency. The ELEXIS part of the research received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015. Many ideas result from the framework of the CA160105 eNetCollect COST Action. We thank all that made our work possible.







# Understanding English Dictionaries: the Experience from a Massive Open Online Course

McGillivray B.<sup>1,2</sup>, Nesi H.<sup>3</sup>, Rundell M.<sup>4</sup>, Süle K.<sup>4</sup>

<sup>1</sup> University of Cambridge, United Kingdom

<sup>2</sup> The Alan Turing Institute, United Kingdom

<sup>3</sup> Coventry University, United Kingdom

<sup>4</sup> Macmillan Education, United Kingdom

## Abstract

We report on the experience from the first Massive Open Online Course (MOOC) dedicated to English dictionaries. The course was created by Coventry University in partnership with The Alan Turing Institute and Macmillan Education, and was provided by the online learning platform FutureLearn. The course ran in late 2019 (with 2477 participants) and again in early 2020 (with 2287 participants). The course relies on a highly interactive approach to knowledge acquisition and consists of articles, videos, interviews, links to further readings, and surveys aimed at stimulating learners' active participation and interaction with the course content and with other participants. In this paper we reflect on our experience of running the course and interacting with participants, and we discuss the results of our quantitative and qualitative analysis of the MOOC. The analysis of the two editions of the course led to very similar results. We found that the majority of learners are female, with a university-level education, and work in the education and teaching sector. The course's participation was comparatively high, and the learners showed a good level of engagement, indicating that there is an interest in accessible courses on lexicographic practice.

**Keywords:** MOOC; massive online open course; data analysis; user data; lexicography; dictionary-making

## 1. The course

This paper reports on the experience from the first Massive Open Online Course (MOOC) dedicated to English dictionaries. The course runs on the *FutureLearn* platform<sup>1</sup> and is a collaboration between Coventry University, The Alan Turing Institute and Macmillan Education. It consists of six week-long units, each providing about four hours of study time. Materials are deliberately kept short (just a few paragraphs for written articles, just a few minutes for video recordings) and every piece of input is followed by at least one interactive task so that participants can construct their own meanings surrounding the information we present. On all *FutureLearn* courses, both qualitative and quantitative records of learner engagement are kept: learners voluntarily supply factual personal information at the start of the course, activity records are collected unobtrusively for each week of study, and we have more open-ended written evidence from learner responses to the tasks, and to the questions in the end-of-course survey.

The Massive Open Online Course (MOOC) concept originated in 2008 at the University of Manitoba, when a course on connectivism with only a few face-to-face students was offered online to everyone, at no charge. It immediately enrolled 1000s of additional learners from all over the world. The idea of offering free online courses was soon picked up by Stanford University, and then MIT. Both these institutions developed their own MOOC companies (*Coursera* and *edX*), while in the UK the Open University developed a rival company, *FutureLearn*. At least 70 million people have now studied with *Coursera*, *edX* or *FutureLearn*, and over 200 universities across the world offer MOOCs (MOOC Lab, 2020). Of these, Delft University of Technology is at the top of the World University Rankings by MOOC Performance (WURMP), followed by the University of Pennsylvania and the University of Illinois. Coventry University was the 4<sup>th</sup> ranked institution on the WURMP 2020 list, and the only institution to offer an online course dedicated to dictionaries.

MOOC providers tend to share an educational philosophy rooted in connectivism, "a learning theory for the digital age" (Siemens 2005). The theory recognizes that in modern times fields of knowledge can expand very rapidly, but that they can also become obsolete very quickly. This means that learning has to be a life-long process, supported by technology and social networks, and that skills and knowledge acquired in the traditional way ("knowing how" and "knowing what") have to be supplemented by an understanding of where new information can be found when it is needed ("knowing where"). This approach to learning suits a dictionary MOOC exceptionally well, as reference works are "where" new knowledge is stored, and as such are an effective means of supporting lifelong learning. True to the spirit of connectivism, many participants on our course were habitual MOOC learners, committed to acquiring "how", "what" and "where" types of knowledge across a variety of fields of study. Our course introduces technology-enhanced information sources that are developing and changing before our very eyes, and participants have opportunities to engage with these resources and at some points even to contribute to their growth.

The MOOC is divided into teaching "Weeks", each covering one of the major themes we address in the course, such as "Why use Dictionaries?" (Week 1), the composition of a dictionary entry (Week 2), the evidence base for dictionaries (Week 3), the inclusion criteria for dictionaries (Week 4), ideas about meaning and definition (Week 5), and the future of

<sup>1</sup> <https://www.futurelearn.com/courses/understanding-dictionaries>



dictionaries (Week 6). Each Week typically begins by asking learners to think about, and share their thoughts on, a fundamental question relating to the theme of the Week. The idea is that they give their own “naive” views on these issues before they have been exposed to any teaching material. This is followed by tasks designed to tease out the complexities of the subject and to demonstrate the challenges lexicographers face – often by requiring learners to work with real language data. The next stage is typically a follow-up video and/or a short article to develop learners’ understanding. Each Week ends with a summary of what has been covered, and participants are asked to reflect on what they have learned – and often to revisit the questions they were asked at the beginning of the Week.

Its first iteration of the MOOC ran from 16 September 2019 to 27 October 2019 and had 2477 enrolled participants. Due to its success, a second iteration of the course started on 20 January 2020, and a third began on 18 May 2020 to run until 26 June 2020.

In Creese et al. (2018), we described the background and motivation of the course, and the design of its very diverse content. The course aimed to introduce the world of English dictionaries to a broad, non-expert audience of language teachers, students and also language enthusiasts. One of our main motivations was to raise awareness of contemporary lexicography, and challenge common misconceptions about dictionary creation. A key feature is the continuous interaction among participants, and between participants and course educators and mentors. These discussions offered us an opportunity to learn a great deal about participants’ expectations and opinions. In this paper, we reflect on our experience of running the course and interacting with participants, and we discuss the results of our quantitative and qualitative analysis of the first two iterations of the MOOC.

## 2. Quantitative analysis

This section reports on the quantitative analysis of the course, for which we focus on three main evidence sources: enrolment and engagement statistics, participants’ demographics, and content from answers to surveys and comments on the platform.

### 2.1 Learners’ participation and engagement

In its first iteration, the MOOC attracted 2477 participants. 1360 (or 55%) of them enrolled after the official start of the course, and only 11 (or 0.4%) of the 2477 participants unenrolled before the start of the course. In its second iteration, the MOOC attracted a slightly higher number of participants (2287), 1454 (or 64%) of them enrolled after the official start of the course, and only 12 (or 0.5%) of the 2287 participants unenrolled before the start of the course. These figures show that, even when they joined late, most learners (2466/2275)<sup>2</sup> stayed in the course until the end in both editions. The rest of our analysis will focus on these groups of participants who stayed in the course until the end. Although there is a wide variation in the length of enrolment (see Figure 1), in both editions of the course on average learners stayed enrolled for 50 and 45 days respectively, which is slightly longer than the duration of the course (42 days) and indicates a high level of interest in the MOOC.

Given the highly interactive nature of the course, we next looked at how active the learners were in participating in the different activities offered. 291 learners (12%) and 198 (9%), respectively in the first and second edition, achieved full participation, meaning that they completed at least 50% of the steps designed for the course. How does this figure compare with other MOOCs? A previous study on MOOC engagement, covering the period 2012-2016 (Chuang and Ho, 2016) reports slightly higher figures for a “typical MOOC”, with 7999 enrolled learners and 1500 achieving full participation. This corresponds to 19% of all participants. However, a more recent study (Reich and Ruipérez-Valiente 2019), covering the years 2017 and 2018, reports a much lower rate, with 3.13% of participants completing their courses. This could indicate a general downward trend in MOOC participation, but it may simply reflect the paucity of evidence, thus far, on users’ engagement with MOOCs. In any case, we can reasonably claim that our course had a good level of participation.

If we further look into the level of engagement, we can see to what extent participants viewed and/or downloaded the videos as the course progressed. As expected, we find a negative significant correlation between the number of views of a video and the video’s position in the course:<sup>3</sup> the later in the course a video appeared, the fewer views and downloads it received. We did not, however, find a statistically significant correlation between the duration of a video and its popularity, measured in terms of number of views or downloads.

<sup>2</sup>In this and all subsequent analyses, the first figure we report refers to the first edition of the course, and the second figure (following “/”) refers to the second edition of the course.

<sup>3</sup> The results of a Spearman correlation test between the number of views of videos and the step position are: correlation coefficient rho = -0.72, p-value < 0.01 for the first edition and correlation coefficient rho = -0.61, p-value < 0.01 for the second edition.



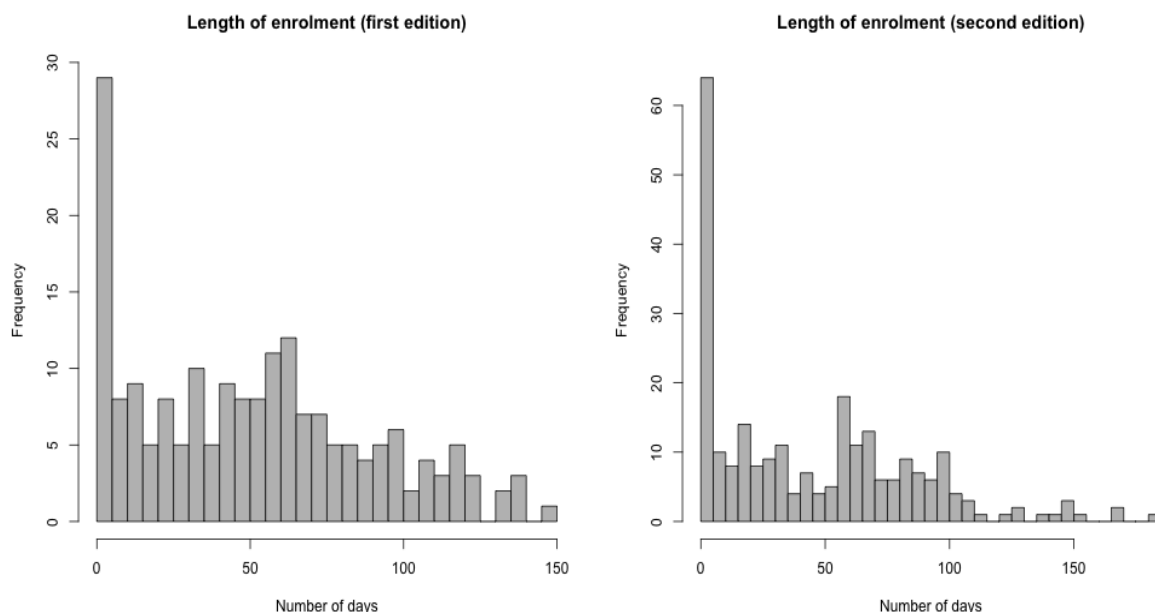


Figure 1: Histogram of the number of days participants stayed enrolled in the first edition (top) and second edition (bottom) of the course. The horizontal axis shows 5-day intervals and the vertical axis shows how many learners stayed enrolled for each of the intervals. For example, in the top figure, the first rectangle corresponds to the interval of 0-5 days and its height shows that 29 learners stayed enrolled for up to 5 days. This analysis is based on information about the dates on which the learners enrolled and unenrolled; the latter dates are available for 189 learners (first edition) and 250 learners (second edition).

## 2.2 Learners' demographics

The demographics of the learners were gathered based on their voluntary responses to questions at the point of signing up to the course, and this information is available for 19% and 16% of the learners (respectively in the first and second editions of the course). These data offer us important insights into the composition of the participants' cohort, their origin, gender, and education level, and allow us to compare them with the expectations we had on the background of the audience when we designed the course. Overall, we found that the majority of learners are female, have a university-level education, and work in the teaching and education sector.

Looking at the gender split, we find that females outnumber males by two to one (67% vs 33%, or 306 vs 150 in the first edition, and 61% vs 38%, or 225 vs 138 in the second edition).<sup>4</sup> This is in contrast with the gender composition of other MOOCs. Chuang and Ho (2016) report a two-to-one male-to-female ratio, but their data is skewed towards computer science courses, which typically attract more male learners.

Regarding the learners' age groups, we have usable information regarding 435 (or 17.6%) of the learners of the first edition and 343 (or 15%) of the learners of the second edition. The data show a relatively even distribution, with an average of 41 years (first edition) and 37 years (second edition), and a median of 36 years for both editions. A comparable proportion of learners (18%/16%) gave information about their education level and employment status, and we have details about the area of employment for 15/13% of the learners. The vast majority have a university degree (82%/79%), most are in employment (60/57%), and a majority (52/43%) are involved in teaching and education (see tables 1 and 2).

<sup>4</sup> It should be noted that in the first edition of the course 2 respondents declared their gender as "non-binary", and one as "Other", and in the second edition of the course 3 respondents responded "non-binary".



First edition			Second edition		
Education level	Count	%	Education level	Count	%
university_degree	194	43	university_degree	154	43
university_masters	148	32	university_masters	100	28
secondary	34	7	secondary	38	10
university_doctorate	32	7	university_doctorate	30	8
tertiary	30	7	tertiary	19	5
professional	13	3	professional	14	4
less_than_secondary	4	1	less_than_secondary	6	2
apprenticeship	1	0			

Table 1: Education level declared by the participants in the first (left) and second (right) edition of the course.

First edition			Second edition		
Employment status	Count	%	Employment status	Count	%
working	264	60	working	203	57
unemployed	82	19	unemployed	84	23
retired	62	14	full_time_student	36	10
full_time_student	32	7	retired	34	10

Table 2: Employment status declared by the participants in the first (left) and second (right) edition of the course.

According to the information provided about the learners' country of origin (available for 67%/58% of the participants), most came from either Europe or Asia (51/25% and 43/31% respectively, see table 3). The most frequent country of origin was Great Britain (19/18% of respondents), which is in line with the over-representation of developed countries found in previous studies (Chuang and Hu 2016). These figures may not be entirely accurate, however, as some Asian participants accessed the course via a virtual private network (VPN) and did not reveal their geographical location.

The preponderance of British participants was probably a reflection of the fact that the course was created in the UK on a British MOOC platform (*FutureLearn*). This may also explain the higher level of active participation observed for participants from Europe<sup>5</sup> and North America: we found that 49/48% and 53/50% (respectively) of all joiners (i.e. those registered for the course and who have given information about their country of origin) from these regions were considered "active learners", meaning that they marked at least one step as complete in the course (see table 3).

Continent	First edition			Second edition		
	Joiners	Active learners	% of active learners	Joiners	Active learners	% of active learners
Africa	79 (5%)	24	30	169 (13%)	34	20
Asia	408 (25%)	137	34	416 (31%)	119	29
Australia	38 (2%)	18	47	25 (2%)	5	20
Europe	834 (51%)	407	49	572 (43%)	274	48
North America	75 (5%)	40	53	38 (3%)	19	50
South America	108 (7%)	49	45	38 (3%)	14	37
unknown	100 (6%)	47	47	73 (5%)	24	33

Table 3: Continent of origin and level of engagement of the participants by continent in the first (left hand side of the table) and second

<sup>5</sup> We only have data about continent provenance for a subset of the course participants. Moreover, the data on active participation at our disposal do not differentiate between the countries within each continent, but we know that participants from Britain constituted a large part of the learners. We conjecture that the high figures of active participation in Europe are to a large extent explained by the participation in the UK. Further analysis would be needed to confirm this.



(right hand side of the table) edition of the course by continent, measured in terms of number of joiners (those who registered in the course), number of active learners (those who marked at least one step as complete in the course), and proportion of active learners (those who marked at least one step as complete in the course) over all learners. This analysis is based on a subset of the full dataset, because continent information is available for 67% (first edition) and 58% (second edition) of all learners.

### 2.3 Learners' responses

As well as being geographically diverse and representing all ages, learners came to the course with a wide range of pre-existing knowledge. While some had backgrounds in linguistics or language teaching, many others belonged to the category we refer to as “language enthusiasts”. To some, concepts such as collocation or the receptive and productive use of dictionaries were entirely unfamiliar. To others, the information we provided about the evidence base of dictionaries was novel and unexpected, often overturning learners' preconceptions both about the lexicographic process and the breadth of information dictionaries contain. One learner began a comment with “Mind blown. I had no idea about...”. For us as educators, these conversations with participants were always interesting and frequently instructive, providing valuable information about what users expect from their dictionaries and how successfully (or not) dictionaries meet users' needs. Learners were encouraged to comment on the course activities and the first edition of the course had twice as many comments as the second edition (7959 vs 4108). There were no specific pedagogical interventions added to the second edition and we did not change the pedagogical approach from the first to the second edition, so this drop may be explained by differences in the cohorts of learners. As expected, in both editions the number of comments per week declined as the course progressed, with 3141/1826 comments in week 1 and 748/276 in week 6. Figure 2 contains a visualisation of the heavily skewed distribution of comments by week in the two editions of the course, and shows that the first week gathered a much higher number of comments compared to the other weeks.

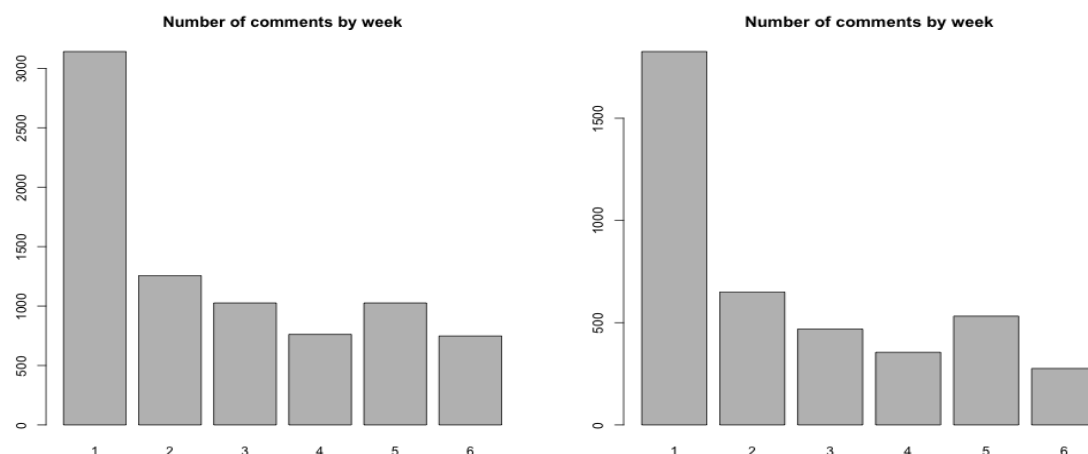


Figure 2: Number of comments per week in the first (top) and second (bottom) edition of the course.

An analysis of the course participants' responses to surveys reveals interesting insights. the majority of the cases (50/37%), the learners who left the course early and responded to the post-course survey offered to them declared that they did not have enough time, while 22/24% recognized that the course did not match their expectations, either because it was too easy, too hard, more time-consuming than expected, or did not match their goals (see table 4). We suspect that in many cases this was due to a misapprehension that the course would provide help with English language learning. Some participants who left comments early in the course had low levels of English and would have found it very difficult to understand the course content.



Reason	First edition		Second edition	
	Count	%	Count	%
I don't have enough time	51	50	45	37
I prefer not to say	10	10	11	9
My access to the course has expired	2	2	23	19
Other	16	16	15	12
The course required more time than I realised	4	4	7	6
The course was too easy	3	3	1	1
The course was too hard	2	2	4	3
The course wasn't what I expected	12	12	12	10
The course won't help me reach my goals	1	1	5	4

Table 4: Results of the survey after leaving the course. The participants were asked to give the reason for leaving the course.

The survey data show a good level of satisfaction with the course among those learners who stayed until the end. Of the learners who completed the course, the overwhelming majority (94/87%) said they acquired new skills or knowledge (see table 5).

Response to the question "Did you acquire new knowledge or skills in the course?"	First edition		Second edition	
	Count	%	Count	%
no response	2	2	1	2
no	3	3	2	3
yes	94	95	54	87
Not sure	0	0	5	8

Table 5: Results of the post-course survey question "to the question "Did you acquire new knowledge or skills in the course?", completed by those learners who completed the course.

Moreover, participants were sent specific surveys aimed at tracking their opinion of the course week by week. Only a small percentage of them (80 or 3% for the first edition and 71 or 3% for the second edition) responded and we cannot take this group as a representative sample of all participants. However, looking at the results of the weekly sentiment surveys for this limited group, we see that their sentiment was generally very positive, with an average of 2.8/2.9, where 1 corresponds to "unhappy" and 3 to "happy" (see figure 3). It should be noted that in the second edition of the course, nobody said they were "unhappy", which is a positive and encouraging result, showing some evidence of an improvement of the course over time.



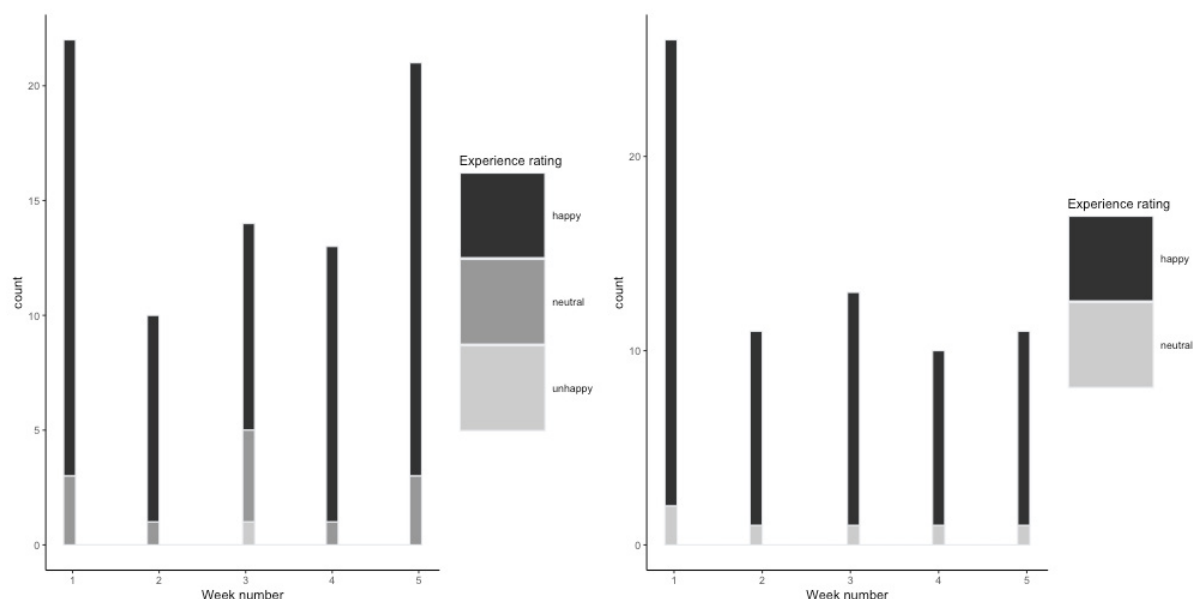


Figure 3: Visualization of the weekly sentiment of the learners in the first (top) and second (bottom) edition of the course. These data are available for 80 (first edition) and 71 (second edition) participants.

### 3. Some qualitative analysis

In this section we will give some examples of how the MOOC format works, and how learners responded to the content, taking examples from Weeks 2 and 5.

Week 2 (“What’s in a Dictionary Entry”) is structured around video clips of genuine dictionary users describing their own experiences and consultation purposes. It starts with a video of three such users trying to make sense of some of the information commonly found in dictionary entries, and asks a question that might be posed by a naive dictionary user: “Why can’t dictionaries just tell us what a word means?” Although the video shows the three users struggling with codes and abbreviations, the majority of the learner responses to this video were in support of the use of grammar pattern codes and IPA, and the ensuing discussion tended to focus on form and denotation, as most learners were not used to considering issues of appropriacy to context, or distinguishing between receptive and productive use. For example a typical comment was “Shouldn’t it be included the type of word, noun, verb, adj, adv etc. and its variations, countable, uncountable, regular verb or irregular, etc.? that’s extremely useful when teaching”, and “A further question of form should be how is this word conjugated or pluralised”. This kind of information is, of course, important for language teachers and learners, but subsequent steps delved more deeply into what is involved in “knowing” a word, and how different aspects of word information might be conveyed in a dictionary entry. Learners were given a Design Task where they could propose additional information that might be added to selected words. Model answers suggested the addition of geographical and register restrictions (*charabanc*, *sidewalk*, *tort*) and notes on connotation or semantic prosody (*skinny*, *utterly*), but also the possibility of including images (for *pelvis*) or a sound file (for *bleat*). Links were provided to online dictionary pages for information about usage labels and word frequency, to *Sketch Engine* for frequency-based word lists, and to Brysbaert et al. (2018) for the concept of word prevalence. This all met with an enthusiastic response from learners, as the concepts were both new (to many) and relevant to their own practice. In keeping with our intention to introduce learners to as broad a range of English dictionaries as possible, learners were invited to examine entries in dictionaries they habitually used, and share and compare their findings. This staged process of reflecting, sharing and exploring illustrates the way the MOOC format can be a particularly powerful means of expanding learners’ horizons, especially in the field of dictionaries where many adult learners start with quite fixed ideas, acquired in their schooldays.

Week 5 (“Meanings and Definitions”) opens by asking learners to give their views on the question: “How do people (people in general – not linguists or lexicographers!) know what words mean?” This is a tough question, and it sparked a lively debate about how we communicate with one another, how we understand what other people say or write. In the first run of the MOOC alone, 79 participants shared their ideas on this topic. Context, real-world knowledge and “repeated exposure in a variety of settings” were mentioned by many, and others reflected on the different experiences of acquiring meanings in our own L1 and when we are learning a second language. One participant commented: “I think of myself as opening a file each time I encounter a new word. My brain adds all the information it can gather from context of use (including who said it and where) and I keep adding information to the file until I feel confident enough to use the word myself” — a process which bears a remarkable resemblance to Michael Hoey’s theory of Lexical Priming (Hoey 2005). Sometimes the discussion would veer into unexpected areas: in Week 5, for example, some participants noted the inbuilt redundancies in many forms of discourse, and this prompted questions about how far AI could successfully reproduce natural language. Throughout the course, the level of engagement with basic questions like these was always high (quantitatively – often



with over 100 responses), and the quality of learners' observations was often impressive and always instructive.

In a follow-up exercise early in Week 5, learners were confronted with some of the issues lexicographers face when creating an entry for a polysemous word. They were shown 10 corpus-derived sentences using the word *overwhelm*, and asked to assign each sentence to one of the numbered senses in a dictionary. For the most part, this was straightforward. But two or three of the sentences, as learners discovered, were more problematic. One of these read: "On their last day they were overwhelmed by farewell messages and gifts". Does this correspond to sense 1 in the dictionary, which emphasizes someone's emotional response ("to affect someone's emotions in a very powerful way")? Or to sense 2, which focuses on the notion of "overwhelming" quantities ("to exist in such great amounts that someone or something cannot deal with them")? As one learner observed "I sometimes had the feeling that a sentence could be reasonably assigned to more than one sense". In this way, participants came to recognise that dictionary senses are not set in stone (as many might have thought), but are to a degree a lexicographic construct which simply aims to provide some useful generalisations about the different ways this verb contributes to the meaning of an utterance. As many of them said, they had no difficulty understanding any of the sentences, so there was no ambiguity at that level. But mapping each sentence to a specific dictionary sense in the dictionary was not always so simple.

This message is then supported in a video interview with Patrick Hanks (which proved very popular: "I love the concept of meaning potentials!" said one learner) and by further tasks and a short article summarising the issues. From our point of view as educators on the course, it was always rewarding to see ideas like these – very familiar to those of us working in the field – gradually dawning on learners who had, in most cases, never given the matter much thought.

In our earlier paper (Creese et al. (2018)), we speculated that the content of the course "may challenge some participants' long-held views about the authority of dictionaries", and might correct popular misconceptions about how dictionaries are created. In this context, it is interesting to compare learners' responses at the beginning and end of each teaching Week – and indeed at the beginning and end of the course as a whole. Week 3, for example, begins with a video addressing the question "Where does the information in dictionaries come from?". This attracted over 100 comments in the first iteration alone. A minority of respondents had some awareness of corpora and their role in lexicography, but most had no clear answers to the question, giving the impression that this was something they had never really thought about. A recurrent assumption was that new dictionaries were mostly based on older ones, with comments such as:

"They probably get their info from past dictionaries, and update every year with a new batch of words", "They build on the previous knowledge and entries in an existing dictionary", "There is obviously a huge database from existing dictionaries but I have no idea how dictionaries originated 'once upon a time'".

Many responses were even more vague, with one learner suggesting that dictionary entries are "possibly sourced from the inputs of professionals like professors, lawyers, scientists, authors and journalists".

Week 3 then continues, through a mix of articles, tasks, videos, and other learning materials, in which different forms of linguistic evidence (introspection, citations, and corpora) are introduced, and learners have an opportunity to work with corpus data in the form of concordances and Word Sketches. At the end of Week 3 learners are asked to share their reflections on what they have learned. Again, this prompted a high number of comments, most of which demonstrated that many earlier assumptions had been overturned. Two typical responses were: "I understand dictionaries substantially better now than before. Exposure to corpus analysis was a wonderful experience... look forward to delving deeper."

This Week "changed my view not only on dictionaries but also on language as a whole."

#### 4. Conclusion

Overall, the course provided valuable insights for lexicography researchers and practitioners, revealing the expectations and the topics of interest of an educated non-expert audience. This is succinctly summarized by one of the course participants, who commented:

I have never imagined there was so much work behind dictionaries! I have learned about different dictionaries, different uses, the parts of a dictionary, the latest technology, their future... Now I see things more clearly and I appreciate dictionaries much more. It has been a fantastic trip.<sup>6</sup>

As we said at the beginning of this article, the course is currently in its third iteration and we will undoubtedly gain further insights into the participants' interest and attitude towards the world of lexicography. At the time of writing, MOOCs are experiencing a drastic increase in their popularity, in a time when a large part of the world's population is forced to stay at home due to the Covid-19 pandemic,<sup>7</sup> and it is reasonable to expect this trend to continue. With this analysis, we have provided the lexicographic community with a new source of evidence of public attitudes towards dictionaries, and we hope that this will be of inspiration in the design and maintenance of current and future dictionaries.

#### 5. References

- Brysbaert, M., Mander, P., McCormick, S.F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. In *Behavior Research Methods* 51. 467-479, pp. 1-13.
- Chuang, I. & Ho, A. (2016). HarvardX and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016 (December 23, 2016). Accessed at: <http://dx.doi.org/10.2139/ssrn.2889436> [29/05/2020].
- Creese, S., McGillivray, B., Nesi, H., Rundell, M. & Sule, K. (2018). Everything You Always Wanted to Know about Dictionaries (But were Afraid to Ask): A Massive Open Online Course. In Čibej, J., Gorjanc, V., Kosem, I., Krek, S.

<sup>6</sup> This comment is reprinted with permission.

<sup>7</sup> <https://www.classcentral.com/report/moocwatch-23-moocs-back-in-the-spotlight> (last accessed 29/05/2020).



- (eds.) *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018, Ljubljana, Slovenia*, 59-66.
- Hoey, M. (2005). *Lexical Priming – A new theory of words and language*. London: Routledge.
- Reich, J. & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science* 383(6423), pp. 130-131.
- Siemens, G. (2005). Connectivism: A learning theory for the digital age. In *International Journal of Instructional Technology and Distance Learning* 2(1), 4-13.

### Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicography and Language Technologies**







# Towards Automatic Linking of Lexicographic Data: the case of a Historical and a Modern Danish Dictionary

Ahmadi S.<sup>1</sup>, Nimb S.<sup>2</sup>, McCrae J.<sup>1</sup>, Sørensen N.<sup>2</sup>

<sup>1</sup> Insight Centre for Data Analytics, National University of Ireland, Galway

<sup>2</sup> Society for Danish Language and Literature (DSL), Copenhagen, Denmark

## Abstract

Given the diversity of lexical-semantic resources, particularly dictionaries, integrating such resources by aligning various types of information is an important task, both in e-lexicography and natural language processing. The current study aims at analyzing the automatic alignment of word senses of the same lemmas across two comprehensive monolingual Danish dictionaries, the historic *Ordbog over det danske Sprog* and the modern *Den Danske Ordbog*. We report our efforts in creating a gold-standard dataset and show that semantic similarity measures can be efficiently used to create statistical models to automatically align senses across dictionaries.

**Keywords:** semantic similarity detection; dictionary linking; natural language processing; e-lexicography

## 1 Introduction

During the past decades, there have been many efforts in applying natural language processing (NLP) techniques for word sense alignment (WSA) where senses of identical words are aligned across various lexical resources. This task is proved to be beneficial to many applications such as semantic role labelling (Palmer 2009) and information extraction (Moro et al. 2013). But also within e-lexicography and the work of publishing a series of online monolingual dictionaries is the alignment of senses of identical lemmas very relevant, opening up for new ways of presenting word information to the users of the resources.

The Society for Danish Language and Literature (DSL) has been publishing scholarly edited Danish dictionaries for more than 100 years, since 2005 also in the form of online dictionaries. Two of these, the modern dictionary *Den Danske Ordbog*<sup>1</sup> (“The Danish Dictionary”, henceforth DDO) covering the senses of more than 100,000 Danish lemmas from 1955 till today, and the historic, retro-digitized dictionary *Ordbog over det danske Sprog*<sup>2</sup> (“Dictionary of the Danish Language” henceforth ODS), covering 220,000 Danish lemmas from 1700 till 1955, are in the Danish society considered to be key lexical resources of the Danish language. Both are available at the same public dictionary site of DSL, *ordnet.dk* which has more than 100,000 daily users. At the site, the lemmas in the two dictionaries are connected at a string-based level (by implying exact string similarity), allowing for hits across the two resources when a word is looked up by a user. In the XML structure, ODS has by the use of semi-automatic methods been supplied with links at lemma level to a number of DSL’s retro-digitized historic dictionaries covering Danish before 1700 (Svendsen et al. 2020). Opposite to this, DSL’s lexical resources for modern Danish are all linked not only at lemma level, but also at sense level. This means that DDO shares sense ID numbers with not only a Danish thesaurus, *Den Danske Begrebsordbog* (Nimb et al. 2014), but also the Danish WordNet DanNet (Pedersen et al. 2009) as well as the Danish FrameNet lexicon (Nimb 2018), see Pedersen et al. (2018). The semantic linking between DDO and the thesaurus constitutes the basis of the compilation of the FrameNet lexicon (Nimb et al. 2017) as well as the presentation of groups of near-synonyms and thematically related words from the thesaurus in the online DDO (Nimb et al. 2018). However, the important but challenging task of linking the modern resources to the historic dictionaries still remains to be carried out. The future online publishing of digital dictionaries at *ordnet.dk* will to a very high degree benefit from such links, opening up for new ways to present the elder vocabulary to dictionary users. For example functions from the modern dictionary like “Ord i nærheden” could be easily transferred to the online ODS and thereby give new insights into the Danish vocabulary and conceptualization in older times.

Linking identical lemmas at sense level in the two key dictionaries DDO and ODS is the obvious first step to take. Once a method is developed for lemmas where we know for sure that there are sense matches between the dictionaries, the method can be applied on the rest of the vocabulary, and senses can be matched across lemmas, not only between identical lemmas. DDO and ODS are similar in many ways since they are compiled by the same institution, DSL, and since DDO from the beginning was planned to be a modern follow-up to ODS. The DDO project was initiated approx. 40 years after the last volume of ODS had been published, and there is an overlap in the different language periods of Danish that they describe: the middle period of the 20th century. Furthermore, DDO is to a high degree inspired by the lexicographic style that ODS had already established for dictionaries being compiled at DSL, both w.r.t. the method, structure and content. ODS has been edited on the basis of approx. 2.5 million manually collected sentences with precise source citation, DDO on the basis of a 40 million text corpus, and both dictionaries use authentic language examples as sense documentation.

In this paper, we will discuss how natural language processing techniques can be applied in the task of aligning the senses of identical lemmas in the two dictionaries, a task which otherwise would be a time-consuming and difficult challenge due to the high number of senses in both of them. The main objective of the study is to evaluate the performance of various automatic methods to carry out the linking, such as string similarity measures and word embeddings. As a preliminary study of its kind for ODS and DDO, we define our alignment task as detecting sense candidate pairs within a combination of all senses for two identical lemmas in two resources.

The rest of the paper is organized as follows. We first describe our lexicographic data, their similarities as well as their dissimilarities in Section 2. Then, in Section 3, we present our methodology where the preparation of the data, the manual

<sup>1</sup> <https://ordnet.dk/ddo>

<sup>2</sup> <https://ordnet.dk/ods>



annotation, and the models are introduced. Section 4 provides the results of our experiments indicating how sense length and various experimental setups change the performance of the alignment task. Finally, the paper concludes in Section 5 where we mention a few future directions in the same field.

## 2 Lexicographic Data

A monolingual dictionary can be considered as a knowledge repository which provides description of the vocabulary of a language with various information, particularly senses. Although two monolingual dictionaries such as DDO and ODS describe the same distinctive and unique ideas in the same language within a certain time period, the way they do it may be very different. The differences are mainly observed as follows:

**Sense structure:** senses in comprehensive dictionaries are typically organized in a hierarchy where semantically related concepts are provided as subsenses to a main sense. However, the sense granularity and the exact distinctions drawn between both main senses and subsenses of a lemma might differ quite a lot across monolingual dictionaries. Closely related concepts, e.g. the many cases of regular polysemy in the language (see among others Buitelaar 2000; Pustejovsky 1998), might be expressed as separate subsenses, but might as well be (indirectly) included in the main senses. This varies not only across dictionaries, but also within the same dictionary. Furthermore, the sense granularity of a dictionary is influenced by the specific editorial guidelines, according to for example the space available in printed versions, however also by the more subjective and individual judgments made by each lexicographer as stated by Kilgariff (2003: 372): “any working lexicographer is well aware that, every day, they are making decisions on whether to ‘lump’ or ‘split’ senses that are inevitably subjective”.

**Definition content:** The description style decided upon by the two dictionaries, as well as the lexicographer’s individual description style, focus on meaning aspect and lexical word choice, may vary quite a lot. When the two dictionaries are compiled in different time periods, such differences become even more significant. In this case, spelling variations over time in the language might also be a factor that must be taken into consideration.

If we compare our two dictionaries, ODS is first of all a historical dictionary covering Danish from 1700- ~1950, where DDO is a modern dictionary covering Danish from around 1950. Its main focus is on the years after 1982 where the first corpus texts that it builds upon date from, see Lorentzen (2004).

ODS was published in 27 volumes describing 188,000 lemmas in the years 1918 to 1954, with a later addition of 5 supplementary volumes with 35,000 lemmas, published 1992-2005. It contains far more dialectal language than DDO, both at lemma and sense level.

DDO was edited in a much shorter period (1994-2003), and published in far less volumes, namely 6, in the years 2003-2005, at that time describing the senses of 66,000 lemmas. Today, the online version describes the senses of 100,000 lemmas. The dictionary focuses on general language, both w.r.t. lemma selection and sense descriptions. DDO still only covers half as many lemmas as ODS. Also at sense level, ODS is more extensive. Since ODS describes Danish in a 250-year period, and DDO only in a 50-year period, ODS covers far more historic senses per lemma.

The figure displays two side-by-side dictionary entries for the Danish noun 'afstand'. The left entry is from the Dansk Ordbog (DDO) and the right entry is from the Ordbog over Det Danske Sprog (ODS).

**DDO Entry (Left):**

- afstand** substantiv, fælleskøn
- Vis oversigt**
- Betydninger**
- 1. rumlig udstrækning der adskiller to punkter, linjer eller flader, målt som længden af en linje eller rute mellem dem**
  - SYNONYMER** distance | sjældent frastand
  - ORD I NÆRHEDED** hul | gab | spring | plads | tom plads | ledig plads...vis mere
  - GRAMMATIK** afstand mellem NOGET/NOGEN og NOGET/NOGEN | afstand af/på NÅL | afstand fra/til NOGET
  - EKSEMPLER** indbyrdes afstand | den geografiske afstand | bedømme/måle afstanden | på lang afstand | i/på behørig afstand | i/på passende afstand | i/på sikker afstand
  - VI passerede skibene i en afstand af ca. 40 meter Hvidov1989**
  - Brevduer kan finde hjem over lange afstande skoleb-fys.91**
- 1.a. tidsmæssig udstrækning der adskiller to begivenheder**
  - ORD I NÆRHEDED** tidsrum | tidsinterval | interval | tidsafstand | tidsmargin...vis mere
  - han vidste at såret til trods for de tyve års afstand endnu ikke var lægt SvMads99**
- 1.b. OVERFØRT mangel på fortrolighed, kontakt eller personligt engagement**
  - SYNONYMER** distance
  - ORD I NÆRHEDED** modsætningsforhold | delte meninger | påstand mod påstand | en strid om ord | ordstrid | uforenelighed...vis mere
  - GRAMMATIK** især i singularis
  - EKSEMPLER** en vis afstand
  - brugen af fagsprog skaber en afstand mellem læge og patient fags-psyk.92a**
- 1.c. OVERFØRT forskel eller modsætningsforhold mellem to parter eller størrelser**
  - ORD I NÆRHEDED** forskellighed | diversitet | skisma | kluft | græft | gab...vis mere
  - afstanden mellem regeringen og oppositionen er mindsket BT1991**
  - Denne afstand mellem løfter og realiteter kalder jeg svindel BerTT1991**

**ODS Entry (Right):**

- Afstand**, en. [ˈau sdan] f. -e [ˈau- sdaːn] (efter ty. afstand (jf. lat. distantia); Moth(8740) har ordet som vbs. til afstaa: "fravigelse. Recensius"; ellers bruges det først ved midten af 18. aarh.; i stedet findes udtr. som Afviggenhed, Distance, Fraviggenhed, Frastand ofl.)
- 1) fjerenhed; længden af mellemrummet (mat.: af en ret linje) mellem to punkter, udfinde Solens Afstand fra Jorden. Heitm. Physik.67. (jf. Steners. Crit.Bet.29 og Mary. Klopstock.Breve.(1760).40). Safarende. have ofte stor Færdighed i at bedømme Afstandene. Heib.Prog.II.369. \*Seer jeg . . en Hatfuld sydet Damp i Tidens Maal, Rummets Afstande flytte. Ploug.VI.13. Afstanden fra Kærsholm til Bostrup Præstegaard var femseks Kilometer. Pont.LP. VII.65. (s.) han vandt Afstand (dus.: kom længere og længere bort) fra (de angribende vilde heste). Rist.FT.28. || X spec. om regelmæssige mellemrum mellem (afdelinger af) soldater, som staar bag ved hinanden. Sal. IX.531. jf.: Der er Gæssene . . med Retning og Afstand som en Trup Soldater. Bogan.I.127. det er daarlæg ridning; her er ikke spor af afstand (dus.: der er ulige stor afstand mellem de enkelte ryttere) || efter præp. i. \*Alt, hvad Naturen . . i Maalles Afstand fra hinanden spredte. Baggens.L. I.156. Munken . . holder sig i en erbedig Afstand. Oehl.IV.161. Medens vi talte, saa jeg i lang Afstand en Dame komme. Goldschm.VI.274. i afstand (ell. † i en afstand. Gylb.III.215. IV.280.331. VIII.223). (s.) ikke tæt ved ell. paa nært hold; (temmelig) langt borte. (nu oftere paa afstand). \*Jeg vendte om, og som en ydmyg Slave | i Afstand troe jeg fulgte Deres Vel. Heib. Poet.VII.272. Vandet saae klart ud nær ved, men seet i Afstand, sort som Blæk. HCAnd.VI.300. Jeg elsker Franskmandene – i Afstand. Goldschm.I.354. De dræbte ham i Afstand med Pile. smst.III. 197. i efter præp. paa. \*Et een af Tillys Mænd er under Vaaben | Paa mange Miles Afstand. Hauch.E.70. Der skulde skydes (dus.: ved en duel) paa 15 Skridts Afstand. JakKnu.A.211. paa afstand, d. s. s. i afstand. Hun havde ogsaa noget Godt i sine Øjne, naar hun var lidt paa Afstand. Schand.BS.118. Aftenklokken . . lod saa smukt paa Afstand. Pont.LP.VII.89. Enhver Voksen der har den mindste Katar bør holde sig paa Afstand fra Bernene. Sundhedstid.1916.266.**

Figure 1: The noun *afstand* (“distance”) in the two Danish monolingual dictionaries, DDO (left) and ODS (right).

We know for a fact that there is an overlap in the lexicographic content of DDO and ODS, both at lemma and sense level. The editors of DDO were generally advised to consult ODS when establishing the DDO descriptions of identical lemmas, since many of the modern senses were already registered and described in ODS, see an example, the noun *afstand* (“distance”), in figure 1, where the first senses are similar. From our studies and the extraction of datasets (see below), we also know that around 86% of the central lemmas in modern Danish are included in ODS (Pedersen et al. 2019).

If we look further into the entries of the two dictionaries that we want to link at sense level, we find many resemblances between them, but also many differences.

**Sense structure:** Both make use of a hierarchical structure with main senses and subsenses, however in different ways.



The order of main senses as well as subsenses is in ODS based on etymology but in DDO on corpus frequency. DDO establishes only main senses proved by concrete textual examples, before the closely related senses to it are listed in the form of subsenses. These might represent either a broader, a narrower, or a figurative use to a higher degree of the main sense, and also have to be manifested in concrete examples in the language. Opposite to this, ODS operates with “main” senses in the structure which are in fact rather a kind of heading or very broad “summing up” sense description for a series of subsenses to be listed, which are then the only ones to be manifested in concrete language. This has of course the consequence that very often two senses in ODS, namely both the heading “main” sense and one of its subsenses are semantically related to the same one sense in DDO. So, in this case, ODS splits in more senses than DDO does. When it comes to sense granularity, they also differ in other ways, since ODS often “lumps” content that DDO would instead express as several senses in the structure, by using formulations like: “også om” (“also about”), “dels .., dels” (“both .. and”), “også uegentl” (“also figurative”) in one and the same definition. So, in these cases DDO splits in more senses than ODS does. Furthermore, the difference in size might have influenced the sense granularities of the two dictionaries. The DDO editors were often encouraged to rather “lump” the senses of the less frequent lemmas due to the limited space in the printed edition, e.g. in the cases of regular polysemy. The editors of ODS had less restrictions on space, which might have had as consequence that they splitted senses more often.

**Definition content:** The time span between the edition of the first volumes of ODS and the most recent edition of lemmas in DDO is 100 years. This leads to many differences in lexicographic description style. The definition style of ODS is very compact, aiming at presenting as many details as possible in one and the same phrase. The editors of DDO focused instead on the communicative qualities of a definition. Where ODS uses many parentheses, additional words and phrases and a deep syntactic structure with many attributives and subordinate phrases in order to try to cover all aspects of a sense, DDO focuses on the prototypical aspects and prefers a more flat syntactic structure (see figure 2, the verb *lukke* (“to close”), and figure 3, the noun *standpunkt* (“view”) for examples). When DDO makes use of supplementary explanations, these are easily identified automatically, always being initiated by a semicolon in the definition text, or being placed in two separate XML-fields, one for connotative, one for encyclopedic information. These fields are not a part of the extracted data to be linked.

	Danish definition	English translation
verb <i>lukke</i> (‘to close’)	<u>ODS</u> : <i>trække, lægge, skyde hen for (over) en aabning, saaledes at denne spærres, udfyldes, tilstoppes; især m. h. t. et dertil beregnet og anbragt (i aabningen passende) spærremiddel, fx. klap, lem, dør; m. h. t. dør olgn. ogs. undertiden: (trække til og) laase ell. stænge</i> <u>DDO</u> : <i>bevæge noget dertil indrettet hen foran eller hen over en åbning så den spærres</i>	<u>ODS</u> : “pull, place, shoot over (over) an opening so that it is blocked, filled out, clogged; in particular w.r.t. a specially designed and arranged (in the aperture) blocking means, e.g. a clap, limb, door; w.r.t. doors or the like also sometimes: (pull and) lock or close” <u>DDO</u> : “move something to the front of or across an opening to lock it”

Figure 2: The verb *lukke* (“to close”) descriptions of the same sense in DDO and ODS differ to a high degree w.r.t. syntax and description style. Where DDO focuses on the prototypical type of closing something, ODS tries to cover all possible ways of doing it, with all types of objects.

Also when it comes to the content of the definition text, we find many differences between the two dictionaries, either due to the time span between the edition of the two dictionaries, or simply to the lexicographer's individual choices in each case. See figure 3 for an example (the lemma *standpunkt* (“view”)) where there is no word at all in common between the two definitions, even though they convey the same meaning.

	Danish definitions in ODS and DDO	English translations
noun <i>standpunkt</i> (‘view’)	<u>ODS</u> : <i>om en persons åndelige stade som forudsætning ell. baggrund for hans anskuelser, synsmåde ell. handlemåde; synspunkt; ogs. om den anskuelse, hvortil man er kommet, det grundsyn, man anlægger på noget, ell. (i videre anv.) om stadium ell. trin i en persons åndelige ell. sociale udvikling ell. i en sags, et forholds udvikling olgn.</i> <u>DDO</u> : <i>opfattelse af og holdning til et bestemt spørgsmål el. anliggende</i>	<u>ODS</u> : “about a person's spiritual state as a prerequisite or background to his views, mode of view or mode of action; point of view. about the view to which one has come, the basic view that one is applying to something, or (further use) about the stage or step of a person's spiritual or social development or the development of a case, a relationship, etc.” <u>DDO</u> : “perception of and attitude to a particular issue or matter”

Figure 3: Different word choice: The two definitions of the noun *standpunkt* (“view”) in ODS and DDO describe the same sense but have no lexical content words in common. The ODS definition is furthermore an example of the complicated definition style of the dictionary.



Figure 4 illustrates that definitions might also focus on different aspects of word meaning, i.e. different qualia roles (Pustejovsky 1995: 76). In ODS, “honey” is described by focusing on how it is produced, the AGENTIVE role: “factors involved in its origin or bringing it about”, in DDO mainly by focusing on how it is used, the TELIC role: “its purpose and function” having as consequence that the resulting definitions become very different.

	Danish definitions in ODS and DDO	English translations
noun <i>honning</i> (‘honey’)	<u>ODS</u> : <i>plantesaft, der er opsuget af bier, omdannet i deres tarmkanal og atter gylpet op</i> <u>DDO</u> : <i>sød klæbrig masse som bier danner af blomsters nektar, og som fx spises på brød eller bruges som ingrediens i mad</i>	<u>ODS</u> : “sap/plant juice which is soaked up by bees, transformed in their intestinal tracts and regurgitated” <u>DDO</u> : “sweet sticky mass that bees form from the nectar of flowers and which for example is eaten on bread or used as an ingredient in food”

Figure 4: Different meaning aspects: In ODS, “honey” is described with the focus on the biological process behind where DDO instead focuses on the resulting food and how it is consumed.

But we also find many quite parallel definitions in the two dictionaries, both w.r.t. syntactic style and lexical choice (however, the lemmas may be in different morphological forms), see Figure 5 for examples.

	Danish definitions in ODS and DDO	English translations
noun <i>klemme</i> , (‘trouble’) - (ODS sense 2)	<u>ODS</u> : <i><b>knibe</b>, forlegenhed; <b>vanskelig situation</b></i> ; <u>DDO</u> : <i><b>vanskelig situation</b>; knibe</i>	<u>ODS</u> : “trouble, embarrassment; difficult situation;” <u>DDO</u> : “difficult situation; trouble”
noun <i>klemme</i> (‘sandwich’) (ODS sense 6)	<u>ODS</u> : <i><b>tykt</b> (og mindre lækkert) <b>stykke smørrebrød</b> (især om <b>sammenlagte</b> (egl.: sammenklemt? ell. mindende om en (tøj)klemmes to led ell. flader?) <b>stykker smørrebrød</b>, der <b>medbringes</b> til arbejdsstedet)</i> <u>DDO</u> : <i><b>tykt stykke smørrebrød</b>; to skiver brød som er <b>lagt sammen</b> omkring et <b>stykke</b> pålæg og <b>medbragt</b> i en madpakke</i>	<u>ODS</u> : “thick (and less delicious) piece of sandwich (especially about combined (maybe in fact squeezed or reminiscent of a clothespin's two joints or surfaces?) pieces of sandwiches brought to the workplace)” <u>DDO</u> : “thick piece of sandwich; two slices of bread that are put together around a piece of topping and brought in a packed lunch”
verb <i>sikre</i> (‘to secure’) (ODS sense 1.1)	<u>ODS</u> : <i><b>beskytte</b> en ell. noget <b>mod angreb</b>, skade, <b>overlast</b>, forstyrrelse olgn. v. <b>hj. af forebyggende foranstaltninger</b></i> <u>DDO</u> : <i><b>beskytte mod angreb, overlast, forringelser e.l. vha. forebyggende foranstaltninger</b></i>	<u>ODS</u> : “protect somebody or something from attack, injury, nuisance, disruption, etc. using preventative measures”→ <u>DDO</u> : “protect against attack, nuisance, deterioration etc. using preventive measures”
noun <i>middag</i> (‘noon’), ODS sense 1.	<u>ODS</u> : <i>det <b>tidspunkt midt på dagen</b> (kl. 12), da solen står højest på himlen</i> <u>DDO</u> : <i><b>tidspunktet midt på dagen hvor solen står højest på himlen</b> (ca. mellem kl. 11 og 13)</i>	<u>ODS</u> : “that time in the middle of the day (12 noon) when the sun is highest in the sky” <u>DDO</u> : “the time in the middle of the day when the sun is highest in the sky (approximately between 11am and 1pm)”
noun <i>søvn</i> (‘sleep’), ODS sense 3.	<u>ODS</u> : <i><b>materie</b> (pus), <b>afsondret i øjet</b> (<b>øjenkrogen</b>) under søvnen</i> <u>DDO</u> : <i><b>materie som afsondres i øjenkrogene mens man sover</b></i>	<u>ODS</u> : “matter (pus), secreted in the eye (eye hook) during sleep” <u>DDO</u> : “matter that is secreted in the corners of the eye while sleeping”

Figure 5: Resemblances in lexical choice: Examples of definitions in ODS and DDO where the two dictionaries make use of identical words or lemmas (in bold), and even identical phrases, to describe the same sense.

An important difference between the two dictionaries is that ODS in some cases presents metainformation in the form of precise sense references (numbers) for words in the definition text itself, typically when it consists of only a synonym or when the lemma is a derivation (e.g. a number reference from a verbal noun to the relevant verb sense). This we never find in the isolated definition data from DDO. We also very often find other types of metainformation inside the ODS definition



text, for example the lexicographers' guess regarding the etymology of the sense of the noun *klemme* ("sandwich") as seen in figure 5. Another very big difference is that ODS sometimes have no definition text at all to a sense in the structure, only examples.

Finally there are some divergences in the orthography of two dictionaries due to a Danish language spelling reform in 1948 where for example the letters "aa" were replaced by a new letter "å". See examples in figure 2: *aabning* → *åbning*, *laase* → *låse*. Many abbreviations are also spelled differently in the two dictionaries: ODS *p. gr. af* (*på grund af* "because/du to") → DDO: *pga.*, ODS *ell. (eller "or")* → DDO: *el.*, ODS: *ogs.(også ("also"))* → DDO *også*, etc. The structure and content of ODS are described (in Danish) in a number of texts at <https://ordnet.dk/ods/>, see for example Jacobsen & Juhl-Jensen (1918).

**Differences in XML structure:** Our task considered, it is important to mention that the dictionaries to a very high degree differ when it comes to the number of markups in the XML structure. DDO was from the very beginning edited in a fine-grained XML structure with isolated content-named elements, e.g. one for the definition, another for the citation etc., constituting the perfect basis for the later online edition. Opposite to this, ODS has been retrodigitized based on the printed version in order to be published online in 2005, and is still in the process of being transformed into a well-defined XML structure. With regard to the semantic part, only the full sense content including citations, etc. has so far been identified automatically in the established digital manuscript, not the exact part of the sense description which constitutes the definition phrase that would be ideal to be compared to the definition phrase of DDO in our task. The definition text from ODS is often initiated by different types of metainformation on for example frequency, chronology, domain, as well as use, which is not part of the DDO definition text that it is compared with. Furthermore, metainformation can even be part of the definition text itself, as described above.

To sum up, in many cases of identical lemmas the two dictionaries differ quite substantially when it comes to structure as well as content, furthermore the extracted ODS definition text used as input for our task is very noisy compared to the extract from DDO.

### 3 Methodology

#### 3.1 Data Preparation

Based on our knowledge of the many differences between the two dictionaries regarding both structure and content, the datasets for linking task are created following these steps:

- Extracting identical lemmas in DDO and ODS: After normalizing the spelling variations, we extract lemmas with identical spelling and with subsequent manual corrections.
- Extracting senses in ODS and DDO: This was a challenging process as different reference keys that are used for senses were dealt with differently. Due to the complexity in extracting senses, we did not take multi-word expressions into account in the extraction process.
- Normalizing orthographies: In ODS, an old Danish orthography is used, as formerly mentioned. We automatically converted that orthography to the modern one using a mapping between characters. The mapping consists simply of 24 mappings like "kjø → kø" and was constructed by philologists at DSL working on 19th century Danish literature (see for example Bjerring-Hansen et al. (2019)).
- Summarizing senses: As described above, the ODS sense descriptions are often very detailed and syntactically complex (see figures 2 and 3 for examples) and the borders between definition text, usage examples and idioms still remain to be fully identified in the XML structure. For the experiments in this study, in addition to the full original text, we create three other datasets in such a way that the number of space-separated tokens is limited to only 15, 20 and 25 tokens. The performance of the alignment task with respect to the number of tokens is shown in Section 4.
- Unifying sense hierarchy: The senses in both dictionaries are provided in a hierarchical form to represent semantically-related concepts. For our task, we bring all the senses along with subsenses at the same level. Having said that, the sense hierarchy structure can explicitly provide information about the semantic relationship between senses and therefore should preferably be considered in later experiments with the data.
- Dataset creation: Entries are linked using a common ID, called metaID, in ODS and DDO. Using this ID, senses of the same headwords in the two dictionaries are brought together for the annotation task.

#### 3.2 Manual Annotation

In the manual linking process where the training data was established, we annotated the senses of a large number of lemmas which were initially linked between ODS and DDO (meaning that they are etymologically the same words). The lemmas were picked out randomly among a selection of "core concept lemmas", already having been identified in DDO, constituting of a total of 4,646 DDO lemmas of which at least one sense constitutes the Danish equivalent of one of the 5000 core/base concept synsets in Princeton Wordnet (Pedersen et al. 2019). Approximately 75% of these DDO core concept lemmas are polysemous, and even though they only constitute 5% of the total number of lemmas in the dictionary, they cover more than 20% of its senses (Pedersen et al. 2019). The lemma selection thereby represents a high degree of polysemy which makes it highly suitable for our task. The DDO core concept lemmas cover both nouns, verbs, adjectives



and adverbs, and 86% of them have a lemma match in ODS, confirming that even though the DDO core concept lemmas were selected via an English selection, they are in fact central lemmas also in the Danish language. We excluded senses from fixed expressions in our dataset. Table 1 summarizes the sense statistics of the annotated data set.

Resource	Nouns	Verbs	Adjectives	Adverbs	Other	All
ODS	2176 (282040)	983 (119163)	36 (60599)	0 (0)	0 (0)	3595 (461802)
DDO	1036 (12326)	383 (4045)	248 (2228)	0 (0)	0 (0)	1667 (18599)

Table 1: The statistics of the annotated data based on (Ahmadi et al. 2020). The numbers in parentheses refer to the overall number of the tokens in senses.

In the manual annotation task, the hierarchical sense structure in ODS, including main sense as well as subsense numbers, is visible to the annotator while the DDO senses are presented in a random linear order with no information on the original sense numbers and hierarchical relations between senses. This facilitated the manual linking process since cases of potentially very different hierarchies in the two dictionaries did not disturb the picture. See figure 6.

pyramide (sb.)-19036640			
1) (massivt) bygningsværk af sten med firkantet grundflade og trekantede sider	exact	4-	1-todimensional figur der har form som en trekant med s
2) (mat.) legeme, hvis grundflade er en polygon, og hvis trekantede sider	exact	3-	2-bygning el. konstruktion med form som et sådant grav
3) hvad der har form af en pyramide ell. kegle; ogs. i videre anv., om hv	exact	2-	3-rumlige geometriske figurer der fremkommer ved at der fra
3. 1 ) om (del af et) bygningsværk (tårn, spir o.lgn.); nu især (jf. Pyramid narrow		2-	4-egyptisk gravmonument, ofte af meget store dimensio
3. 2 ) (især gart.) om træer (sjældnere andre planter). Pyramideaster, d narrow		2-	
3. 3 ) om (lille) pyramideformet ting ell. figur; fks. til havepynt: små Pyra narrow		2-	
3. 4 ) opstabling, opstilling af ting, der har form som en pyramide, tilspid narrow		2-	
3. 5 ) om (del af) møbel (hylde, opsats o.lgn.), der tilspidses opefter; spe			
3. 6 ) (fagl.) krystalform bestående af to mod hinanden vendte pyramide			
3. 7 ) (anat.) om forsk. fremspring o.lgn. d. s. s. Nyrepyramide . Anat.(18			
3. 8 ) om (konkylier med stærkt opsvulmet nederste vinding af) forsk. f			

Figure 6: The senses of the noun *pyramide* (“pyramid”) in ODS (column 1 to the left) and DDO (column 4 to the right) in the sheet used for the linking task. The linking values (relation, e.g. “exact” and sense number, e.g. “4”) are annotated in the columns 2 and 3. In ODS the original sense numbers and sense order is kept, in DDO the sense numbers are ad hoc, and the order does not correspond to the one in the dictionary.

We operate with the following types of relations between senses in the two dictionaries:

- **none:** There is no match for this ODS sense in DDO
- **exact:** The sense in ODS corresponds to the sense in DDO, for example, the definitions are simply paraphrases, as seen in the examples in Figure 5, or they describe the same concept in rather different ways, as seen in the examples in figure 2, 3 and 4. Senses are also considered to be exact matches in cases where the only difference is due to the modernization of society. E.g. the ODS sense of the noun *passager* (“passenger”) “person traveling with mail coach etc.”, was considered an exact match to the DDO sense “person traveling with private or public means of transportation”.
- **broad:** The sense in ODS completely covers the meaning of the sense in DDO, but is also applicable to further meanings. E.g. the ODS sense of the noun *værge* (“guardian”): “a guardian of anything or anybody” is a broader sense of the DDO sense restricted to “a guardian in legal context” (i.e. a guardian for a child not yet legally competent or for an incapacitated adult).
- **narrower:** The sense in ODS is entirely covered by the sense of DDO, which is also applicable to further meanings. In ODS the adjective *spids* (“sharp”) has, for example, two specific senses, one about a sound and another one about a smell, where DDO covers both senses in one definition: “pungent in an unpleasant way (about smell, taste or sound)”. Therefore, both ODS senses are considered to be narrower than the “lumbered” DDO sense.
- **related:** There are cases when the senses may be related even though the definitions in ODS and DDO differ in key aspects. For example, the property of “being able to sleep”, a sense of the noun *søvn* (“sleep”) in ODS is considered “related” to “the state of sleeping” sense in DDO, however not identical. The noun “*bamse*” (teddy bear) is in ODS, described as a “fat, clumsy person, especially a child”, is in DDO described as a “fat, good-natured person”, and these two senses are also considered to be related. Also, cases of regular polysemy are considered to be “related” matches. E.g. ODS has only one sense for the noun “*ambassade*” (“embassy”), namely the organization sense, while DDO has two: the organization sense as well, but also the building sense. While the organization sense is an exact match to the sense in ODS, the building sense is considered to be only “related” to it.



### 3.3 Models

Using the annotated data, we predict the similarity scores between senses using a similarity function. The similarity function is a trained model based on the following similarity features given that  $A$  is a sense in the first resource, ODS, and  $B$  is a sense in the other resource, DDO:

#### 1. String metrics

- **Longest common substring:** the length of the longest substring that exists in both senses
- **Length ratio:** the ratio of the number of space-separated tokens in each sense
- **Average word length ratio:** the average length of words in each sense
- **Jaccard, Dice, and Containment:**

$$J(A, B) = |A \cap B| / |A \cup B|,$$

$$D(A, B) = 2|A \cap B| / (|A| + |B|),$$

$$C(A, B) = |A \cap B| / \min(|A|, |B|).$$

- **Smoothed Jaccard:** this metric is an improved formulation of the Jaccard coefficient that makes the optimization possible and can be adjusted to distinguish matches on shorter texts (McCrae et al. 2017). It is defined as follows:

$$J_{\sigma}(A, B) = \frac{\sigma(|A \cap B|)}{\sigma(|A|) + \sigma(|B|) - \sigma(|A \cup B|)}$$

where  $\sigma(x) = 1 - \exp(-\alpha x)$  and  $\alpha$  is a constant.

2. **Word Embeddings:** with the current progress in the field of NLP, representing words within vector-spaces has been widely used and is proved to be beneficial in various applications. To evaluate the usability of word embeddings in the task of WSA, we also train a model based on ODS and DDO data using the Global Vectors for Word Representation (GloVe) model (Pennington et al. 2014). We took as a starting point the word embeddings model trained at DSL in the DDO project using a corpus of approximately one billion running words of modern Danish. The model is trained with 500 features, a window size of 5 and a minimum occurrence of 5 (any types below this threshold are discarded), and used the Skip-Gram version of the model. See Sørensen & Nimb (2018) for details about the model<sup>3</sup>.

Given  $V_A$  and  $V_B$ , the corresponding vector representations of each word in our word embeddings for senses  $A$  and  $B$ , we calculate the similarity between vectors using the cosine similarity as follows:

$$\text{Similarity based on the word embeddings: } \cos(\theta) = \frac{V_A \cdot V_B}{|V_A| |V_B|}$$

where  $\theta$  is the angle between two vectors projected in a multi-dimensional plane.

3. **Automatic feature extraction:** In this model, we automatically extract useful features from the input data in such a way that the performance of the extracted features is maximal among the whole combination of features.

Once the similarity scores are extracted, we automatically align senses in a bijective and greedy approach where the sense pairs are ordered based on the similarity score and then aligned in such a way that a sense is linked to only one other sense in the other resource. Although this bijective constraint ignores polysemous senses, it yields a more diverse combination of sense matches.

## 4 Evaluation

We evaluated the performance of the models using NAISC (McCrae & Buitelaar 2018). NAISC<sup>4</sup> is a tool for automatic alignment of lexical and ontological data which can be configured based on various semantic similarity extraction techniques including the ones described in Section 3.3. We use precision, recall and F-measure as our evaluation metrics as described by Nakache et al. (2005).

As discussed in Section 2, senses in ODS are long and unstructured. Therefore, in addition to the original ODS data, we create three other datasets where the number of space-separated tokens is limited to 15, 20 and 25. The performance of our similarity detection models with respect to each dataset is provided in Table 2.

Although the precision of the models in automatically detecting the similarity of two senses varies in a close range of 50.3% (All-auto) and 66.7% (15-Word embeddings), there is more significant difference between the recall of each dataset and so, in F-measure. The lowest recall appears in aligning DDO with ODS with its original senses. In other terms, when senses with all the composing parts, such as usage examples and idioms, are aligned with DDO, all the three models can predict a link over 50% correctly. However, they only succeed in less than 10% of cases to retrieve relevant senses. Truncating senses from 25 tokens to 15 significantly improves both the precision and recall, proving our initial observation of the noisiness of senses in ODS. Figure 7 illustrates the correlation of senses sizes with F-measures in all the models.

ODS sense size	Model	Precision	Recall	F-measure
----------------	-------	-----------	--------	-----------

<sup>3</sup> The paper describes the training of a previous version of the model. However, the only differences are that corpus material for 2018 and 2019 have been added and that the skip-gram version is chosen instead of CBOW.

<sup>4</sup> The tool is openly available at <https://github.com/insight-centre/naisc>



15	String metrics	65.3%	<b>48.1%</b>	55.4%
	Word Embeddings	<b>66.7%</b>	48.0%	<b>55.8%</b>
	Auto	64.0%	46.6%	54.0%
20	String metrics	61.5%	44.3%	51.5%
	Word Embeddings	64.7%	46.7%	54.3%
	Auto	63.3%	45.8%	53.2%
25	String metrics	57.5%	21.9%	31.7%
	Word Embeddings	55.9%	21.2%	30.8%
	Auto	58.5%	22.2%	32.1%
All	String metrics	54.7%	9.8%	16.7%
	Word Embeddings	50.7%	9.7%	16.3%
	Auto	50.3%	9.4%	15.8%

Table 2: The performance of our similarity detection models for automatic alignment of DDO and ODS within a specific limit of space-separated tokens (15, 20, 25 and all tokens).

The highest F-measure of 55.8% belongs to the ODS dataset with a maximum of 15 tokens and trained with the word embeddings model. In comparison to the baselines presented by Kernerman et al. (2020) where an F-measure of 4.3% is reported, such an improvement is promising.

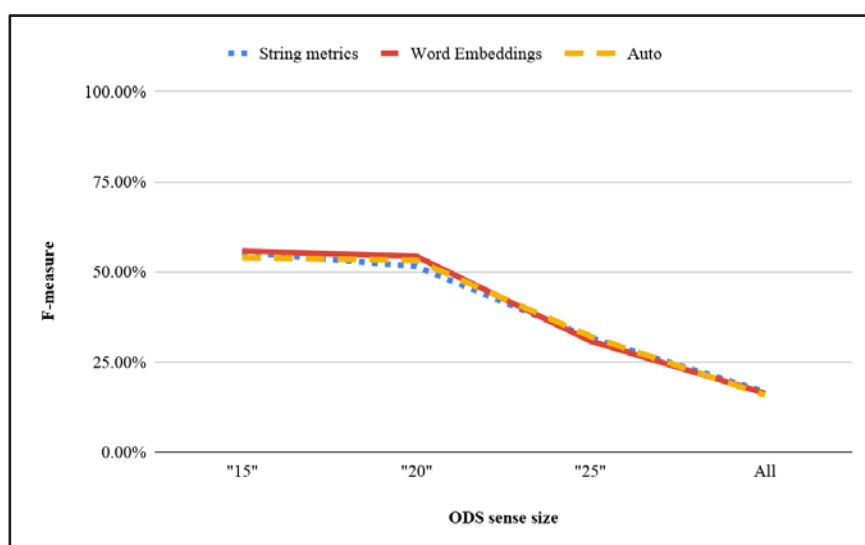


Figure 7: The correlation of sense sizes in ODS with F-measure using various methods.

## 5 Conclusion

In this paper, we studied the automatic alignment of senses across two Danish dictionaries, ODS and DDO. We demonstrate



that basic string similarity metrics along with word embeddings and automatic feature extraction models can be efficiently used to align senses of identical lemmas across these two resources. Converting printed historical dictionaries into structured electronic forms is an expensive and burdensome task. As future work, we are interested in exploring unsupervised methods to detect sense boundaries in dictionaries such as ODS. Moreover, we would like to explore further methods to automatically detect the type of the semantic relationship that may exist between two senses, also of non-identical lemmas, and study to which degree manual markups of the meta-information in the ODS improve the method, as well..

## 6 References

- Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S. & Troelsgård, T. (2020). A multilingual evaluation dataset for monolingual word sense alignment. In *Proceedings of the 12th Language Resource and Evaluation Conference (LREC 2020)*, volume 1, 3232–3242.
- Bjerring-Hansen, J., Jelsbak, T., Sørensen, N. H. & Fischer, F. (2019). 'Nodes and Edges in Literary History: Modelling 19th Century Literary Landscapes', *Digital Humanities*, Utrecht, 9-12 July, 2019, Accessed at: <https://georgbrandes.dk/research/3explorations/brandes-poster-dh2019-utrecht.pdf> [30/05/2020]
- Buitelaar, P. (2000). 'Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification'. In *Proceedings of the ANLP2000: Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems*. Seattle, USA. Accessed at: <http://dfki.de/~paulb/anlp00.html> [30.03.2010].
- Dahlerup, V. (1918-54). *Ordbog over det danske sprog, volume 1-28; Supplement til Ordbog over det danske Sprog, volume 1-5 (1992-2005)*. Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: [ordnet.dk/ods](http://ordnet.dk/ods)
- Hjorth, E., Kristensen, K. (2003-2005). *Den Danske Ordbog, volume 1-6*, Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: [ordnet.dk/ddo](http://ordnet.dk/ddo)
- Jacobsen, L., Juul-Jensen H. (1918). Indledning til Bind 1 in *Ordbog over det danske Sprog, volume 1*, 1918, Det Danske Sprog- og Litteraturselskab, Copenhagen. Accessed at: <https://ordnet.dk/ods/tekster-fra-den-trykte-ordbog> [30/05/2020]
- Kernerman, I., Krek, S., McCrae, J. P., Gracia, J., Ahmadi, S. & Kabashi B. (eds) (2020). Globalex Workshop on Linked Lexicography. *European Language Resources Association (ELRA) - LREC 2020 Workshop Language Resources and Evaluation Conference*, volume 1, 115.
- Kilgarriff, A. (2003). "I don't believe in word senses". In B. Nerlich, D. D. Clarke, Z. Todd & V. Herman (eds.), *Polysemy - Flexible Patterns of Meaning in Mind and Language*, Series: Trends in Linguistics. Studies and Monographs [TiLSM], 142, De Gruyter Mouton, pp. 361–392.
- Lorentzen, H. (2004). "The Danish Dictionary at large: presentation, problems and perspectives". In W. Geoffrey & S. Vessier (eds.) *Proceedings of the 11th EURALEX International Congress*, Vol. 1, pp. 285-294, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, Lorient, France.
- McCrae, J. P., Arcan, M. & Buitelaar, P. (2017). "Linking knowledge graphs across languages with semantic similarity and machine translation." *Foreword by the chairs*, Vol. 1, p. 31.
- McCrae, J. P., Buitelaar, P. (2018). Linking datasets using semantic textual similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123.
- Miles, A., Bechhofer, S. (2009). SKOS simple knowledge organization system reference. W3C recommendation, 18:W3C. Accessed at: <https://www.w3.org/TR/skos-reference/> [30/05/2020]
- Moro, A., Li, H., Krause, S., Xu, F., Navigli, R. & Uszkoreit, H. (2013). Semantic rule filtering for web-scale relation extraction. In *International Semantic Web Conference*, volume 1, pp. 347–362. Springer.
- Nakache, D., Metais, E. & Timsit, J. F. (2005). Evaluation and NLP. In *International Conference on Database and Expert Systems Applications*, volume 1, pp. 626-632. Springer, Berlin, Heidelberg.
- Nimb, S. (2018). The Danish FrameNet Lexicon: method and lexical coverage. In *Proceedings of the International FrameNet Workshop at LREC 2018*, vol. 1, p. 51-55, Miyazaki, Japan.
- Nimb, S., Lorentzen, H., Theilgaard, L. & Troelsgård, Th. (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab & Syddansk Universitetsforlag.
- Nimb, S. Sørensen N. H. & Troelsgård, T. (2018). "From standalone thesaurus to integrated related words in the Danish Dictionary". In: *Proceedings from Euralex 2018*, volume 1, p. 183, Ljubljana, Slovenia.
- Nimb, S. Trap-Jensen, L. & Lorentzen H. (2014). "The Danish Thesaurus: Problems and Perspectives". In: A. Abel, C. Vettori & N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen: EURAC Research, volume 1, pp. 191-199
- Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the generative lexicon conference*, volume 1, pp. 9–15. GenLex-09, Pisa, Italy.
- Pedersen, B.S, Nimb, S., Asmussen, J. Sørensen, N., Trap-Jensen, L. & Lorentzen H. (2009). "DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary". In: *Language Resources and Evaluation, Computational Linguistics Series*, Volume 3, 269-299.
- Pedersen, B. S., Nimb, S., Olsen & S. Sørensen, N. H. (2018). "Combining Dictionaries, Wordnets and other Lexical Resources - Advantages and Challenges". In *Globalex Proceedings 2018*, volume 1, p. 102-105, Miyasaki, Japan.
- Pedersen, B. S., Nimb, S., Olsen, I. R. & Olsen, S. (2019). "Linking DanNet with Princeton WordNet". In *Global WordNet 2019 Proceedings*, volume 1, 10 p. Wroclaw, Poland.
- Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, volume 1, pp. 1532-1543.
- Pustejovsky, J. (1995). *The Generative Lexicon*, Cambridge, MA: MIT Press.
- Pustejovsky, J. (1998). "The semantics of lexical underspecification". In *Folia Linguistica* 32(?), pp. 323– 347.



- Sørensen, N. H., Nimb, S. (2018). "Word2Dict–Lemma Selection and Dictionary Editing Assisted by Word Embeddings". In: *Proceedings from Euralex 2018*, volume 1, p. 148, Ljubljana, Slovenia.
- Svendsen Møller, M.-M., Sørensen, N.H., Troelsgård T. (2020). "An automatically generated Danish Renaissance Dictionary. Building a period dictionary by reducing and merging relevant existing dictionary resources". In *Proceedings of the LREC 2020 Globalex Workshop on Linked Lexicography* (I. Kernerman, S. Krek, J. P. McCrae, J. Gracia, S. Ahmadi & B. Kabashi), European Language Resources Association (ELRA), volume 1, pp. 29-32, Paris, France.

### Acknowledgements

This work has received funding from the EU's Horizon 2020 Research and Innovation programme through the ELEXIS project under grant agreement No. 731015. The authors would like to thank Thomas Troelsgård, Society for Danish Language and Literature (DSL) who contributed to the linking of identical lemmas between ODS and DDO and Sussi Olsen, CST, University of Copenhagen who contributed to the manual annotation task.



# Interlinking Slovene Language Datasets

Bajčetić L.<sup>1</sup>, Declerck T.<sup>1,2</sup>

<sup>1</sup>Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Austria

<sup>2</sup>DFKI GmbH, Multilinguality and Language Technology Lab, Germany

## Abstract

We present the current implementation state of our work consisting in interlinking language data and linguistic information included in different types of Slovenian language resources. The types of resources we currently deal with are a lexical database (which also contains collocations and example sentences), a morphological lexicon, and the Slovene WordNet. We first transform the encoding of the original data into the OntoLex-Lemon model and map the different descriptors used in the original sources onto the LexInfo vocabulary. This harmonization step is enabling the interlinking of the various types of information included in the different resources, by using relations defined in OntoLex-Lemon. As a result, we obtain a partial merging of the information that was originally distributed over different resources, which is leading to a cross-enrichment of those original data sources. A final goal of the presented work is to publish the linked and merged Slovene linguistic datasets in the Linguistic Linked Open Data cloud.

**Keywords:** Slovenian Language Data; interlinking; OntoLex-Lemon; LexInfo

## 1 Introduction

In the context of approaches aiming at the generation of a densely linked dataset of language resources, we are considering different types of Slovenian language resources, consisting of lexical, morphological, and conceptual data. The Slovene data sets we are including in our current work are:

1. Slovene Lexical Database (SLD) - a lexical database, which deals with collocations (Gantar & Krek 2011)
2. Sloleks - a morphological lexicon (Dobrovolic et al. 2017)
3. sloWNet - Slovenian WordNet (Fišer et al. 2012)

Linking and merging processes have been implemented for those three resources, and we are planning to extend the work to other (types of) resources. For the cross-linking and merging of the different types of language data we are making use of the OntoLex-Lemon framework (Cimiano et al. 2016). For the present work, we focus only on nouns. The use of OntoLex-Lemon for representing lexicographic data has been previously presented and discussed in (Declerck et. 2017; Tiberius & Declerck 2017). The relevance of OntoLex-Lemon for the representation of WordNet data and for interlinking and merging WordNet and lexicographic data for Romance languages has been demonstrated in (Racioppa & Declerck 2019) and we anticipated similar results for the Slovenian language.

In the following sections we present first the selected Slovenian resources. We continue with a brief description of the OntoLex-Lemon model and the LexInfo ontological vocabulary, which is providing data categories for the model. We then present some results of the linking and merging processes, as they are represented in OntoLex-Lemon. We close the paper with a description of the planned next steps.

## 2 The selected Slovene Datasets

In the following sections we briefly present the Slovenian data sets we are currently dealing with, and which are representing a wide coverage of different types of linguistic information. For all those datasets we give as an example the way they encode information related to the word “alergija” (*allergy*).

### 2.1 The Slovene Lexical Database (SLD)

The Slovene Lexical Database (SLD) is a lexical-conceptual resource, structured as a network of interrelated semantic and syntactic information about a word. SLD contains dictionary-type of information on words and word combinations, for example senses, collocations, example sentences, syntactic patterns, grammatical information, etc.

The database was compiled from the Gigafida corpus (Logar & Kosem 2011), a recent generation of Slovene corpora which contains 1,134,693,933 words from 38,310 texts of different genres, including Internet content. The core element of the resource is the lexical unit which includes all senses of the headword, multi-word expressions and phraseological units. SLD contains two types of information which are designed for two types of users: the first is the lexico-grammatical information that is intended for human users and comes in the form of sense descriptions as well as collocations and typical examples from the corpus, which are both attributed to particular senses and syntactic patterns of the lemma. The second type of information is devised for natural language processing tools. It includes the formal encoding of syntactic patterns at the clause and phrasal level (syntactic structures) as well as the formal encoding of



semantic arguments and their types. For our purpose, we have extracted all the lemmas with collocations and example sentences stored alongside the syntactic pattern which they exhibit. This way our integrated resource can be enriched with a multitude of example sentences which can potentially be used to improve disambiguation. The original encoding for the word *alergija* in SLD is given in Table 1 just below:

```
<zapis>alergija</zapis>
<iztocnica>alergija</iztocnica>
</oblika>
<zaglavje>
<besvrs>samostalnik</besvrs>
</zaglavje>
</glava>
<geslo>
<pomen>
<indikator>preobčutljivost organizma</indikator><oznaka tip="podrocje">zdravje</oznaka>
<pomenska_shema><definicija1>zdravstveno stanje, ki se kaže kot preobčutljivost organizma na določeno snov ali hrano, s katero pride v stik</definicija1>
<skladijske_skupine>
<skladijska_struktura>
<struktura>sbz0 SBZ2</struktura>
<kolokacije>
<kolokacija><k>zdravljenje, simptom, znak, ugotavljanje, diagnoza, odkrivanje</k> alergije</kolokacija>
<kolokacija><k>povzročitelj, sprožilec</k> alergije</kolokacija>
<kolokacija><k>nastanek, pojavljanje, pojav, porast, izbruh</k> alergije</kolokacija>
<kolokacija><k>razvoj, posledica, preprečevanje, vzrok, pogostnost, preprečitev</k> alergije</kolokacija>
<kolokacija><k>oblika, vrsta, primer, tip</k> alergije</kolokacija>
<kolokacija><k>nevarnost, napad, čas</k> alergije</kolokacija>
</kolokacije>
<zgledi>
```

Table 1: The entry for the word *alergija* in the Slovene Lexical Database

As the reader can observe, the used descriptors are in Slovene. There is a need to map those descriptors (or tags) to an interoperable vocabulary, which in our work is given by LexInfo (see Section **Fehler! Verweisquelle konnte nicht gefunden werden.**2). This mapping onto LexInfo is particularly relevant for the descriptor used within the “<struktura>” tag for marking the syntactic structure of a collocation, which is represented in SLD as an abstract pattern. See (Fišer et al. 2012) for a more in-depth explanation of the Slovene Lexical Database, where the authors already present an approach for relating SLD and sloWNet.

## 2.2 The Slovene Morphological Lexicon Sloleks

Sloleks is a large open-source machine-readable morphological lexicon for the Slovene language (Dobrovoljc et al. 2008). The lexicon provides inflectional, derivational, and grammatical information, formally represented within the XML serialization of the standardized LMF framework, which is described in (Francopoulou et al. 2006). For morphologically rich languages, such as Slovene, describing morphological paradigms of inflected parts of speech is a crucial step in creating the language model. Therefore, databases such as Sloleks are incredibly valuable. Sloleks is designed as a digital resource for computational linguistics, which means it contains a huge collection of systematically described morphological patterns in machine-readable format, resulting in almost 2 800 000 inflected forms. An entry includes the basic form (the lemma) of the word, its inflected forms (the inflectional paradigm) and related morphological information. Since we only focus on nouns for now, we have extracted all the noun entries which contain the lemma and all the inflected forms marked with case and number. Other information we have extracted is gender and pronunciation information, which can hopefully be used in the future, as well as some information of corpus frequency, also included in Sloleks. Table 2 just below displays the (simplified) Sloleks encoding of “alergija”, in a tabular format we designed for the purpose of this paper.

<i>Alergija</i>	singular		dual		plural	
case	word form	morpho-syntactic code	word form	morpho-syntactic code	word form	morpho-syntactic code
nominative	alergija	ncfsn	alergiji	ncfdn	alergije	ncfpn
accusative	alergijo	ncfsa	alergiji	ncfda	alergije	ncfpa
genitive	alergije	ncfsg	alergij	ncfdg	alergij	ncfpg
dative	alergiji	ncfsd	alergijama	ncfdd	alergijam	ncfpd
locative	alergiji	ncfsl	alergijah	ncfdl	alergijah	ncfpl
instrumental	alergijo	ncfsi	alergijama	ncfdi	alergijami	ncfpi

Table 2: The morphological variants of the entry “alergija” in Sloleks, with their original abbreviated encoding, which has been mapped onto LexInfo (see Section 3.2, Example 1)



## 2.3 The Semantic Lexicon of Slovene: sloWNet

sloWNet is the Slovene WordNet, developed in the expand approach: it contains the complete Princeton WordNet 3.0 and over 70,000 Slovene literals. These literals have been added automatically using different types of existing resources, such as bilingual dictionaries, parallel corpora, and Wikipedia. For the scope of this work we have extracted only the Slovene literals with their synset ids. For future work it would be interesting to see to what extent the semantic information encoded in PWN can be used for Slovene. How the lemma *alergija* is represented in sloWNet is shown in Table 3 just below:

```
<LexicalEntry id='w1167167'>
  <Lemma writtenForm='alergija' partOfSpeech='n'/>
  <Sense id='w1167167_05653475-n' synset='slv-10-05653475-n'/>
  <Sense id='w1167167_14532816-n' synset='slv-10-14532816-n'/>
  <Sense id='w1167167_14533796-n' synset='slv-10-14533796-n'/>
</LexicalEntry>
```

Table 3: The encoding of the word *alergija* in sloWNet, in the XML serialization of the LMF model

In Table 3, we can see that the linguistic information is poor in this resource. Only the part-of-speech is indicated for the corresponding lemma. Our work consists in linking the sloWNet concepts to the full lexical description available in Sloleks. This can be straightforwardly done, once both resources have been transformed onto the OntoLex-Lemon model.

## 3 OntoLex-Lemon and LexInfo

We present briefly the instruments used for transforming the original Slovene datasets into a harmonized shared representation. On the one hand we have the OntoLex-Lemon model, which we deploy for representing the lexical and conceptual data, and on the other hand the LexInfo vocabulary, which was, among others, developed for providing a set of data categories for use in OntoLex-Lemon.

### 3.1 OntoLex-Lemon

OntoLex-Lemon is a further development of the “Lexicon Model for Ontologies” (*lemon*, see McCrae et al. 2012). Both *lemon* and the OntoLex-Lemon model, which is resulting from a W3C Community Group, were originally developed with the aim to provide a rich linguistic grounding for ontologies, meaning that the natural language expressions used in the labels, definitions or comments of ontology elements are equipped with an extensive linguistic description (Cimiano et al. 2016). This rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or to specialized vocabularies. The main organizing unit for those linguistic descriptions is the *LexicalEntry* class, which enables the representation of morphological patterns for each entry (a multiword expression, a word, or an affix). The connection of a lexical entry to an ontological entity is marked mainly by the *ontolex:denotes* property or is mediated by the *LexicalSense* or the *LexicalConcept* classes, as this is represented in Figure 1, which displays the core module of the model. A major difference between *lemon* and OntoLex-Lemon is that the latter includes an explicit way to encode conceptual hierarchies, using the SKOS<sup>1</sup> standard. As can be seen in Figure 1, lexical entries can be linked via the *ontolex:evokes* property to such SKOS concepts, which we use to represent WordNet synsets. This structure is paralleling the relation between lexical entries and ontological resources, which is implemented either directly by the *ontolex:reference* property or mediated by the instances of the *ontolex:LexicalSense* class.

OntoLex-Lemon comes with additional modules, one of them being used in our transformation and harmonization exercise: the *decomp* module. This module supports the representation of elements of a multi-word or compound lexical entry. As can be seen in Figure 2 below, the *decomp* module makes use of the *decomp:subterm* property in order to indicate that a multi-word lexical entry contains other entries. The *decomp:Component* class is there for collecting the components of the lexical entry as the individual tokens (particular realizations of lexical entries) that compose that compound lexical entry. This module is relevant for representing the collocations that are included in SLD, as one view on collocations can be that they are a type of Multi Word Expressions (MWE).

<sup>1</sup> SKOS stands for “Simple Knowledge Organization System”. SKOS provides “a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary” (<https://www.w3.org/TR/skos-primer/>) [31.07.2020]



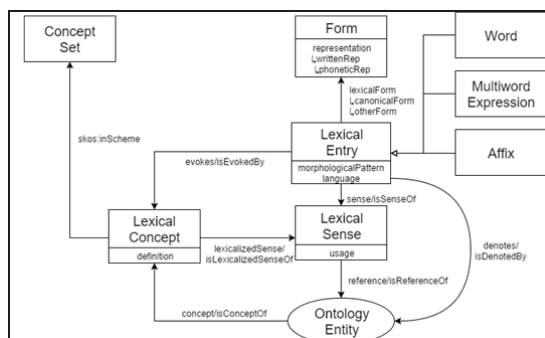


Figure 1: The core module of OntoLex-Lemon (taken from <https://www.w3.org/2016/05/ontolex/>)

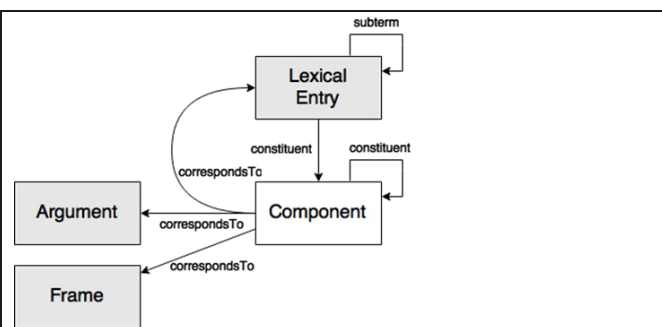


Figure 2: The decomp module of OntoLex-Lemon (taken from <https://www.w3.org/2016/05/ontolex/>)

### 3.2 LexInfo

As already stated, LexInfo is an ontology that was defined to provide data categories for the *lemon* model, and which has been updated with the new OntoLex-Lemon model of the W3C Ontolex community group. LexInfo is designed as an ontology and is written using the same W3C standards as OntoLex-Lemon, being RDF, RDF(s) or OWL, making thus use of resource-describing graphs as the basic representation instrument. We deploy LexInfo to harmonize all the descriptors (tags, metadata, etc.) used in the different Slovenian language data resources we are dealing with. Just to give an example, we map the morpho-syntactic descriptors used in Sloleks to the standardized LexInfo representation, shown in Example 1, taken from the Python code realizing the transformation (not considering for the time being the gender information):

```
slo_lexinfo_table["Ncfsn"] = ("lexinfo:singular", "lexinfo:nominativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfsd"] = ("lexinfo:singular", "lexinfo:dativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfda"] = ("lexinfo:singular", "lexinfo:accusativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfsl"] = ("lexinfo:singular", "lexinfo:locativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfsl"] = ("lexinfo:singular", "lexinfo:instrumentalCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfdn"] = ("lexinfo:dual", "lexinfo:nominativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfdg"] = ("lexinfo:dual", "lexinfo:genitiveCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfd"] = ("lexinfo:dual", "lexinfo:dativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfda"] = ("lexinfo:dual", "lexinfo:accusativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfdl"] = ("lexinfo:dual", "lexinfo:locativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfdi"] = ("lexinfo:dual", "lexinfo:instrumentalCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpn"] = ("lexinfo:plural", "lexinfo:nominativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpg"] = ("lexinfo:plural", "lexinfo:genitiveCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpd"] = ("lexinfo:plural", "lexinfo:dativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpa"] = ("lexinfo:plural", "lexinfo:accusativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpl"] = ("lexinfo:plural", "lexinfo:locativeCase", "lexinfo:feminine")
slo_lexinfo_table["Ncfpi"] = ("lexinfo:plural", "lexinfo:instrumentalCase", "lexinfo:feminine")
```

Example 1: Mapping the abbreviated morpho-syntactic code of Sloleks to the LexInfo vocabulary, which is equipped with unique reference points (URIs), as the name space "lexinfo" is standing for: "http://www.lexinfo.net/ontology/3.0/lexinfo"

## 4 The integrated Slovenian Language Data in the OntoLex-Lemon Representation

In this phase of our work we focused on nouns. From Sloleks, 50,873 nouns have been transformed onto instances of the `ontolex:LexicalEntry` class, while 854,813 morphological variants are now encoded as instances of the `ontolex:Form` class. The entries are declaratively linked to the forms via the properties `ontolex:canonicalForm` (linking the entry to its singular nominative form) or `ontolex:otherForm` (linking the entry to all other form variants). The forms are encoding the number information (while the gender and the pronunciation information are still to be added).

We have also mapped 18,735 *sloWNet* nominal entries onto the `LexicalConcept` class of OntoLex-Lemon. Some nominal entries included in *sloWNet* could not be considered, as we could not map their lemmas to Sloleks entries. This concerns mainly multiword items, which are not considered in Sloleks, but which might be present in SLD. This is a topic which stays next on our agenda. The *sloWNet* concepts, now encoded as part of a SKOS scheme, are linked to the Sloleks entries via the `ontolex:isEvokedBy` property (on the other way round by the `ontolex:evokes` property).

This way, both original resources, Sloleks and *sloWNet* are interlinked and merged in one and the same representation space. A benefit is for example that all form variants of Sloleks are now related to a *sloWNet* element.



In the following tables, we display the resulting representation for *alergija* at the lexical, morphological, and conceptual levels. In the first table (Table 5) the lexical entry representation is displayed. In Table 6 few examples of the form variants the entry is linking to are displayed (as the `rdfs:label` property is widely used in the Linked Data community, we keep it in parallel to the `ontolex:writtenRep` property). In Table 7 we can see the corresponding synset of *sloWNet*.

<pre>:Lex_646 rdf:type ontolex:LexicalEntry ; rdfs:label "alergija"@slv ; ontolex:canonicalForm :Form_646_0 ; ontolex:evokes :eng-30-14533796-n ; ontolex:otherForm :Form_646_1 ; ontolex:otherForm :Form_646_10 ; ontolex:otherForm :Form_646_11 ; ontolex:otherForm :Form_646_12 ; ontolex:otherForm :Form_646_13 ; ontolex:otherForm :Form_646_14 ; ontolex:otherForm :Form_646_15 ; ontolex:otherForm :Form_646_16 ; ontolex:otherForm :Form_646_17 ;</pre>	<pre>ontolex:otherForm :Form_646_2 ; ontolex:otherForm :Form_646_3 ; ontolex:otherForm :Form_646_4 ; ontolex:otherForm :Form_646_5 ; ontolex:otherForm :Form_646_6 ; ontolex:otherForm :Form_646_7 ; ontolex:otherForm :Form_646_8 ; ontolex:otherForm :Form_646_9 ;</pre>
---	--

Table 5: the OntoLex-Lemon representation of the entry *alergija*, with internal links to the forma variants and to the related *sloWNet* concept

<pre>:Form_646_10 rdf:type ontolex:Form ; lexinfo:case lexinfo:locativeCase ; lexinfo:number lexinfo:dual ; rdfs:label "alergijah"@slv ; ontolex:writtenRep "alergijah"@slv ;</pre>	<pre>:Form_646_12 rdf:type ontolex:Form ; lexinfo:case lexinfo:nominativeCase ; lexinfo:number lexinfo:plural ; rdfs:label "alergije"@slv ; ontolex:writtenRep "alergije"@slv ;</pre>
---	---

Table 6: Two example of form variant for the Sloleks entry *alergija*  
Information on gender and pronunciation will be added soon.

<pre>:eng-30-14533796-n rdf:type ontolex:LexicalConcept ; skos:inScheme :SlowNet ; ontolex:isEvokedBy :Lex_33897 ; ontolex:isEvokedBy :Lex_646 ;</pre>
--

Table 7: The *sloWNet* synset corresponding to the entry *alergija*. This synset is also evoked by another lexical entry.

The Tables 5-7 show how Sloleks and *sloWNet* are now in the same (harmonized) representation space and how they enrich each other.

While working with *sloWNet*, we noticed that only (if at all) very few definitions (glosses) and examples in Slovenian language are associated with the synsets. This a reason why we also started to work with the data included in SLD, as there a richer combination of semantic and syntactic aspects is described, also with links to corpora data. Our goal is then to extract such examples and definitions and to associate those with the *sloWNet* synsets, but also with the morphological forms which are included in the examples.

A first step towards this goal was to extract all examples and to be able to encode them in the OntoLex-Lemon environment. We defined for this a class “Examples” and we store examples associated with an entry as instances of this class, as this is displayed in Table 8 below.



```

:Example_45
rdf:type :Examples ;
rdfs:comment "for the lemma alergija"@en ;
rdfs:label "Zdravljenje <i>alergij</i> se zadnja leta izboljšuje z boljšimi sredstvi za diagnosticiranje, kot so krvni in kožni testi "@sl .

:Example_46
rdf:type :Examples ;
rdfs:comment "for the lemma alergija"@en ;
rdfs:label "Znaki <i>alergije</i> pa so predvsem odvisni od mesta in organa, kjer se začne alergična reakcija. "@sl .

:Example_47
rdf:type :Examples ;
rdfs:comment "for the lemma alergija"@en ;
rdfs:label "Najbolj običajni simptomi <i>alergije</i> so izpuščaji na koži in driska. "@sl .

```

Table 8: Example of the current implementation of the integration of examples taken from SLD.

Current work is being pursued in disambiguating some of the examples included for their appropriate linking to the concepts. As the reader can see, we still have for now tags around the form variants, as we want to extract those for relating the examples not only to the lemma, but to the concrete forms.

SLD is also containing a very rich list of collocations (the example sentences are in fact examples of such collocations, as encountered in large corpora). Current work consists in representing those collocations in OntoLex-Lemon. We started for this also a discussion within the W3C Community Group "Ontolex". An option would be to interlink the form variants that are in a collocation relation. Another option would be to consider collocations as kind of multiple word expressions. We are consulting for this also an additional Slovene lexical resource, MWElex, described in (Ljubesic et al. 2015), which also has a cross-lingual perspective. Representing collocations as MWE in OntoLex-Lemon can be done with the help of its decomp module.

```

:MWE_diagnoza_alergije
rdf:type ontolex:MultiWordExpression ;
rdfs:label "diagnoza alergije"@slv ;
<http://www.w3.org/ns/lemon/decomp#constituent> :alergije_comp ;
<http://www.w3.org/ns/lemon/decomp#constituent> :diagnoza_comp ;
<http://www.w3.org/ns/lemon/decomp#subterm> :Lex_6400 ;
<http://www.w3.org/ns/lemon/decomp#subterm> :Lex_646 ;

:MWE_sprožilec_alergije
rdf:type ontolex:MultiWordExpression ;
rdfs:label "sprožilec alergije"@slv ;
<http://www.w3.org/ns/lemon/decomp#constituent> :alergije_comp ;
<http://www.w3.org/ns/lemon/decomp#constituent> :sprožilec_comp ;
<http://www.w3.org/ns/lemon/decomp#subterm> :Lex_41850 ;
<http://www.w3.org/ns/lemon/decomp#subterm> :Lex_646 ;

:alergije_comp
rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
lexinfo:case lexinfo:genitiveCase ;
lexinfo:number lexinfo:singular ;
rdfs:label "alergije"@slv ;
<http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_646 ;

:diagnoza_comp
rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
rdfs:label "diagnoza"@slv ;
<http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_41850 ;

:sprožilec_comp
rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
rdfs:label "sprožilec"@slv ;
<http://www.w3.org/ns/lemon/decomp#correspondsTo> :Lex_6400 ;

```

Table 9: The Lemon-OntoLex encoding of the collocations "diagnoza alergije" and "sprožilec alergije"



Table 9 shows how the components of the MWE expressions are linked to the lexical entries, and so indirectly to the corresponding sloWNet entries (as can be seen in Table 5). This opens an interesting field of investigations, as we could aim this way at generating new sloWNet entries based on the semantic composition of the components of the collocations.

## 5 Current and future Work

Current work is dedicated to the extension of the number and types of Slovenian data sets. Therefore, we started the analysis of terminological resources, as the relevance of OntoLex-Lemon for representing the lexical elements of a terminology has also been shown in (Cimiano et al. 2015). We consider here the Slovene terminology “Evroterm”.

The downloadable version of Evroterm mentions 130.000 terms (for the date 10.10.2017). It is containing a mix of information, but not making a proper use of XML syntax. This aspect can be solved relatively easily, and we already mapped the data into a clean and machine interpretable XML structure. This resource is relevant to our project, as it also contains definitions in Slovene language that could be linked to the sloWNet data, which are lacking definitions in the Slovene language. Evroterm also contains term equivalents in various languages.

No part-of-speech information is assigned to the expressions realizing the terms. This is something we can easily add by linking the terms to Sloleks, at least for the terms that are not realized by multi-word-expressions but by sole words. Table 10 just below shows how the word *alergija* is encoded in the downloadable version of Evroterm.

```
<Entry Number>75845
<Subject>medicine ME
<Subj>medicina
<EN>allergy
<Definition>A condition of abnormal sensitivity in certain individuals to contact with substances such as proteins, pollens, bacteria, and certain foods. This contact may result in exaggerated physiologic responses such as hay fever, asthma, and in severe enough situations, anaphylactic shock.
<DefRef>KOREN
<SL>alergija
<TermRef>Besednjak Gemet - http://eionet-si.arso.gov.si/kpv/Gemet
<Definition>Nenormalna občutljivost pri nekaterih posameznikih na stik z določenimi snovmi, npr. beljakovinami, cvetnim prahom, bakterijami in določeno vrsto hrane. Ta stik lahko privede do pretirane fiziološke reakcije, kot je seneni nahod, astma ter v težkih primerih tudi do anafilaktičnega šoka.
<DefRef>KOREN
<DA>allergi
<CS>alergie
<DE>Allergie
<ES>alergia
<FI>allergia
<FR>allergie
<IT>allergia
<NL>allergie
<PL>alergia
<Definition>swoista reakcja ustroju na pewne zwi'zki chemiczne znajduj'ce się m.in. w powietrzu, w pokarmach, w bakteriach, béd'ca przyczyn' wielu chorób, np. astmy, pokrzywki i i
<PT>alergias
<SK>alergia
<SV>allergi
```

Table 10: The representation of the term *alergija* in Evroterm

The cross-lingual aspects present in Evroterm can be dealt with in OntoLex-Lemon with the help of its vartrans module. A cross-lingual resources, called MWElex, for multiple word expressions is also described in (Ljubecic et al. 2015), where MWEs are aligned across Serbian, Croatian, and Slovenian. The transformation of this resource onto OntoLex-Lemon and its linking to the other transformed and integrated Slovenian resources is also part of our current work.

## 6 Conclusions

We presented on-going work consisting in transforming a series of different rich Slovene lexical resources onto the OntoLex-Lemon framework. The current state of work shows the benefits of such an approach, as the lexical information from different resources is now available in one and the same representation format, supporting their interlinking and even merging.

The next steps will consist of representing the collocation information included in the cross-lingual MWElex resource and correctly linking it to elements of the current OntoLex-Lemon integrated data set. We will also continue working on the Evroterm data, as this resource is not only providing for terminological knowledge, but also with term translations.



## 7 References

- Cimiano, P., Buitelaar, P., McCrae, J. & Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics First Look*. 52 pages.
- Cimiano, P., McCrae, J. & Buitelaar, P. (2016). Lexicon Model for Ontologies. *W3C Community Report*.
- Cimiano, P., McCrae, J., Rodriguez-Doncel, V., Gornostaya, T., Gomez-Perez, A., Siemoneit, B., & Lagzdins, A. (2015). Linked Terminology: Applying Linked Data Principles to Terminological Resources. In *Proceedings of eLex 2015*.
- Declerck, T., McCrae, J., Navigli, R., Zaytseva, K. & Wissik, T. (2018). ELEXIS - European Lexicographic Infrastructure: Contributions to and from the Linguistic Linked Open Data. In *Proceedings of the 2nd GLOBALEX Workshop*.
- Declerck, T., Tiberius, C. & Wandl-Vogt, E. (2017). Encoding lexicographic Data in lemon: Lessons learned. In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. Galway, Ireland, CEURS, 8/2017
- Dobrovoljc, K., Krek, S. & Erjavec, T. (2017). The Sloleks Morphological Lexicon and its Future Development. In *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana University Press, Faculty of Arts.
- Evroterm. Accessed at <https://evroterm.vlada.si/evroterm>. [31/05/2020]. The terminology data in text format can be downloaded at: <http://podatki.vlada.si/evroterm.aspx> [31/05/2020].
- Fišer, D., Novak, J. & Erjavec, T. (2012). sloWNet 3.0: development, extension and cleaning. In *Proceedings of the 6th International Global Wordnet Conference*. The Global WordNet Association.
- Fišer, D., Gantar, P. & Krek, S. (2012). Using explicitly and implicitly encoded semantic relations to map Slovene wordnet and Slovene lexical database. In *Proceedings of the 8th Conference on Language Resources and Evaluation*.
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel., Pet.M. & Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.
- Gantar, P. & Krek, S. (2011). Slovene lexical database. In *Proceedings of the sixth international Conference on Natural language processing, multilinguality*.
- Krek, S., Kosem, I., McCrae, J.P., Navigli, R., Pedersen, B.S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts*.
- Lexicon Model for Ontologies. Accessed at <https://www.w3.org/2016/05/ontolex/> [31/05/2020].
- LexInfo. Accessed at <https://lexinfo.net/ontology/3.0/lexinfo> [31/05/2020].
- Ljubecic, N., Dobrovoljc, K., & Fiser, D. (2015). \*MWElex - MWE Lexica of Croatian, Slovene and Serbian Extracted from Parsed Corpora. *Informatica (Slovenia)*, 39.
- Logar Berginc, N. & Kosem, I. (2011): Gigafida – the new corpus of modern Slovene: what is really in there? In *Proceedings of the Slavicorp conference*. Dubrovnik.
- McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.
- Morphological lexicon Sloleks 2.0. Accessed at: <https://www.clarin.si/repository/xmlui/handle/11356/1230> [31/05/2020]
- Racioppa, S. & Declerck, T. (2019). Enriching Open Multilingual Wordnets with Morphological Features. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari, Italy, CEUR, 10/2019
- Semantic lexicon of Slovene sloWNet. Accessed at <https://www.clarin.si/repository/xmlui/handle/11356/1026> [31/05/2020].
- Slovene Lexical Database. Accessed at <http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza> [31/05/2020].
- Tiberius, C. & Declerck, T. (2017). A lemon Model for the ANW Dictionary. In *Proceedings of the eLex 2017 conference*, Pages 237-251. Leiden, Netherlands, Lexical Computing CZ s.r.o., INT, Trojina and Lexical Computing, Brno, Czech Republic

## Acknowledgements

Work by DFKI was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 825182 through the Prêt-à-LLoD project, and the COST Action CA18209, NexusLinguarum: European network for Web-centred linguistic data science. Work by the Austrian Centre for Digital Humanities and Cultural Heritage was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 through the ELEXIS project. We would also like to thank the anonymous reviewers for their helpful comments on the paper.



# Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues

Ducassé M.

Univ Rennes, INSA Rennes, CNRS, IRISA -UMR 6074, France, mireille.ducasse@irisa.fr

## Abstract

The Georgian language has a complex verbal system, both agglutinative and inflectional, with many irregularities. Inflected forms of a given verb can differ greatly from one another and it is still a controversial issue to determine which lemmas should represent a verb in dictionaries. Verb tables help people to track lemmas starting from inflected forms but these tables are tedious and error-prone to browse. We propose Kartu-Verbs, a Semantic Web base of inflected Georgian verb forms. For a given verb, all its inflected forms are present. Knowledge can easily be traversed in all directions: from Georgian to French and English; from an inflected form to a *masdar* (a verbal noun, the form that comes closest to an infinitive), and conversely from a *masdar* to any inflected form; from component(s) to forms and from a form to its components. Users can easily retrieve the lemmas that are relevant to access their preferred dictionaries. Kartu-Verbs can be seen as a front-end to any Georgian dictionary, thus bypassing the lemmatization issues.

**Keywords:** Georgian verbs; Inflected forms; Dictionary front-end; Semantic web tool; Prolog

## 1 Introduction

Georgian is a Caucasian language, mother tongue of about 5 million people. It has its own alphabet.<sup>1</sup> Georgian grammar has a complex verbal system. Some issues are illustrated below from a beginner's perspective (for more details see for example Anderson 1984; Tuite 1998; Assatiani & Malherbe 2011; Gérardin 2016). There are numerous irregular verbs and the language is both agglutinative and inflectional. Conjugation can modify both the beginning and the ending of verbs. For example, the verb “to work” (*mushaoba* - მუშაობა), in the first-person plural of the present tense produces “*vmushaobt*” (ვმუშაობთ). Note the “v” at the beginning of the verb to mark the first person, and the “t” at the end to mark the plural. Some tenses, such as the future, often introduce a preverb. For example, for the verb “to work”, the first-person singular future is “*vimushaveb*” (ვიმუშავებ). An “i” has been inserted after the “v” marker of the first person; it is somewhat regular for a large set of verbs. Note that “ob” has changed into “eb”; it is typical of a smaller subset of verbs. The apparition of post-radical “v” is more exceptional. Many verbs have different stems in different tenses. For example, the third-person singular forms of the verb “to see” are respectively “*khedavs*” (stem, “*khed*”, ხედავს) in present and “*nakha*” (stem, “*nakh*”, ნახს) in aorist. Indications of directions are given by prefixes. For example, “I go” = “*mivdivar*” (მივდივარ), “you go down” = “*chadikhar*” (ჩადიხარ), “she comes” = “*modis*” (მოდის). Note the different prefixes, and the very different markers of persons. A dozen prefixes can be used. Table 1 gives the conjugation tables for 3 tenses, present, future and present perfect. Beyond the above comments, note the mechanism in present perfect, the preradicals are very different. Those preradicals are used in other tenses, for other groups. For example the 3 persons singular at present of group 2 verb “to love” (“*siqvaruli*”, სიყვარული) are respectively, “*miqvars*” (მიყვარს), “*giqvars*” (გიყვარს) and “*uqvars*” (უყვარს).

To work /მუშაობა		Present		Future		Present perfect	
I	მე	<i>vmushaob</i>	ვმუშაობ	<i>vimushaveb</i>	ვიმუშავებ	<i>mimushavia</i>	მიმუშავია
you	შენ	<i>mushaob</i>	მუშაობ	<i>imushaveb</i>	იმუშავებ	<i>gimushavia</i>	გიმუშავია
s.he	ის	<i>mushaobs</i>	მუშაობს	<i>imushavebs</i>	იმუშავებს	<i>umushavia</i>	უმუშავია
we	ჩვენ	<i>vmushaobt</i>	ვმუშაობთ	<i>vimushavebt</i>	ვიმუშავებთ	<i>gvimushavia</i>	გვიმუშავია
you	თქვენ	<i>mushaobt</i>	მუშაობთ	<i>imushavebt</i>	იმუშავებთ	<i>gimushaviat</i>	გიმუშავიათ
they	ისინი	<i>mushaoben</i>	მუშაობენ	<i>imushaveben</i>	იმუშავებენ	<i>umushaviat</i>	უმუშავიათ

Table 1: Three tenses of the verb “to work/მუშაობა”.

The preceding examples are not exhaustive. They only aim at illustrating the difficulty of morphosyntactic analysis of Georgian verbs and pave the way to introduce some issues of verb lemmatization in Georgian Dictionaries (details can be found in (Margalitadze 2020; Gippert 2016). Georgian has no infinitive. Most dictionaries use the “*masdar*” (a verbal noun that is the form closest to an infinitive) as lemma to represent a verb.<sup>2</sup> However, for neophytes, going from a conjugated form to a *masdar* can be a real challenge. For example, for “*chamodikhar*” (ჩამოდიხარ, “you come down”), the *masdar* is “*mosvla*” (მოხვლა, “coming”). Many projects give samples of inflected forms as lemma(s). For example, the third-person singular future is used in Daraselia and Sharoff (2016). The “Comprehensive Georgian-English Dictionary” presents, for all verbs, *masdar* and 3rd person singular in present and future tenses, both active (transitive)

<sup>1</sup> We use a transliteration in Latin characters, in this article and in Kartu-Verbs, to ease non-native Georgian speaker's reading. The transliteration is currently “French” oriented for historical reasons.

<sup>2</sup> In our system and in this report, a *masdar* is currently improperly called “Georgian infinitive” because it is easier to understand for the (French or English non-linguist) target users.



form and passive (intransitive) form, with markers for the indirect object in the third person. This is more exhaustive than in any previous bilingual Georgian dictionary (Rayfield et al., 2006). It is, however, still difficult for a neophyte to track the above-mentioned “chamodikhar”. The Georgian-German dictionary by Tschenkeli et al. (1965) uses *the abstract verbal root under which all subparadigms are listed. It can result in an extremely complex structure of entries* (Gippert 2016). While this representation is very informative for linguists, it is too cumbersome for beginners, especially as many roots consist of only one or two characters.

Some linguists provide exhaustive tables of inflected forms, for example the *Georgische Verbtabelle* by Chotiwari-Jünger et al. (2010) or the “Biliki series” books by Nana Shavtvaladze.<sup>3</sup> The latter contain conjugation tables of several types in appendix of the lessons. The first type of tables (henceforth “whole conjugation tables”) concerns the verbs introduced in a given lesson. They are systematically conjugated in all the tenses that have been introduced in the lessons so far. In these tables, *masdar* and English translation are also given. The second type of tables (henceforth “sample tables”) gives a list of conjugation samples, one line per verb. A line contains firstly an English translation and, for each tense introduced in the book, an inflected form at the third person singular. Those tables contain invaluable information; they are a tremendous help for neophytes. However, learners have to browse through different books to find relevant information. Finding an inflected form (in Georgian) in order to translate from Georgian to English is difficult. Indeed, the lines are sorted by English translation. When searching for an inflected form, learners have to check each one of the more than 10 000 entries. Furthermore, the inflected forms use the Georgian alphabet, which is a big hurdle for beginners. Exceptions, which are quite common, cannot always be anticipated from the sample tables. Verbs introduced in the first books do not have a complete “whole conjugation” table because few tenses have been presented at the time these verbs are introduced. Searching is thus tedious; it takes time and it is not granted that users find an entry.

We propose Kartu-Verbs, a Semantic Web base of Georgian inflected verb forms that can be seen as a front-end to any dictionary, thus bypassing the lemmatization issues.<sup>4</sup> When a verb is in the base, all its inflected forms are present and users can retrieve the lemmas relevant to access their preferred dictionary. As illustrated in depth in Section 2, knowledge can easily be traversed in all directions: from Georgian to French and English and conversely; from an inflected form to a *masdar* and from a *masdar* to any inflected form; from component(s) to forms and from a form to its components. In order to build the base, conjugation rules, taking exceptions into account, are built in Prolog, a programming language designed for language processing (Colmerauer 2011). The generated forms are integrated within a Semantic Web tool, Sparklis, which can retrieve information from their facets, and which allows users to smoothly refine their queries by giving them suggestions (Ferré 2017). The base currently contains over 15 000 inflected forms related to 278 verbs for 10 tenses.<sup>5</sup> As discussed in Section 3, in comparison with related work and to our best knowledge, our tool is the only one of its type.

## 2 Using Kartu-Verbs, Our Georgian Verb Form Base

This section illustrates how to use our base of Georgian verb inflected forms and demonstrates the power of the tool. As the base is primarily meant to be a companion of the “Biliki” books already mentioned, we use the knowledge structures defined by Nana Shavtvaladze (groups, subgroups, morphological decomposition, etc.). We are aware that the morphological decomposition is simplified, for example with respect to the work of Kevin Tuite (1998). Section 2.1 illustrates how to find information about an inflected form, our initial goal for the project. Section 2.2 shows that it is equally easy to get information starting from an English infinitive. Section 2.3 shows how to build a sample of conjugation. Section 2.4 describes how to gain conjugation information from a given ending. Sections 2.5 and 2.6 illustrate how to gain knowledge by comparing similar forms or stems. Section 2.7 shows how to check hypotheses about preradicals using logical operators. Section 2.8 discusses more sophisticated queries to gain meta-knowledge about the base using aggregates.

### 2.1 Finding Information About an Inflected Form

The user interface of Kartu-Verbs consists of 3 areas related respectively to the query, the suggestions and the results. Figure 1 shows two of those areas: the query area on the left-hand side and the “Suggestions” area on the right-hand side. The displayed query enables the user to find 12 features of inflected forms: its form in Georgian alphabet, person, number, tense, ..., French infinitive. The “Suggestions” area is itself divided in two areas. On the left, the “Types and Relations” area suggests features that can still be added to the query; on the right, the “Identities or Values” area suggests some of the verb forms that match the query. Let us assume that the user is interested in the “inadirebdnen” verb form and that he would also like to have information about its “Georgian infinitive” feature. He can click on both suggestions. Figure 2 displays the query and result areas after those selections. The query has been automatically updated. At the top there is no longer “give me every verb” but “inadirebdnen”, and “Georgian infinitive” has been added in the list of features. In the result area, now, the 13 requested features of “inadirebdnen” correspond to 13 columns. We can see, for example, its form in Georgian alphabet, “ინადირებდნენ”, and that it corresponds to the third person plural of both conditional and future conjunctive.

Any field could be used to search the base. As opposed to paper tables, there are no predefined uses. As illustrated above, all queries are built using suggestions. Users do not have to invent anything. They can use filters to help Sparklis propose relevant suggestions, then queries are built solely by clicking on suggestions that are necessarily relevant. The benefits

<sup>3</sup> Biliki, Georgian Language For English Speakers. See <http://lsgeorgia.com>.

<sup>4</sup> The base is available at <https://www-semis.irisa.fr/software/georgian-verb-inflected-forms-base/>

<sup>5</sup> Present, imperfect, conjunctive, future, conditional, future conjunctive, aorist, optative, present perfect, past perfect



are threefold, firstly it is easier to find something in a list than typing it, secondly users cannot mistype, and lastly, as a direct consequence, the queries can never give an empty result. That is a very strong property.

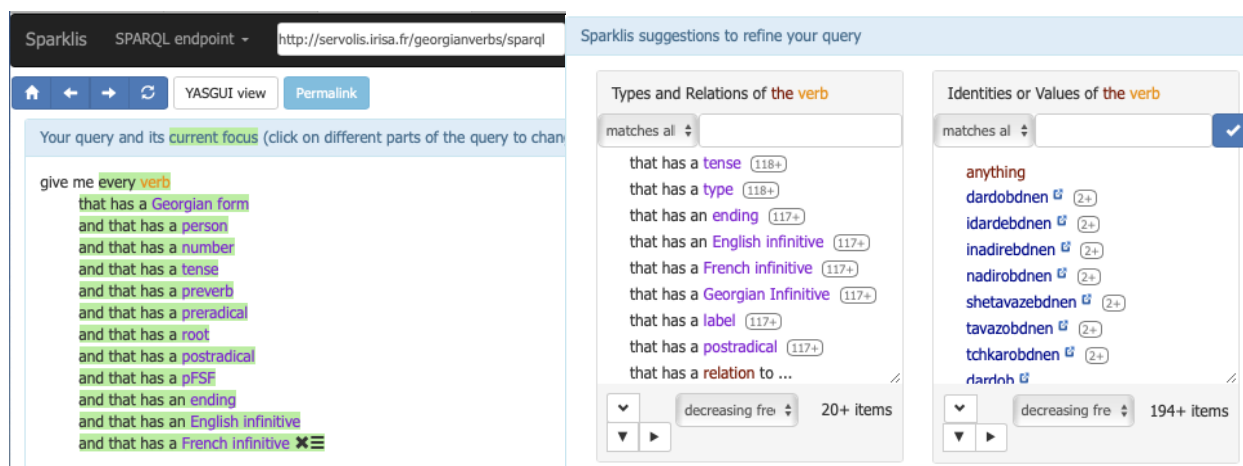


Figure 1: Defining a query with the help of suggestions.

Your query and its **current focus** (click on different parts of the query to change it)

**inadirebdnen** [↗](#)

- has a Georgian form
- and has a person
- and has a number
- and has a tense
- and has a preverb
- and has a preradical
- and has a root
- and has a postradical
- and has a pFSF
- and has an ending
- and has an English infinitive
- and has a French infinitive
- and has a Georgian Infinitive [✕](#) [≡](#)

Sparklis suggestions to refine your query

Results of your query

Table Map Slideshow

2 results Show 10 results

	Georgian form	person	number	tense ②	preverb	preradical	root	postradical	pFSF	ending	English infinitive	French infinitive	Georgian Infinitive
1	ინადირებდნენ	third	plural	conditional	-	i	nadir	-	eb	dnen	hunt	chasser	nadiroba
2	ინადირებდნენ	third	plural	futureConjunctive	-	i	nadir	-	eb	dnen	hunt	chasser	nadiroba

Figure 2: Getting information about an inflected form.

## 2.2 Finding Information From an English Infinitive

Figure 3 shows how to use criteria to search for information. Values are given to some features: the English infinitive has to be “to live”) and the tense has to be “present”. The features that do not have a value specified in the query (here “Georgian form”, “person”, “number”, “ending” and “Georgian infinitive”) are called “open”. They correspond to requested information and they produce the columns of the result area. The table in the result area gives the conjugation at the 6 persons at the present tense, with the requested information, in particular that it corresponds to Georgian verb “tskhovreba”. The features with a value in the query (here “English infinitive” and “Tense”) are not repeated in the result area. One needs the query in order to interpret the results. To get all the tenses, users can specify them in turn in the query in order to get them one after the other, or they can leave the tense feature “open” and all the -tenses will be given. This view is equivalent to a “whole conjugation table” of the Biliki books mentioned in the Introduction. Note that, at any moment, users can add or retract any value or feature from the query. The result and suggestion areas are then automatically updated by Sparklis.



Your query and its **current focus** (click on different parts of the query to change it)

give me every **verb**  
 that has a **Georgian form**  
 and that has a **person**  
 and that has a **number**  
 and whose **tense is present** ✕≡  
 and that has an **ending**  
 and whose **English infinitive is live**  
 and that has a **Georgian Infinitive**

Sparklis suggestions to refine your query

Results of your query

Table Map Slideshow

6 results Show 100 results

	verb ⑥	Georgian form ⑥	person ③	number ②	ending ④	Georgian Infinitive
1	vtskhovrob	ცხოვრობ	first	singular	-	tskhovreba
2	tskhovrob	ცხოვრობ	second	singular	-	tskhovreba
3	tskhovrobs	ცხოვრობს	third	singular	s	tskhovreba
4	vtskhovrobt	ცხოვრობთ	first	plural	t	tskhovreba
5	tskhovrobt	ცხოვრობთ	second	plural	t	tskhovreba
6	tskhovroben	ცხოვრობენ	third	plural	en	tskhovreba

Figure 3: The conjugation of the six persons at a given tense for a verb given in English.

## 2.3 Building a Sample of Conjugation

Your query and its **current focus** (click on different parts of the query to change it)

give me every **verb**  
 that has a **Georgian form**  
 and whose **person is second**  
 and whose **number is singular**  
 and that has a **tense**  
 and whose **Georgian Infinitive is tskhovreba**  
 and that has an **English infinitive** ✕≡

Results of your query

Table Map Slideshow

10 results Show 20 results

	verb ⑩	Georgian form ⑩	tense ⑩	English infinitive
1	tskhovrob	ცხოვრობ	present	live
2	tskhovrobdi	ცხოვრობდი	imperfect	live
3	itskhovreb	ცხოვრებ	future	live
4	itskhovre	ცხოვრე	aorist	live
5	itskhovro	ცხოვრო	optative	live
6	itskhovrebdi	ცხოვრებდი	conditional	live
7	itskhovrebde	ცხოვრებდე	futureConjunctive	live
8	getskhovra	გეცხოვრა	pastPerfect	live
9	gitshkovria	გიცხოვრია	presentPerfect	live
10	tskhovrobde	ცხოვრობდე	presentConjunctive	live

Figure 4: All 10 tenses at the second person singular of a verb given in Georgian.

Figure 4 shows how to conjugate Georgian verb “tskhovreba” in all the tenses known by the base for the second person singular. This is equivalent to a “Sample table” of the “Biliki” books. The advantage is that users can chose the person(s) they want or any criteria. Note that this time, we have specified the verb by its Georgian infinitive but we could have given one of its English or French infinitives.

## 2.4 Finding Possible Tenses From a Given Ending

Let us, now, assume that the user searches a verb form that is not present in the base but for which the user thinks that the ending is “da”. It could already be interesting to know the possible tenses. Figure 5 shows a query that sets the ending and asks for many features (“Georgian Form”, person, number, tense, preverb, PFSF and English infinitive), in order to try to map the searched verb to what is currently in the base. The result area shows 10 forms out of the more than 200 ones that match the query. The forms on display all correspond to a third person singular, in imperfect or conditional and with a PFSF being either “eb” or “ob”. It gives interesting trends. The user can check the remaining forms in order to confirm them (not illustrated here).



Results of your query

Table Map Slideshow

results 1 - 10 of 200+ Show 10 results

Your query and its current focus (click on the focus)

give me every verb  
that has a Georgian form  
and that has a person  
and that has a number  
and that has a tense  
and that has a preverb  
and that has a pFSF  
and whose ending is da X  
and that has an English infinitive  
and that has a Georgian Infinitive

	verb (147+)	Georgian form (147+)	person	number	tense (2+)	preverb (9+)	pFSF (4+)	English infinitive (94+)	Georgian Infinitive
1	tavazobda	თავაზობდა	third	singular	imperfect	-	ob	offer	shetavazeba
2	shetavazebda	შეთავაზებდა	third	singular	conditional	she	eb	offer	shetavazeba
3	dardobda	დარდობდა	third	singular	imperfect	-	ob	feel_sorrow	dardi
4	idardebda	იდარდებდა	third	singular	conditional	-	eb	feel_sorrow	dardi
5	nadirobda	ნადირობდა	third	singular	imperfect	-	ob	hunt	nadiroba
6	inadirebda	ინადირობდა	third	singular	conditional	-	eb	hunt	nadiroba
7	tchkarobda	ჩქარობდა	third	singular	imperfect	-	ob	hurry	sitchkare
8	itchkarebda	იტყვარობდა	third	singular	conditional	-	eb	hurry	sitchkare
9	khumrobda	ხუმრობდა	third	singular	imperfect	-	ob	joke	khumroba
10	ikhumrebda	იხუმრებდა	third	singular	conditional	-	eb	joke	khumroba

Figure 5: Finding possible tenses from a given ending.

## 2.5 Comparing Similar Forms

When learning, it is often useful to confront similar forms. For example, let us assume that the user realizes that he is confused about “to have someone” and “to resemble”. Figure 6 illustrates how to display the third singular present form for both verbs, using feature “French infinitive” and the “or” logical operator. The result area shows that the difference between the two forms consists in only one character. Thus, the user has learnt that “to have someone” has a “q” (“ყ”) as second letter and “to resemble” a “g” (“გ”).

Your query and its current focus (click on the focus)

give me everything  
that has a Georgian form  
and that is a verb  
and whose person is third  
and whose number is singular  
and whose tense is present X  
and that is something  
and whose French infinitive is  
resembler  
or avoir\_quelqu\_un  
and that has an English infinitive  
and that has a Georgian Infinitive

Results of your query

Table Map Slideshow

2 results Show 100 results

	verb (2)	English infinitive (2)	Georgian Infinitive (2)
1	hqavs	have_someone	qola
2	hgavs	resemble	damgvaneba

Figure 6: Comparing two similar forms.

## 2.6 Investigating Similar Stems

Similarly, let us assume that the user is confused about verbs containing “gheb” (“გებ”) in their form, not knowing exactly which type of morpheme it is. Figure 7 illustrates how to use the suggestion area to help on this matter. The query requests the verb to be third singular present and asks information about English and Georgian infinitives as well as stem/root. The green underlining in the query area indicates that the focus for the suggestions is on the “Georgian form”. The user has typed “გებ” in the suggestion area and Sparklis has automatically produced 3 suggestions (“იღებ”, “ღებ” and “გებ”). The result area shows 8 results for verbs whose Georgian form at third singular present matches “გებ”. For verb “to dye” the stem/ root is exactly “gheb” and the PFSF is “av”, while for verbs of the “to take/receive” family the stem/root is “gh” and “eb” is the PFSF. Thus, the user has learnt that “to dye” and “to take/receive” are not acquainted.

Note that “იღებ” and “ღებ” each give several answers. “იღებ” corresponds to 2 different Georgian infinitives and 2 different English translations.<sup>6</sup> “ღებ” corresponds to 3 different English translations. Figure 7 shows the complete display of the “Suggestions” area. The “Types and Relations” area (on the left) and the “Identities or Values” area (in the middle) have already been introduced. Let us remind here that they suggest, respectively, features that can still be added to the query and some of the values that match the query. The “Aggregation and Operators”

<sup>6</sup> Actually, “agheba” means “to take” and “migheba” means “to receive”. There should be 2 lines for “ighebs” and not 4. While it is a powerful feature to be able to display several lines for a given inflected form, the system shall be enhanced to remove irrelevant products.



area (on the right) allows users to build more sophisticated queries as illustrated in the following sections.

YASGUI view   Permalink   parts of the query to change it)

give me everything  
that has a Georgian form ✖  
and that is a verb  
and whose person is third  
and whose number is singular  
and whose tense is present  
and that is something  
and that has an English infinitive  
and that has a Georgian Infinitive  
and that has a root

Sparklis suggestions to refine your query

Types and Relations of the Georgian form

matches z

that is the Georgian form of ... (3)  
that has a label (3)  
that has a relation to ...  
that has a relation from ...  
according to which there is ...

decreasing fr 5 items

Identities or Values of the Georgian form

matches z ღებ ✓

anything  
იღებს (4)  
უღებს (3)  
იღებაღს

decreasing fr 4 items

Aggregations and Operators

matches all of

for each thing give me ...  
give me the number of thing  
give me a sample of thing  
if else  
if else  
is a blank node  
is a IRI  
is a literal  
is a number

14 items

Results of your query

Table Map Slideshow

◀ 8 results ▶ Show 20 results

	verb (3)	English infinitive (6)	Georgian form (3)	Georgian Infinitive (4)	root (2)
1	ighebs	take	იღებს	agheba	gh
2	ighebs	take	იღებს	migheba	gh
3	ighebs	receive	იღებს	agheba	gh
4	ighebs	receive	იღებს	migheba	gh
5	ughebs	make_copy	უღებს	gadagheba	gh
6	ughebs	take_a_picture	უღებს	gadagheba	gh
7	ughebs	give_another_helping	უღებს	gadagheba	gh
8	ighebavs	dye	იღებაღს	shegheba	gheb

Figure 7: Investigating similar stems.



## 2.7 Learning About Preradicals

Let us assume that the user believes that the first person always has preradical “v” or “vi”. Figure 8 shows a query, using logical operators “and” and “not”, that searches for forms in the first person (singular or plural, as number is not specified) and whose preradical is neither “v” nor “vi”. The addition of “something” in the query tells the system that preradical values are of interest. The suggestion area immediately shows that there are at least 6 other possibilities. For example, as illustrated in Table 1 “mi” and “gvi” are used in the present perfect for some verbs. The user has to refine his knowledge!

The screenshot shows the Sparklis interface. On the left, under 'Your query and its current focus (click)', a query is built: 'give me every verb that has a Georgian form and whose person is first and whose preradical is not v and not vi and something'. On the right, under 'Identities or Values of the preradical', a list of suggestions is shown: 'anything', 'va' (70+), 'gve' (30+), 'gvi' (30+), 'me' (30+), 'mi' (30+), and '-' (10+).

Figure 8: Learning preradicals at first person, using logical operators “and” and “not”.

## 2.8 Meta-Knowledge About the Base

The screenshot shows the Sparklis interface with a complex query: 'give me every verb that has a group and that has an ending and that has a pFSF and that has a preverb and that has a Georgian Infinitive and for each group give me a sample of verb and the number of verb and the number of pFSF and the number of preverb and the number of Georgian Infinitive'. Below the query, there are sections for 'Sparklis suggestions to refine your query' and 'Results of your query'. The results are displayed in a table with 4 results, showing the distribution of verbs across groups (g1 to g4) based on various features.

	group	sample of verb	number of verb	number of pFSF	number of preverb	number of Georgian Infinitive
1	g4	khar	1626	13	18	32
2	g2	mqopnis	599	3	7	11
3	g3	movip'ove	3051	7	8	56
4	g1	vtavazob	9988	15	13	179

Figure 9: Gaining knowledge about the group distribution, their numbers of PFSF, preverbs and Georgian infinitives using aggregation operators.

Sparklis also enables the user to gain knowledge about the current base. For example, Figure 9 shows a more sophisticated query, using aggregates, to gain information about the distribution of the 4 Biliki groups. The query shows that the features of interest are group, ending, pFSF, preverb and Georgian infinitive. It requests that all the verbs with these features are grouped according to their group (“g1” to “g4”). For each group, a sample should be given and the number of verb inflected forms, the number pFSF, the number of preverbs and the number of Georgian infinitives are requested. In the result area, we can see that group g1 is the largest one. In the current state of the base, it gathers 179 Georgian infinitives, 9 988 inflected forms, 15 different PFSF and 13 preverbs. The result area also shows a sample of



each group. Note that this query, as all the previous ones, was built solely by clicking. Here the right hand part of the “Suggestions” area had been used (see previous section).

### 3 Discussion, Perspectives and Conclusion

To our best knowledge, our tool is the only one of its type. We have, for example, found nothing specific for the Georgian language on the MultiTAL platform,<sup>7</sup> expert in automatic language processing (TAL) focused on Eastern and/or poorly endowed languages (Sadoun et al. 2016). The Georgian Wiktionary<sup>8</sup> is aimed at Georgian-speaking people. It is of no help to people who are beginning to learn the language. Google translate<sup>9</sup> is still doing quite poorly to translate Georgian verbs. INESS:XLE-Web,<sup>10</sup> the system of Paul Maurer (2007), is aimed at linguists. It is able to parse sentences and produce syntax tree of a number of languages, including Georgian. While its linguistic power is much larger than what Kartu-Verbs offer, the information that we need is buried in the syntax tree and not really accessible to beginners. Furthermore, there are no transliterations, no translations, and last but not least, none of our querying possibilities.

Our project is still under development. Currently, the base contains over 15 000 inflected forms related to 278 verbs for 10 tenses (present, imperfect, conjunctive, future, conditional, future conjunctive, aorist, optative, present perfect, past perfect). The forms have been generated and tested by students who are native Georgian speakers. At least all the verbs of the “Biliki” books are covered. One can expect that the most useful verbs for everyday life are already present. According to Tuite (1998), 5 tenses are missing: present iterative, imperative, permansive, mixed conjunctive present, perfect conjunctive.

The short-term perspectives are as follows. The current verb forms are being systematically tested. We are still in the process of analyzing exceptions and irregularities. A library of usual queries is under construction. Phonetic and English-oriented transliterations are planned in order to help non-French users. More verbs will be added. Links to an actual electronic dictionary will be inserted (for example, to the Comprehensive Georgian-English Dictionary by D. Rayfield, on the site of the National Parliamentary Library of Georgia.<sup>11</sup>)

In the medium term, we have to slightly revise the ontology that is structuring our form description in order to use a vocabulary more standard in linguistics and to be able to accommodate other types of words (nouns, adjectives, ...). We have to adapt the generation rules in order to be able to build forms with direct and indirect object markers (see for example (Assatiani and Malherbe 2011)), a feature that is especially confusing for French and English speakers. For example, “I do” (without other indication) = “vak’eteb” (ვაკეტებ), “I do for me” = “vik’eteb” (ვიკეტებ), “I do for you” = “gik’eteb” (გიკეტებ) [note the disappearance of the first person marker, “v”], “you do for me” = “mik’eteb” (მიკეტებ), etc.

In the longer term, we intend to complete the system to help users: 1) enter new verbs, 2) validate the newly produced inflected forms, and 3) update the conjugation program when exceptions are detected by experts. At some point, it will be important to ensure that the tool is collaborative, and that any user can suggest modifications and new entries in the database in a safe way.

Kartu-Verbs is an on-going project with many perspectives. In its current state it is already a successful proof of concept. In this paper we have shown how versatile and powerful its querying mechanisms are and how they can help users to easily get information about verbs that they encounter in Georgian text whatever their form. Kartu-Verbs can be used as a front-end to any Georgian dictionary, whatever lemmatization principles that dictionary uses for verbs.

### 4 References

- Anderson, S. R. (1984). On representations in morphology case, agreement and inversion in Georgian. *Natural Language & Linguistic Theory*, 2(2):157–218.
- Assatiani, I. and Malherbe, M. (2011). *Parlons Géorgien*. L’Harmattan.
- Chotiari-Jünger, S., Melik’išvili, D., and Wittek, L. (2010). *Georgische Verbtabelle*. Buske.
- Colmerauer A. (2011). From natural language processing to Prolog. *Presentation at the Academy of Mathematics and Systems Science*, Beijing, April 8, 2011. <http://alain.colmerauer.free.fr>.
- Daraselia, S. and Sharoff, S. (2016). Enriching Georgian dictionary entries with frequency information. In Margalitadze, T. and Meladze, G., editors, *Proceedings of the 17th EURALEX International Congress*, pages 321–327, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.
- Ferré, S. (2017). Sparklis: An expressive query builder for SPARQL endpoints with guidance in natural language. *Semantic Web: Interoperability, Usability, Applicability*, 8(3):405–418.
- Gippert, J. (2016). Complex morphology and its impact on lexicology: the Kartvelian case. In Margalitadze, T. and Meladze, G., editors, *Proceedings of the 17th EURALEX International Congress*, pages 16–36, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.
- Gérardin, H. (2016). *Les verbes intransitifs primaires et dérivés en Géorgien : Description morphosyntaxique, sémantique et dérivationnelle*. PhD thesis, Institut National des Langues et Civilisations Orientales. UMR 7192 – « Proche-Orient-Caucase : langues, archéologie, cultures ».
- Margalitadze, T. (2020). Lexicography of Georgian: a brief overview. *Language@Leeds Working Papers in Linguistics*,

<sup>7</sup> <http://multital.inalco.fr>

<sup>8</sup> [https://meta.wikimedia.org/wiki/List\\_of\\_Wiktionaries\\_by\\_language\\_family\\_-\\_Kartvelian\\_languages\\_\(8,074\\_-\\_1\\_-\\_1\)](https://meta.wikimedia.org/wiki/List_of_Wiktionaries_by_language_family_-_Kartvelian_languages_(8,074_-_1_-_1))

<sup>9</sup> <https://translate.google.fr>

<sup>10</sup> <http://clarino.uib.no/iness/xle-web>

<sup>11</sup> დიდი ქართულ-ინგლისური ლექსიკონი. <http://www.nplg.gov.ge/gwdict/index.php?a=index&d=46>



1. Forthcoming.

- Meurer, P. (2007). A computational grammar for Georgian. In *International Tbilisi Symposium on Logic, Language, and Computation*, pages 1–15. Springer.
- Rayfield, D., Apridonze, S., Chanturia, A., Amirejibi, R., Broers, L., Chkhaidze, L., and Margalitadze, T., editors (2006). *A Comprehensive Georgian-English Dictionary*. Garnett Press.
- Sadoun, D., Mkhitarian, S., Nouvel, D., and Valette, M. (2016). The MultiTal NLP tool infrastructure. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 156–163.
- Tschenkéli, K., Marchev, Y., and Flury, L. (1965). *Georgisch-Deutsches Wörterbuch*, volume 2. Amirani-Verlag Zürich.
- Tuite, K. (1998). *Kartvelian morphosyntax: Number agreement and morphosyntactic orientation in the South Caucasian languages*. Lincom Europa Munich.

## Acknowledgements

We are thankful to Irma Grdzeldze and Tina Margalitadze from Ivane Javakhishvili Tbilisi State University for providing more hindsight of this work. We are indebted to Tina for pointing out the verb lemmatization issue. Thanks to the Erasmus+ International Credit Mobility program, the following Georgian students contributed to the project: Ketī Meipariani, Mariam Asatiani, Mikheil Maisuradze, Tamari Kldiashvili, Tamar Sharabidze, Beka Chachua, Ana Idadze, Veriko Nikuradze, Tornike Tchanturia, Ana Elchishvili, and Aleksandre Jajanidze. We are grateful to Ketī Meipariani who managed the working teams of two semesters and who remotely helps the following teams. Last but not least, we want to warmly thank Sébastien Ferré for his support to use Sparklis.







# Making Dictionaries Visible, Accessible, and Reusable: The Case of the Greek Conceptual Dictionary API

Giouli V., Sidiropoulos N.F.

*Institute for Language and Speech Processing/ "Athena" Research Centre, Greece*

## Abstract

Language resources of any type are of paramount importance to several Natural Language Processing applications; developing and maintaining, however, quality lexical semantic resources is still a laborious and costly task that presents various challenges. In this respect, there is an ever-growing demand for resources that are visible, easily accessible, inter-operable and re-usable. The paper presents work in progress aimed at the development of a web service and the integration of a semantic lexical resource for Modern Greek in it, with a view to enabling robust ‘search and retrieve’ case scenarios. Given a lexeme, the intended service returns lexical semantic information encoded in the conceptual dictionary. The web service and the dictionary jointly form an infrastructure that can be exploited not only by researchers interested in studying the lexicon of the Modern Greek language, but also in application scenarios involving deep semantic information.

**Keywords:** conceptual dictionary; RESTful API; web service; accessing, querying, and re-using dictionaries

## 1 Introduction

Lexicographic data stored in databases constitute valuable resources that might be useful not only to human end-users but also to researchers and application developers alike. In this respect, the overwhelmingly big datasets of today ask for robust and efficient tools and services that will facilitate easy access to all sorts of language data. Currently, a variety of similar services and tools exist for well-resourced languages since most key players in the Lexicography industry have already opted to provide access to their data under a variety of licences and business models. However, it is still difficult to spot similar infrastructures for less-resourced languages. This paper presents a web Application Programming Interface (API) that enables robust “search and retrieve” case scenarios over a conceptual dictionary of Modern Greek (MG).

The remainder of the paper is structured as follows: In section (2) we provide a brief overview of trends in the development of semantic lexical resources that can be exploited for Natural Language Processing (NLP) applications and the initial endeavours of stakeholders in the industry of lexicography to provide quality data. The conceptual lexical resource that constitutes the basis of the intended infrastructure is described in section (3); the approach taken towards developing the dictionary API is presented in section (4) along with extended examples of the functionalities provided. Finally, our conclusions and plans for future research and development are outlined in section (5).

## 2 Background and Objectives

Over the last decades, semantic representation at word, phrase and sentence level has been the focus of attention in the field of NLP. In this context, the development of lexical resources (semantic lexica, thesauri, ontologies) coupled with information about the words and their meaning take linguistic theories of semantic representation into account. Over the last years, machine-readable hand-crafted datasets have been created primarily for the English language. Among these, WordNet (Fellbaum 1998), Verbnet (Kipper-Schuler 2005), and FrameNet (Fillmore et al. 2001) have been successfully employed in applications, whereas similar projects world-wide have resulted in resources in other languages as well.

From another perspective, the ontological approach to the representation of lexical semantics (ontology-driven lexical semantics), is largely based on the principles of Artificial Intelligence. In this context, an ontology formally specifies the concepts based on their referential status as well as the relationships that hold between these concepts. Therefore, the ontological approach to the lexical meaning is aimed at representing formally our knowledge of the entities and the relations that hold between them. In a sense, this approach to the representation of the lexical meaning reflects our grasping of entities and abstract concepts that surround us in the world as we perceive it. Prominent examples of ontologies include the OntoWordNet Project (Gagnemi et al. 2003) which has defined relations between WordNet synonym sets (synsets) as well as a set of logical, ontological and contextual commitments; similarly, SUMO (Suggested Upper Merged Ontology) ontology (Niles & Pease 2001; Pease 2011) consists of a basic upper-level ontological scheme and a number of domain-specific ontologies; it is thus one of the largest standard ontologies used in a variety of NLP applications. Similarly, Mikrokosmos (Mahesh & Nirenburg 1995a; Mahesh & Nirenburg 1995b) is a language independent ontology that is organized around the upper concepts Event, Object and Property.

Moreover, crowdsourcing techniques have been extensively employed with a view to boosting the creation of large-scale language resources and catering, thus, for the so-called knowledge acquisition bottleneck (Gale et al. 1992). The



multi-lingual wiktionary dictionary (Meyer & Gurevych 2012) is developed in accordance with the principles set by Wikipedia; similarly, DBpedia is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an open knowledge graph which is available for everyone on the Web (Morsey et al. 2011). The resources thus constructed are freely available for research purposes via dedicated APIs.

Still, the quest for hand-crafted quality language data that are made accessible and open to interested parties is still present. In this context, stakeholders in the industry of Lexicography provide their data to prospect developers; Oxford and Meriam Webster Dictionaries APIs are pioneers in this respect.

In this framework, Greek is still under-represented, and therefore, our objective in creating the current API was two-fold: on the one hand, we wanted to render the conceptual dictionary of MG that we are developing open, visible, accessible and re-usable; on the other hand, we wanted to provide high quality lexicographic data to prospect application developers.

### 3 The Conceptual Dictionary of Modern Greek

The core of the infrastructure is a computational dictionary of Modern Greek (Fotopoulou & Giouli 2015). It is a conceptually organized dictionary that builds on the format proposed by an already existing lexical resource model (Markantonatou & Fotopoulou 2008). In the current implementation, the initial model was appropriately extended and modified where needed. As a result, the new architecture was defined along the following axes: (a) to ensure compatibility and interoperability with standardized tools and resources, (b) to define a robust – yet extensible – taxonomical system that could be applied consistently throughout the dictionary, (c) to account for the efficient description of the semantic properties of words and the relations that hold between them, and (d) to ensure the functionality of the final resource and the user-friendliness to the lexicographer while encoding. In this respect, the whole ontology was re-designed and extended. Initially, we adopted widely accepted standards for lexicon description (namely, the Lexical Markup Framework ISO (TC37/SC4)) and the appropriate mappings were performed through-out the resource.

In our approach, we have chosen the ontological approach to the representation of lexical semantics. Following the initial model adopted, lexical entries are represented in the dictionary as instances modelling the Saussurian notion of the linguistic SIGN and its two inseparable facets, namely, the SIGNIFIER and the SIGNIFIED (Markantonatou & Fotopoulou 2008). The final resource forms a linguistic ontology in which the linguistic SIGN is instantiated as an instance in the ontology that is represented as the unique combination of a word and a lexical concept. In this approach, words (word forms) are instances in the SIGNIFIER class; these are further specified for (a) morphosyntactic properties: grammatical category or Part of Speech (POS), gender, argument structure, and word specific information; (b) lexical relations such as word families, allomorphs, syntactic variants etc.; and (c) features pertaining to lexical semantic relations (i.e., synonymy, antonymy). Values for these features are assigned to both single- and multi-word entries in the lexicon. Finally, following common lexicographic practices, all entries are coupled with rich linguistic information, that is, gloss, one or more usage examples, register, and prior polarity information. The afore-mentioned lexicographic information is encoded via a set of relations (i.e., `has_gender`, `has_pos`, `has_allomorph`, `has_synonym`, `has_antonym`, etc).

#### 3.1 The Hierarchical Ontological Schema: the SIGNIFIED Class

Similarly, word meanings (or lexical concepts) are instances in the SIGNIFIED class. Each instance in the SIGNIFIER class is mapped onto a concept, the latter represented as an instance in the SIGNIFIED class. Moreover, depending on the kind of relations defined thereof, concepts in the SIGNIFIED class are organized under two main classes: the class IS-A and the class ABOUT-A. In the current implementation, the two classes were re-designed and defined from scratch. The former class comprises a set of sub-classes that define a taxonomy hierarchically organized; concepts in this taxonomy are primarily connected via the hyponymy-hyperonymy or Is-a relation. Meronymy or part-whole relations are also encoded in this class as well. The latter class comprises a schema adopted from well-defined thesauri, namely Roget's Thesaurus (Hullen 2003) and Onomastikon (Vostantzoglou 1962). Sub-classes in the ABOUT-A class are semantically homogenous; they are defined via and populated by groups of concepts that are related via a set of semantic and pragmatic relations. In this way, words are inter-linked via a dense semantic net within or across POS categories and classes. The ontological hierarchy provides relations of inheritance, from general classes to classes that are lower in the hierarchy. As a result, the organization attempted provides for a functional and effective encoding interface, that facilitates the encoding process. The upper ontology defined within the current project is depicted in Figure 1.

Being defined as two-place predicates involving two lexical items, all relations between concepts in the ontology are further defined as inverse relations as well. For example, the relation `is_the_quality_of` has been paired with the inverse relation `has_the_quality`, and vice versa, as depicted in (1):

(1) `is_the_quality_of` ↔ `has_the_quality`

As a result, lexical instances like *εξυπνάδα* (=intelligence) and *έξυπνος* (=intelligent) are interlinked via both relations:

(2) `is_the_quality_of`(*εξυπνάδα*, *έξυπνος*)

(3) `has_the_quality`(*έξυπνος*, *εξυπνάδα*)

As noted above, the resource developed builds on existing models for semantic representation and extends the existing language resource infrastructure by integrating, harmonizing, and extending already existing lexical resources. However, the approach taken differs from them in several ways. Being primarily an ontological approach to the representation of meaning, the relations depicted are extended beyond mere lexical semantic ones. In this regard, the resource at hand differs from WordNet, a large lexical database of English in which nouns, verbs, adjectives, and adverbs are grouped into



sets of cognitive synonyms (synsets), each expressing a distinct concept. As a result, our dictionary also differs from lexical resources that were built following the WordNet paradigm for other languages – including Greek. Similarly, the conceptual dictionary adopts a FrameNet-like cognitive approach to the representation of meaning – yet the relations defined are not limited to the ones entailed by the argument structure of the predicates encoded.

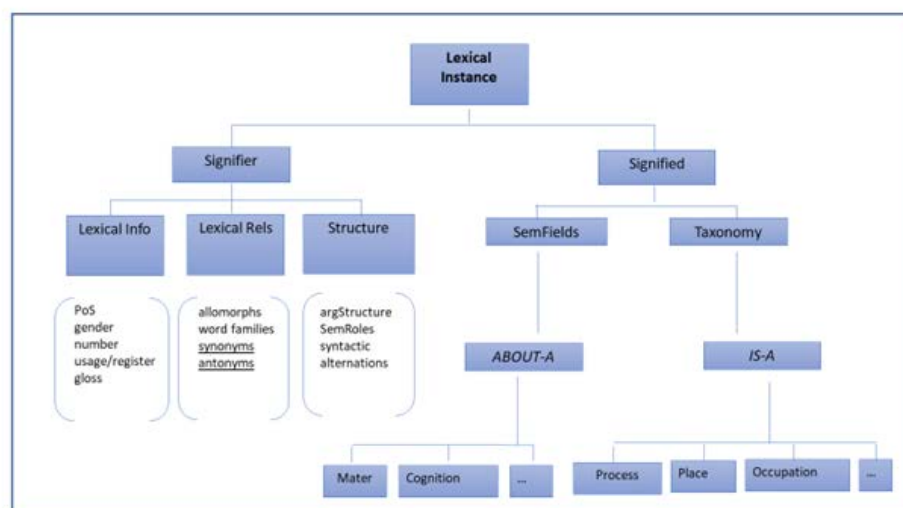


Figure 1: The search for the Greek word άνθρωπος.

### 3.2 The Dictionary in Numbers

The conceptual dictionary comprises lexical entries that belong to the grammatical categories Adjective, Adverb, Noun and Verb. It features both single and multi-word lexical entries, defined as lexical items which are composed by more than one lexical items and are characterized by semantic non-compositionality, syntactic non-modifiability, and the non-substitutability of components by semantically similar words. MWEs are further coupled with rich linguistic information that is useful for their robust representation in the lexical resource. To date, the dictionary comprises c. 26,5 lexical instances (entries) which have been classified under c. 250 classes (semantic fields). These instances are mapped onto ~13K concepts, whereas a set of c. 120 relations defines a dense network of linked lexical instances. The integration and handling of the relations and lexical instances developed in the API is still in progress.

POS	Single	MWEs	Total
Adjective	5214	70	5284
Adverb	1068	439	1507
Noun	13659	933	14592
Verb	3939	1191	5130
Total	23880	2633	26513

Table 1: Distribution of lexical instances per Part of Speech (PoS).

## 4 Creating the Dictionary API

An application programming interface (API) is seen as a computing interface which defines interactions between multiple software intermediaries. In this regard, a dictionary API is the intermediate between developers who might be in need of language data and the data providers. Therefore, an API defines the kinds of calls or requests that can be made to the data at hand, the format of the data, the conventions to follow, etc. In this section, we describe the work towards building the dictionary API and the searches foreseen so far.

### 4.1 Data Preparation

For the purpose of creating an interoperable tool between computer systems which would exploit the conceptually organized lexical resource, we selected the RESTful API architecture (Fielding 2000) as an easy way to enable robust and flexible “search and retrieve” case scenarios. Since there was no need to manipulate the resource data, but only make



them available, RESTful architecture seemed ideal.

As the dictionary was built on Protégé and later moved to WebProtégé, a web application running on Tomcat Apache server in Ubuntu 14, we had to find a way to provide access to dictionary data. WebProtégé uses MongoDB, a non-relational database, making access difficult to implement, since MongoDB uses its own query language. So, we decided to find a way to migrate the data to a relational database. PHP provided the RAP-RDF API which enabled that migration. The process involved the exportation from WebProtégé of a text file in RDF/XML format. The file was next loaded into a PHP scripting page, running in an IIS web server, and then imported in a MySQL 5.5 relational database using RAP-RDF API. The main dictionary information was saved in a three-field table called 'statements' with corresponding columns "subject", "predicate" and "object". This structure is in accordance with the triplets' structure that an ontology dictionary is based on.

## 4.2 API Functionality

In order to develop a RESTful API we used the same IIS Web Server running on Windows. PHP was the scripting language while the use of Slim framework for PHP provided the RESTful functionality. The framework simply translates plain URL paths following the RESTful logic to corresponding SQL queries to the MySQL database, while the response to each search is provided in JSON format. Since the dictionary material is important as a digital asset that needs to be protected and regularly maintained and updated, the RESTful API will be running under an API Gateway Platform, Tyk serving also other API services. Tyk is only regulating the use of the Conceptual Dictionary API and not the data per se; this was important in view of ensuring a proper and sufficient use of the service. The Tyk platform provides the use of basic authentication parameters (username and password) to discourage improper use. For this paper and for testing our API, we have excluded the Tyk Gateway, until we analyze its usage and determine the running parameters corresponding to the load of and behavior towards our server.

A JSON result consists of blocks of data in a form of attribute-value pairs, namely the search parameter name and its corresponding value as extracted from the dictionary. In case a value consists of several value data, then a sub-block of data appears in the value place. This node-like structure permits a limitless depth of data presentation following a hierarchical presentation. Examples of JSON results will be presented in the following section where we describe all our API searches that are available as functionalities.

### 4.2.1 Searching the Dictionary

As mentioned above, our API provides several functionalities (searches) in the form of simple URL requests, as RESTful architecture demands. All our API requests are performed with GET method, so there is no need of sending parameters except the ones included in our URL path. There is a general word search that provides all information related to the word matches but also specific searches that retrieve part of information according to our needs.

The most simple or general search of a word is performed by a simple URL:

*GET http://www.xanthi.ilsp.gr/apis/polytropon/word/<word\_text>*

where the <word\_text> is the Greek word we search for. So, our API URL for the word άνθρωπος (anthropos, =man) will assume the following form:

*GET http://www.xanthi.ilsp.gr/apis/polytropon/word/άνθρωπος*

The above URL performs a search in the dictionary for the word "άνθρωπος" along with other matches of words starting with the same characters. Each matching result is returned as a block of information containing the text of the word and all dictionary information related to that word (category type, synonyms, antonyms, morphologically related words). The response is formatted in JSON (Figure 2) so it can be easily used in every environment.

To provide error control, we have the "error" parameter which appears in each result set and has a "false" value whenever the search result is successful. This is very helpful for developers that will use the API to check the status of their request result and conduct error handling. Moreover, a "message" parameter includes all necessary data retrieved by our dictionary. Each match is presented in a form of JSON block that has the same structure. This structure contains a set of parameters that contain the information encoded in the dictionary. In the current release of the dictionary API, the following parameters can be retrieved:

- **wordtxt**: Contains the word string on which search matching is performed. The number in parenthesis notes the number of similar word records with different meaning.
- **params**: Includes all possible dictionary information.
- **word\_type**: Returns the value "single" for a single word and "mwe" for a phraseological unit or multi-word entry that includes the searched word.
- **PoS**: Returns the Part of Speech characterization of the matching word. Returned values are No (Noun), Ad (Adverb), Aj (Adjective), Pp (Preposition) and Vb (Verb).
- **wordstring**: Returns the plain word string.
- **number**: Returns the value "Sg" for singular and "Pl" for plural. This information in the dictionary is retained only for lexical instances that belong to the grammatical category Noun and are only plural.
- **gender**: Returns the value "Ma" for Masculine, "Fe" for feminine and "Ne" for neuter. This information is valid for lexical instances that belong to the grammatical category Noun.



- **degree**: Returns the value “Ba” for basic, “Co” for comparative and “Su” for superlative. As expected, this information is valid only for lexical instances that belong to the grammatical category Adjective or Adverb.
- **morphologically\_related**: Returns an array of words that are morphologically related to the search word. It should be noted that lexical instances that are encoded as morphologically related are linked via derivation relations.
- **synonyms**: Returns an array of words characterized as synonyms of the matching word. In the current implementation, only lexical items that are absolute synonyms as opposed to near synonyms (Cruse 1986) have been encoded. To this end, the dictionary specifications cater for the distinction of the two types of synonyms, and stylistic, expressive, and structural variations are taken into account.
- **antonyms**: Returns an array of words characterized as antonyms of the matching word. According to the lexicographic specifications set, both gradable and reciprocal antonymy relations have been encoded in the dictionary.
- **instantiates**: returns lexical concepts which the word form instantiates. In case a word is polysemous, multiple lexical concepts are returned.

```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxt": "άνθρωπος (1)",
6       "params": [
7         {
8           "name": "word_type",
9           "value": "single"
10        },
11        {
12          "name": "PoS",
13          "value": "No"
14        },
15        {
16          "name": "wordstring",
17          "value": "άνθρωπος"
18        },
19        {
20          "name": "number",
21          "value": "Sg"
22        },
23        {
24          "name": "gender",
25          "value": "Ma"
26        },
27        {
28          "name": "morphologically_related",
29          "value": [
30            "ανθρωπάκι",
31            "ανθρωπιτά",
32            "ανθρωπινά",
33            "ανθρωπινός",
34            "ανθρωπισμός",
35            "ανθρωπιστικός",
36            "ανθρωποειδής",
37            "ανθρωποκεντρικός",
38            "ανθρωποκεντρισμός",
39            "ανθρωπότητα",
40            "ολιγάριθμος"
41          ]
42        }
43      ]
44    }
45  ]
46 }

```

Figure 2: JSON output of a general Dictionary API search for the Greek word *άνθρωπος*.

The dictionary information structure contains only the existing parameters for each matching record depending in its PoS.

#### 4.2.2 Specific Searches

Moreover, the dictionary provides several specific search functionalities. These specific request types are used to perform faster searches and retrieve small amount of data. Such a specific functionality is the type search. In this case the URL path is:

*GET* [http://www.xanthi.ilsp.gr/apis/polytropon/word/<word\\_text>/<type>](http://www.xanthi.ilsp.gr/apis/polytropon/word/<word_text>/<type>)

where <word\_text> is the greek word we search for, and <type> must have one of the following values: synonyms, antonyms, morpho and instantiates.

For synonyms of the word *καλός* (kalós, =good) we perform the following request:

*GET* <http://www.xanthi.ilsp.gr/apis/polytropon/word/καλός/synonyms>

The API result is presented in Figure 3. In this use case, the service returns the word *αγαθός* (agathós, =naive) as synonym.



```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxt": "καλός (1)",
6       "type": "synonyms",
7       "words": [
8         [
9           "αγαθός (1)"
10        ]
11      ]
12    }
13  ]
14 }

```

Figure 3: An example of synonym search result.

For the same word καλός (kalós, =good), the antonym search is performed by the URL:

*GET <http://www.xanthi.ilsp.gr/apis/polytropon/word/καλός/antonyms>*

As expected, the API returns the code in the following figure.

```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxt": "καλός (1)",
6       "type": "antonyms",
7       "words": [
8         [
9           "κακός (1)"
10        ]
11      ]
12    }
13  ]
14 }

```

Figure 4: An example of antonym search result.

An example of a search for only instantiates of the word βαφή (vafí, =paint) is the following URL:

*GET <http://www.xanthi.ilsp.gr/apis/polytropon/word/βαφή/instantiates>*

This request results are presented in Figure 5.

```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxt": "βαφή (1)",
6       "type": "instantiates",
7       "words": [
8         [
9           "βάψιμο",
10          "χρωμάτισμα"
11        ]
12      ]
13    },
14    {
15       "wordtxt": "βαφή (2)",
16       "type": "instantiates",
17       "words": [
18         [
19           "βάψιμο",
20          "χρωμάτισμα"
21        ],
22        [
23          "μπογιά"
24        ]
25      ]
26    }
27  ]
28 }

```

Figure 5: An example of instantiate search result.

Finally, a search example for morphologically related words of the word καρδιά (karðjá, =heart) will be performed by the URL:

*GET <http://www.xanthi.ilsp.gr/apis/polytropon/word/καρδιά/morpho>*

Such a request will result a JSON code presented in Figure 6:



```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxt": "καρδιά (1)",
6       "type": "morphologically_related",
7       "words": [
8         [
9           "ekfrassi_tzortzi6_Class84",
10          "καρδιά (2)",
11          "καρδιαγγειακός",
12          "καρδιογράφημα",
13          "καρδιογράφος",
14          "καρδιογραφία",
15          "καρδιολογία",
16          "καρδιολογικός",
17          "καρδιολόγος",
18          "καρδιοπάθεια",
19          "καρδιοπαθής",
20          "μυοκάρδιο",
21          "περικάρδιο"
22        ]
23      ]
24    },
25    {
26      "wordtxt": "καρδιά (2)",
27      "type": "morphologically_related",
28      "words": [
29        [
30          "ekfrassi_tzortzi6_Class84",
31          "καρδιά (2)",
32          "καρδιαγγειακός",
33          "καρδιογράφημα",
34          "καρδιογράφος",
35          "καρδιογραφία",
36          "καρδιολογία",
37          "καρδιολογικός",
38          "καρδιολόγος",
39          "καρδιοπάθεια",
40          "καρδιοπαθής",
41          "μυοκάρδιο",
42          "περικάρδιο"
43        ],
44        [
45          "καρδιά (1)"
46        ]
47      ]
48    }
49  ]
50 }

```

Figure 6: JSON result for morphologically related word search.

In order to provide advanced functionality as a resource beyond a simple dictionary, we have also opted for implementing a collective search. As it has already been mentioned, concepts in the dictionary are linked via hyperonymy/hyponymy relations. Being a transitive relation, hyperonymy and its inverse relation, hyponymy, link lexical instances in an hierarchical order and under a single node. This information is deemed useful for many applications that call for reasoning and inference-making. In this regard, searches for words that are encoded as hyponyms or hypernyms of a selected word have also been enabled returning a chain of lexical instances that are linked via the hyperonymy/hyponymy relation at any depth.

This functionality is illustrated in the following URL that provides a repetitive search for hyponyms of the word “χρώμα” (chroma, =colour):

*GET <http://www.xanthi.ilsp.gr/apis/polytropon/word/χρώμα/hyponyms>*

The proposed search will return all words marked as hyponyms of the word ‘χρώμα’ and will return all entries denoting a colour, namely: “κόκκινο” (kokino, =red), “πράσινο” (prasino, =green), “καφέ” (kafe, =brown). The process will go on searching for hyponyms of every matching word in a depth of 2 levels by default from the initial word. The outcome is one or more lexical chains comprising the hyponyms of the searched word (Figure 7).

In case we want to perform a quicker search, we can set the depth level by submitting it in the URL request as a number, in our case:

*GET <http://www.xanthi.ilsp.gr/apis/polytropon/word/χρώμα/hyponyms/1>*



```

1 {
2   "error": false,
3   "message": [
4     {
5       "word": "χρώμα",
6       "level": 0,
7       "hypos": [
8         {
9           "word": "άσπρο χρώμα",
10          "level": 1,
11          "hypos": "No hyponyms"
12        },
13        {
14          "word": "χρυσό χρώμα",
15          "level": 1,
16          "hypos": "No hyponyms"
17        },
18        {
19          "word": "μπλε χρώμα",
20          "level": 1,
21          "hypos": "No hyponyms"
22        },
23        {
24          "word": "μαύρο χρώμα",
25          "level": 1,
26          "hypos": "No hyponyms"
27        },
28        {
29          "word": "παραλλαγή βασικού χρώματος",
30          "level": 1,
31          "hypos": "No hyponyms"
32        },
33        {
34          "word": "κόκκινο χρώμα",
35          "level": 1,
36          "hypos": "No hyponyms"
37        },
38        {
39          "word": "καφέ χρώμα",
40          "level": 1,
41          "hypos": "No hyponyms"
42        },
43        {
44          "word": "γκρι χρώμα",
45          "level": 1,
46          "hypos": "No hyponyms"
47        },
48        {
49          "word": "κίτρινο χρώμα",
50          "level": 1,
51          "hypos": "No hyponyms"
52        },
53        {
54          "word": "πράσινο χρώμα",
55          "level": 1,
56          "hypos": "No hyponyms"
57        },
58        {
59          "word": "μοβ_1 χρώμα",
60          "level": 1,
61          "hypos": "No hyponyms"
62        },
63        {
64          "word": "μπορντό",
65          "level": 1,
66          "hypos": "No hyponyms"
67        }
68      ]
69    }
70   ]
71 }

```

Figure 7: Hyponyms of word χρώμα in JSON format.

This search will provide only the words been marked as direct hyponyms of the word “χρώμα” (=colour) and them. The higher the number the heavier the search is, so a maximum of 3 level search has been proposed to be used. Similarly, the dictionary can be searched for all relations encoded in the dictionary, and for each entry hypernyms, allomorphs, synonyms, near-synonyms, antonyms, etymologically related and semantically related words can be retrieved. A set of pre-defined relations enables semantic searches. For this kind of search we use a URL of the following syntax:

*GET* [http://www.xanthi.ilsp.gr/apis/polytropon/word/<greek\\_word>/relations/<relation\\_string>](http://www.xanthi.ilsp.gr/apis/polytropon/word/<greek_word>/relations/<relation_string>)

The <relation\_string> string must use one of the values as presented on Table 2.



Relation string	Description	Relation string	Description
affliction	Disease that afflicts	has_parts	What are its parts
afflicted	Part of body that gets afflicted	is_part_of	Where do this part belongs to
affliction_entity	The one that gets afflicted	has_participant	Who is taking part
entity_affliction	He gets afflicted by	participates_in	Where is he participating
drives	What does he drives	treats	What diseases this doctor treats
driven_by	It is driven by	is_treated_by	Which doctor treats this disease
has_garment	He has that garment	practices	What does he practice
garment_used_by	From whom this garment is used for	is_practiced_by	Who practices it
has_agent	Who is the agent of this action	location_of	What takes place in this
is_agent_of	What is the action of this agent	takes_place	Where is this taking place in
has_habitat	What is the habitat of this person	workplace_of	Who works here
is_habitat_of	Whom is this habitat of	works_in	Where is he working here
has_bodypart	What are the body parts of this body part	used_by	What tool does he use
is_bodypart_of	Which body part contains this part	uses_tool	Who uses this tool
has_fruit	Which is the fruit that is being produced by	used_in	Which action uses this tool
is_fruit_of	Where is this fruit originated from	needs_tool	What tool is used in this action
has_colour	What is its color	effect	Effect
is_colour_of	Whom is this color	cause	Cause
has_member	What are its members	uses_vehicle	What vehicle does he use
is_member_of	Where do this member belongs to	vehicle_used_by	Who uses this vehicle
has_office	What is this person's office	has_property	What property does he have
is_office_of	Whom office is this	property_of	Whose property is this

Table 2: Pre-defined relation types used in API.

For example if we want to search the Greek word “ορνιθοτροφείο” (orniθotrofío, =poultry farm) with the “location\_of” relation and find as the description says (“What takes place in this”), we should perform the following URL:

*GET* [http://www.xanthi.ilsp.gr/apis/polytropon/word/ορνιθοτροφείο/relations/location\\_of](http://www.xanthi.ilsp.gr/apis/polytropon/word/ορνιθοτροφείο/relations/location_of)

This request will result the following JSON code (Figure 8) where the related word “εκτροφή” (ektrofí, =breeding) is returned from our dictionary.

```

1 {
2   "error": false,
3   "message": [
4     {
5       "wordtxtword": "ορνιθοτροφείο",
6       "relation": "location_of",
7       "related_words": [
8         "εκτροφή"
9       ]
10    }
11  ]
12 }
```

Figure 8: Relation type result in JSON format.



## 5 Conclusion

We have presented an infrastructure, currently under development, for exploiting a variety of lexical resources. Currently, the infrastructure will facilitate easy and robust access to a conceptual dictionary of MG. The service is targeted to researchers of MG and to application developers in need of linguistically aware lexical resources. In the future, we envisage the creation of an interface so that the resource is usable by end-users as well. In this context, we are planning to enable interactive access to the data and extending the resources and services via crowdsourcing. Finally, the integration of other lexical resources into the infrastructure is also planned; in this context, linking of the resource at hand with other language data will be examined.

## 6 References

- Fellbaum, C. (ed.). (1998). *WordNet: An electronic lexical database*. The MIT Press.
- Fielding, R. (2000). Architectural styles and the design of network-based software architectures, Ph.D. thesis, University of California.
- Fillmore, C., Wooters, C. & Baker, K. (2001). Building a Large Lexical Databank Which Provides Deep Semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*. Hong Kong. Accessed at: <http://framenet.icsi.berkeley.edu/papers/dsemlex16.pdf> [21 /06/ 2007].
- Fotopoulou, A., Giouli, V. (2015). From Ekfrasis to Polytropon: conceptual design of Lexical Resources. In *Proceedings of the 12th International Conference in Greek Linguistics (ICGL12)*. Berlin, Germany.
- Gale, W. A., Church, K.W. & Yarowsky, D. (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6),415.
- Gangemi A., Navigli R. & Velardi P. (2003). The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In R. Meersman, Z. Tari, D.C. Schmidt (eds) *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003*. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg.
- Guarino, N., Welty, C. (2004). An Overview of OntoClean. In S. Staab, R. Studer (eds.) *The Handbook on Ontologies*. pp. 151-172. Berlin:Springer-Verlag.
- Hullen, W. (2003). *A History of Rogets Thesaurus: Origins, Development, and Design*. Oxford University Press.
- Kipper-Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis, Computer and Information Science Dept., University of Pennsylvania, Philadelphia, PA.
- Mahesh, K., Nirenburg, S. (1995a). A situated ontology for practical NLP. In *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95)*, Montreal, Canada.
- Mahesh, K., Nirenburg, S. (1995b). Semantic classification for practical natural language processing. In *Proceedings of the 6th ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting*, Chicago, IL.
- Markantonatou S., Fotopoulou, A. (2007). The tool Ekfrassi. In *Proceedings of the 7th International Conference on Greek Linguistics, The Lexicography Workshop*, September 8-10, 2007, University of Ioannina, Greece [in Greek].
- Meyer, C. M., Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In S. Granger, M. Paquot (Eds.): *Electronic Lexicography*, Oxford: Oxford University Press, pp. 259–291.
- Navigli, R., Ponzetto, S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, pp. 217-250.
- Niles, I., Pease, A. (2001). Toward a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*.
- Pease, A. (2011). *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.
- Vostantzoglou, Th. (1962). *Antilexicon i Onomasticon tis neoellinikis*, Athens.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their comments that contributed to improving the manuscript. The research leading to the results presented here was partially funded by the project “Computational Science & Technologies: Data, Content and interaction” (MIS 5002437), which was co-financed by Greece and the EU, and the project “AIO IEL – Strategic Planning and Infrastructures of ILSP”.



# Principled Quality Estimation for Dictionary Sense Linking

Grosse J., Saurí R.

*Oxford University Press, United Kingdom*

## Abstract

Estimating the quality of lexical data automatically linked on the sense level is challenging, as the quality of the predicted sense links can differ significantly across various datasets. This variability is especially problematic when quality estimation is limited to general statements about an extensive collection of sense pairs, such as the links between two entire dictionaries. We argue that estimating probabilities for individual sense pairs is a superior method for quality estimation for two reasons: Firstly, it allows us to draw more nuanced conclusions about the quality of linked lexical data. Secondly, it opens the door for merging automated with manual means of sense linking by pointing lexicographers towards sense pairs that are especially difficult to classify. We propose a method for generating such probability estimates for a supervised machine learning approach. We show that these probabilities successfully dissect the sense pairs based on the certainty of the classification algorithm, thereby enabling lexicographers to analyse and improve the quality of automatically linked lexical data effectively.

**Keywords:** Word sense linking; language data construction; semi-automated annotation; data quality estimation; probability estimation

## 1 Introduction

Research into automating lexicographic processes for the creation of dictionary content has gained significant traction in recent years. This research has been partly facilitated by the European Union, which currently funds two large programmes in the area: ELEXIS (European Lexicographic Infrastructure Programme)<sup>1</sup> and Prêt-à-LLOD.<sup>2</sup> One of the goals of the programmes is to link linguistic linked data on the sense level<sup>3</sup> and make such data a useful tool for real-life problem-solving. To date, automatically sense-linked datasets, dictionaries in specific, have been a topic of research but are rarely seen in the production context. The obstacles preventing the full leveraging of automated sense linking boil down to the issue of quality: automatically linked dictionaries do not yet satisfy the high-quality expectations of the market. At the same time, linking dictionaries manually, while ensuring superior link quality, takes a long time and is often unfeasible financially.

One potential solution to this problem is to combine automated means with human expertise to create cost-efficient high-quality linked datasets. A necessary step towards such semi-automated linking, we argue, is to quantify the quality of individual linked sense pairs in terms of probability estimates, as such probability estimates allow us to distinguish certain sense links from uncertain ones. Accordingly, in this paper, we describe an algorithm estimating the link probability of previously predicted sense links. These probability estimates allow us to more accurately predict the quality of newly generated sense links between two dictionaries. Additionally, they enable the joint application of automated and manual means of sense linking.

## 2 Related Work

There exist various approaches to linking lexical content on the sense level, including using a pivot language (e.g. Kaji, Tamamura, & Erdenebat 2008; Tanaka & Umemura 1994; Varga & Yokoyama 2009; Wushouer et al. 2014), translation graphs with circles (e.g. Alper, 2017; Mausam et al. 2009; Villegas et al. 2016) or weights (e.g. Proisl et al. 2017), dictionary triangulation (e.g. Gollins & Sanderson 2001; Massó et al. 2013), neural machine translation (e.g. Arcan et al. 2019), and supervised machine learning (e.g. Donandt, Chiarcos, & Ionov 2017; Saurí et al. 2019). Notably, even though there has been groundbreaking recent progress in machine translation and natural language processing, the 1994 seminal paper by Tanaka and Umemura remains a highly competitive baseline for dictionary linking on the sense level (Gracia, Kabashi, & Kernerman 2019). This observation, paired with the fact that researchers have explored numerous approaches, is a testament to the difficulty and uniqueness of the problem of sense linking. Therefore, we must ask: How viable is automated sense linking? What is the level of quality we can expect for a given set of sense links?

The diversity of approaches to sense linking necessitates a comparable diversity of approaches in quality estimation. While some have settled with extrapolating quality audits of modest data samples, others have aimed to rank potential sense links by obtaining a proxy for the certainty of the respective linking algorithm's prediction: Villegas et al. (2016) assume that the density of translation cycles corresponds to the likelihood of a correct sense link; Mausam et al. (2009) probabilistically evaluate possible paths in a translation graph to estimate the percentage of complete cycles, which in turn serves as the certainty proxy; Arcan et al. (2019) simply take the output of the machine translation algorithm as the

<sup>1</sup> <https://elex.is/>

<sup>2</sup> <https://www.pret-a-llod.eu/>

<sup>3</sup> Note that linguistic linked data often refers specifically to data in the RDF format. Here, however, we mean a format agnostic interpretation of linguistic linked data as language data that is linked on the sense level.



confidence score; Massó et al. (2013) quantify the reliability of a sense link by counting the number of dictionaries included in possible triangulations; Shezaf & Rappoport (2010) obtain certainty in the form of a similarity score for a pair of words in source and target language.

The previous work on linking lexical content described above, despite vast variations in approach, faces a common challenge: differentiating between different degrees of certainty for candidate sense links. As shown above, all approaches aiming to address the challenge make use of a proxy for the certainty that a given sense pair is a sense link. The issue with the general method of estimating quality through proxies for certainty is that these proxies alone fall short of providing principled probability estimates. While the proxies are correlated with the underlying probabilities (after all this correlation is what makes the proxies valid in the first place), no previous work that we are aware of has convincingly quantified this correlation to estimate the underlying probabilities. Failing to obtain probability estimates means that, while (limited) general conclusions can be drawn about the quality of sufficiently large datasets, the actual probability of any individual predicted sense link being correct remains opaque. Furthermore, any conclusions drawn about the quality of a dataset based on a certainty threshold rely on the assumption that different datasets have an identical, or at least highly similar, distribution of certainty scores. However, due to the linguistic idiosyncrasies of languages, this assumption rarely holds. For instance, the degree of polysemy of words is (negatively) correlated with the certainty of prediction. This correlation means that we can expect languages with higher polysemy to result in less certain sense links than languages with lower polysemy. The lower certainty consequently results in lower data link quality, even if the certainty cut-off remains the same, rendering the original quality estimates imprecise.

### 3 Motivation

Supervised machine learning offers a promising opportunity for obtaining the desired probability estimates that solve the challenges mentioned above. For one, many machine learning algorithms, such as logistic regression, inherently output predictions in form of fractions in the unit interval. These outputs have widely been interpreted as probabilities. Even for machine learning algorithms where such an interpretation of the output is impossible or unprincipled, numerous calibration methods have been proposed for obtaining the underlying probabilities (e.g. Niculescu-Mizil & Caruana 2005). Therefore, machine learning approaches to sense linking lend themselves well to a natural quantification of link certainty, consequently allowing a ranking of sense pairs by link likelihood.

In our previous work, applying supervised machine learning to sense linking (Sauri et al. 2019), however, we are denied the straightforward ways of obtaining probability estimates. The reason lies within the dependency of the data. Defining the problem of sense linking as a binary classification task, thus making a machine learning approach feasible, entails assuming that each potential sense pair is independent of any other potential sense pair. This assumption is flawed; the link likelihood of any given sense pair depends, at least in part, on the connection the two senses have to other senses. Knowing that the two senses of a sense pair are already linked to other senses makes the sense pair less likely to be a link. Inversely, knowing that the two senses of a sense pair are not linked to any other senses makes the sense pair more likely to be a link. To account for the existing dependency, Sauri et al. (2019) employed a second algorithmic layer where the initial prediction of the machine learning algorithm is compared to a threshold quantifying the effect of the dependency across sense pairs. This additional classification layer significantly improves results. Unfortunately, it has the side effect that we can no longer take advantage of the probability estimates (pure or calibrated) that could be calculated in the first algorithmic layer because these estimates are tied to the assumption of independence between sense pairs. In other words, because the second algorithmic layer changes the predictions (therein accounting for the dependencies across sense pairs and improving results), the initial probability estimates lose validity. For instance, we find that a sense pair with an initial link probability of 0.3 may be classified a link because neither sense is linked to any other sense, while a sense pair with an initial link probability of 0.8 may be classified a non-link because both senses are already linked to numerous other senses.

This paper addresses these challenges by introducing an alternative method for obtaining principled link probabilities for Sauri et al. (2019).

## 4 Method

### 4.1 Starting Point

The algorithm for calculating probability estimates takes as input the results of the sense linker developed in Sauri et al. (2019). The results of the sense linker come in the form of a binary value for each sense pair: link or non-link. The sense linker predicts the binary category through a two-step process:

- a. The first level of the algorithm applies a machine learning based classifier, returning an independent prediction  $p$  for each sense pair.
- b. The second level of the algorithm determines the final classification (link or non-link) by comparing the prediction  $p$  from the first level to a threshold  $t$ . The threshold  $t$  is calculated for each sense pair using:
  - The number of senses that the two senses of the sense pair have already been linked to in the (respective) other dictionary, and
  - The total number of senses that the sense pair's lexeme has in either dictionary

Only if the prediction  $p$  exceeds the threshold  $t$  for a given sense pair is the sense pair classified as a link; else it is classified as a non-link. Given the sense linker's binary classification as well as the initial prediction  $p$  and the threshold  $t$ , we are tasked with calculating the link probability. Calculating this probability is complicated by the two-level structure described above. The issue is that the prediction  $p$  from the first level does not provide clear cut-offs for when a link is



made. In fact, because of the threshold  $t$  from the second level, there exist sense pairs that are not linked despite having a higher prediction  $p$  in the first level than other sense pairs that are linked. For example, a sense pair with an initial prediction  $p=0.8$  and a threshold  $t=0.9$  is not linked while a sense pair with an initial prediction  $p=0.3$  and a threshold  $t=0.2$  is linked.

## 4.2 Measuring Certainty

A proper measure of the certainty of prediction must be strictly monotonic regarding the actual certainty of the algorithm (i.e. the predicted link-likelihood). That is, a higher measure directly translates to higher certainty, and a lower measure directly translates to lower certainty. Also, and as a result of the first condition, there must be a clear, numeric cut-off point for creating a binary prediction. For instance, every sense pair with a certainty score equal to or larger than  $x$  is labelled a link while every sense pair with a certainty score lower than  $x$  is labelled a non-link (where  $x$  falls within the range of the certainty score).

The initial prediction  $p$  meets neither of these conditions because it fails to account for the role of the threshold  $t$  in deciding the final prediction. Therefore, we chose to quantify the certainty instead by looking at the distance of the initial prediction  $p$  from the threshold  $t$ , because this distance directly translates to how *close* a sense pair is to being labelled a link or non-link. Further, this measure satisfies the condition of being directly correlated to the certainty of the algorithm, with a clear cut-off point for the different predictions. We thus have the certainty score  $c$ :

$$c = p - t$$

Because the ranges of both  $p$  and  $t$  are  $[0, 1]$ , the certainty score  $c$  ranges from  $[-1, 1]$ . Values toward -1 correspond to high certainty of a non-link, values toward 1 correspond to high certainty of a link, and values around 0 correspond to low certainty. Importantly, 0 is the true cut-off for the prediction. That is, a positive score for certainty  $c$  indicates a predicted link while a negative score indicates a predicted non-link.

After obtaining the estimated certainty value  $c$ , the next step is to translate this measure to percentage-wise probability estimates for the outcome. That is, beyond a certainty measure, we want the underlying probability that a given sense link is correct.

## 4.3 Bucket approach

To estimate the link probability of a sense pair, we take advantage of the large amount of annotated data we used for training the predictive algorithm. Using the annotated data, we can compare a new sense pair to annotated sense pairs that are similar. That way, we can estimate the probability outcome for the new sense pair based on the results of the similar sense pairs. Similarity, in this case, is defined across the certainty axis (i.e. the value of  $c$ ). That is, we consider sense pairs to be similar if they received a comparable certainty score  $c$  by the predictive algorithm.

The certainty variable  $c$  is divided into several ranges, or buckets, to choose similar sense pairs. To estimate the probability that a given new sense link is correct, we take the relative frequency of true, annotated sense links among all annotated sense pairs in the same respective bucket. For instance, consider a new sense pair with a certainty score of 0.78. If we chose to divide the certainty variable  $c$  into 20 equidistant (evenly spaced) ranges, or buckets, then the relevant bucket for the new sense pair is the range  $[0.7, 0.8)$ . Therefore, to estimate the link probability of the new sense pair, we calculate the percentage of true links among the annotated sense pairs with a certainty score  $c$  between 0.7 and 0.8. This percentage serves as the link probability of the new sense pair.

In our previous work (Saurí et al., 2019), we noted a sizeable variability in data distributions across the different parts of speech, leading us to divide the dictionaries into sense pairs by part of speech. Since this observation also affects the predictions, we again consider each part of speech separately in applying the bucket approach to estimating link probabilities. Figure 1. shows the distribution of the sense pairs regarding the certainty score  $c$  for each part of speech when dividing  $c$  into 10 equidistant buckets (evenly spaced ranges). It also visualises the bucket approach: Each bar is a bucket, and the proportion of true links quantifies the link probability for a new sense pair falling into the bucket.



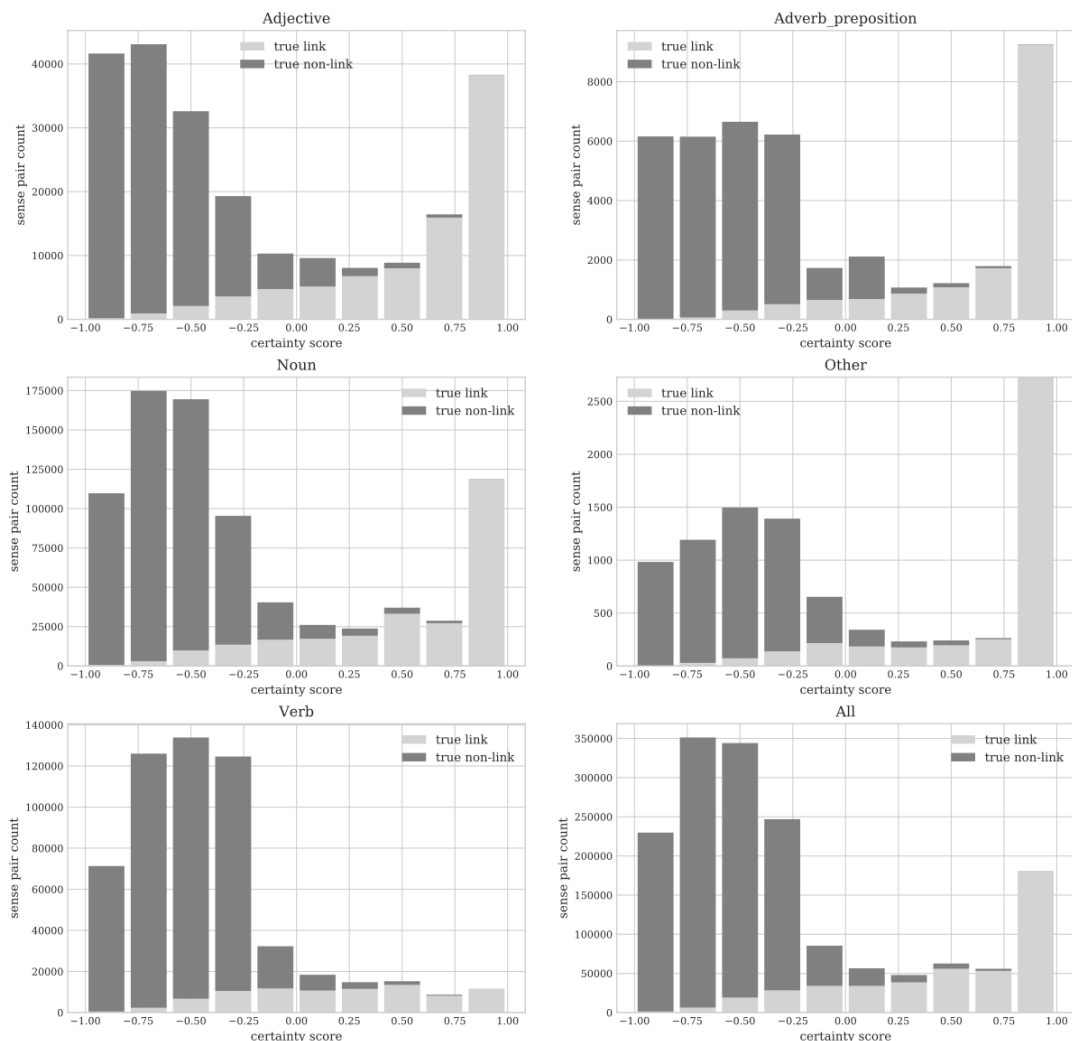


Figure 1: Distribution of sense pairs according to certainty score by part of speech using 10 equidistant buckets

The accuracy of link probabilities estimated with the bucket approach crucially depends on the choice of bucket sizes. When a bucket is too broad, new sense pairs that have considerably unlike certainty scores receive the same probability estimate. The resulting insufficiently fine discrimination of certainty scores is an example of under-fitting. That is, by literally throwing too many sense pairs into the same bucket, we fail to adequately represent the underlying differences in link probabilities that exist in the data. On the flip side, when a bucket is too narrow, the number of annotated sense pairs may be too small, thus rendering the relative frequency measures unrepresentative as well. A scenario exemplifying such kind of overfitting is a sudden, yet random, drop in the link probability for a bucket with a range comprising high certainty scores.

Finding the proper bucket sizes is an empirical matter. Therefore, we experimented with several ways of partitioning the certainty score  $c$  into buckets. One approach, coined *equidistant buckets*, divides the certainty score  $c$  into several buckets that each cover a range of the same size. For instance, a division into 20 equidistant buckets means that the first bucket includes all the links with a certainty score  $c$  in the range  $[-1, -0.9)$ , the second bucket includes all the links with a certainty score  $c$  in the range  $[-0.9, -0.8)$ , and so on to the last bucket covering the range  $[0.9, 1]$ . This approach has the advantage of being straightforward and independent of the training data. The width of each bucket is impervious to the distribution of the training data. The disadvantage of the approach is that number of links per bucket can vary hugely between buckets, as some certainty scores may occur more frequently than others. As shown in Figure 1., the distribution of the data is heavily skewed towards extreme certainty scores (close to -1 or close to 1) for every part of speech. This skew has the effect that buckets near the extremes have more sense pairs than the buckets near a certainty score of 0, which biases the relative frequency measures in the different buckets.

The second approach, coined *quantile buckets*, divides the certainty score  $c$  into several buckets that each contain the same number of sense pairs. For instance, if there are 2000 sense pairs in total and we set the number of buckets to 20, then the quantile approach involves first finding the cut-off points for the buckets across  $c$  such that each bucket covers 100 sense pairs. These cut-off points then define the extent, or width, of each bucket. As a result, the buckets cover ranges of variable sizes but containing the same number of sense pairs. The advantage of this approach is that there is no risk of



disproportionate buckets since, by definition, every bucket has the same number of links. The main disadvantage is that the determination of the buckets is dependent on the distribution of the training data, which may or may not correspond to the distribution of the data that the approach is applied to.

Figure 2. visualises the two different approaches, with all sense pairs with the part of speech *verb* divided into six equidistant buckets (left) and six quantile buckets (right) for comparison. Note that the bucket width is constant for equidistant buckets and varies for quantile buckets, while the number of sense pairs per bucket varies for equidistant buckets and is constant for quantile buckets.

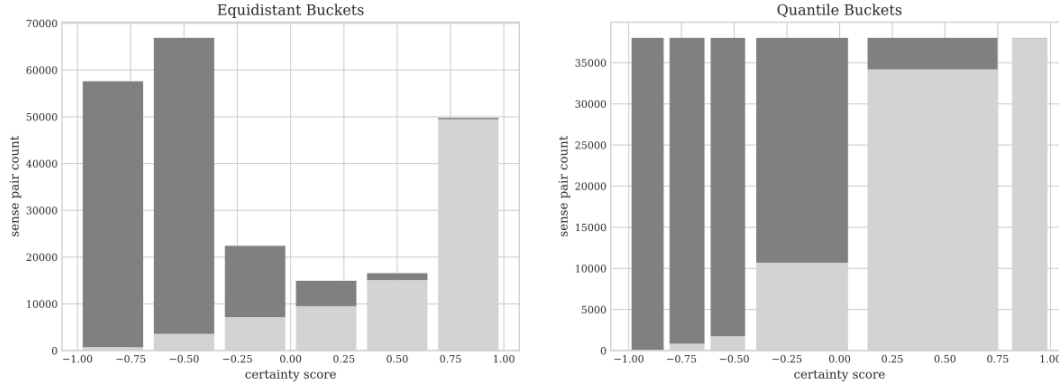


Figure 2: Comparison of equidistant buckets and quantile buckets using the example of *verb* sense pairs divided into six buckets

We further experimented with the absolute number of buckets. For every part of speech, we tried out 10, 50, 100, 500 and 1000 equidistant and quantile buckets.

#### 4.4 Validation

The ‘goodness of fit’ of each bucket approach variety was quantified by calculating the logistic likelihood ratio (with base 10) for the resulting sense link probabilities:

$$LLR_{10} = N \log_{10} 2 + \sum_1 \log_{10} p_i + \sum_0 \log_{10} (1 - p_i)$$

$N$  is the number of sense pairs in the entire dataset, and  $p$  is the predicted link probability for a given sense pair. The summations are for all true links (1) and all true non-links (0) respectively. Due to a large number of sense pairs, the resulting likelihood ratios blow out of proportion, which is why we additionally calculated the average likelihood ratio for one sense pair as follows:

$$avgLR = 10^{LLR_{10}/N}$$

$avgLR$  ranges from (0, 2] and represents how much better or worse the predictions are (on average) compared to chance. Chance is defined as a 50-50 guess, a flip of a fair coin, to make the prediction. An  $avgLR$  score of 1 implies that the probability estimates are no better (or worse) than chance. Optimal predictions, where every link is predicted with  $p = 1$  and every non-link is predicted with  $p = 0$ , result in an  $avgLR$  score of 2, meaning that, for each sense pair, the prediction was twice as good as chance. This upper-bound is logical, as pure chance classifies a sense pair correctly with probability  $p = 0.5$  and we cannot do better than probability  $p = 2 * 0.5 = 1$ .

Random guessing is a low bar to compare our probability estimates against, especially since the algorithm in Sauri et al. (2019) does the heavy lifting of prediction. Therefore, we also compare the results against a well-informed baseline. The baseline is calculated by taking the true positive rate of all predicted links in the training data as the probability  $p$  for all predicted links, and the false negative rate of all predicted non-links in the training data as the probability  $p$  for all predicted non-links. For instance, if in the training data 84% of all predicted links were true links and 12% of all predicted non-links were true links, then the link probability estimates are 0.84 for predicted links and 0.12 for predicted non-links. Note that we can expect the baseline to perform much better than chance, as it takes into account the prediction of the algorithm. That is, the baseline probabilities are informed by what the algorithm has learned about sense pairs and sense links. Note further that this baseline is equivalent to the bucket approach with two equidistant buckets (one for predicted links and one for predicted non-links), and therefore can be seen as the most basic version of the approach proposed in this paper.

The bucket approach was cross-validated using each of the six dictionaries used in the training of the binary classifier from Sauri et al. (2019) as one fold. That is, five dictionaries were used to create the buckets and calculate the relative frequencies of true links. These relative frequencies were then used to estimate the link probability of every sense pair in



the sixth dictionary. For each part of speech separately, we identified the best performing bucket approach as the one that had the highest mean *avgLR* score across the six dictionaries. Table 1. summarises the cross-validation results, showing the baseline as well as the best performing bucket approach for each part of speech.

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline ( <i>avgLR</i> )	1.543	1.605	1.530	1.524	1.531
Best buckets ( <i>avgLR</i> )	<b>1.656</b>	<b>1.710</b>	<b>1.638</b>	<b>1.634</b>	<b>1.614</b>
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 1: Average Likelihood Ratio (*avgLR*) for the baseline and the best performing bucket approach for each part of speech

As expected, the baseline performs reasonably well, predicting a sense pair on average around 1.5-1.6 times better than chance. For every part of speech, the best performing bucket approach outperforms the baseline, with an *avgLR* score consistently higher at around 1.6-1.7. Thus, cross-validation confirms that the bucket approach does enable us to generate more precise probability estimates for each sense pair individually. Quantile buckets emerged as the best approach only for the part of speech *other*, and equidistant buckets for all other parts of speech. The optimal number of buckets ranges from 50 to 500, plausibly with some correlation to the size of the dataset (e.g. *verb* has many more sense pairs than *other*).

## 5 Results

The sense linker developed in Sauri et al. (2019) in conjunction with the algorithm presented in this paper were used to link an English-Arabic bilingual dictionary and a Portuguese-English bilingual dictionary to a monolingual English dictionary. For both Arabic links and Portuguese links, we took a random sample of around 1000 sense pairs for each part of speech and obtained the gold standard from expert lexicographers. Table 2. and Table 3. show the results of the external validation using the random samples compared to the baseline for Arabic and Portuguese links, respectively. Recall that an *avgRL* score of 1 represents performance equivalent to chance, while an *avgRL* score of 2 corresponds to perfect predictions (i.e. links with a probability of  $p = 1$  and non-links with a probability of  $p = 0$ ). The baseline, again, is equivalent to the bucket approach with two equidistant buckets.

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline ( <i>avgLR</i> )	<b>1.419</b>	1.454	1.557	1.154	1.545
Best buckets ( <i>avgLR</i> )	1.319	<b>1.500</b>	<b>1.584</b>	<b>1.260</b>	<b>1.598</b>
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 2: Average Likelihood Ratio (*avgLR*) results compared to the baseline for a random sample of Arabic links

Part of speech	Adjective	Adverb / Preposition	Noun	Other	Verb
Baseline ( <i>avgLR</i> )	1.383	1.477	1.470	1.272	1.479
Best buckets ( <i>avgLR</i> )	<b>1.437</b>	<b>1.528</b>	<b>1.563</b>	<b>1.304</b>	<b>1.537</b>
Best bucket approach	500 equidistant	100 equidistant	100 equidistant	50 quantile	500 equidistant

Table 3: Average Likelihood Ratio (*avgLR*) results compared to the baseline for a random sample of Portuguese links

The improvements over the baseline are more modest than was the case with cross-validation. This result was to be expected since the optimal bucket approach was chosen during cross-validation, while only the bucket approach identified during cross-validation was used for external validation. For the links of the Portuguese-English bilingual dictionary, the previously determined optimal bucket approach consistently outperforms the baseline, albeit to a lesser extent than during cross-validation, with a mean *avgLR* score improvement close to 0.05 across the different parts of



speech. For the links of the Arabic-English bilingual dictionary, we obtain similar results for four out of the five parts of speech. However, for adjectives, the baseline performs better than the previously determined optimal bucket approach on the random sample. Potential reasons are addressed in the Discussion section.

## 6 Discussion

### 6.1 Interpretation of Results

The results show that, using the bucket approach described in this paper, we can obtain sense link probability estimates for individual sense pairs that outperform more general approaches like the baseline. The part of speech group *other* proves most challenging to predict. A potential reason may be that *other* subsumes many different parts of speech (e.g. conjunctions, pronouns, determiners), which may not generalise well to different languages. Another possible explanation is simply that *other* has the fewest samples (i.e. sense pairs) in the training data, thus making the relative frequency measures used to estimate probabilities less reliable. Another noteworthy observation is that the baseline performs better than the more nuanced bucket approach (only) for the part of speech group *adjective* in the English-Arabic dictionary. This exception suggests that the probability estimates for Arabic adjectives were overly optimistic (as false predictions with high certainty are punished disproportionately, strongly affecting the outcome), possibly because the features that are highly indicative of sense links for adjectives in general are less relevant for Arabic adjectives in specific. Of course, given the nature of using random samples for validation, fluctuations in performance may also, to a certain degree, be explained by randomness.

### 6.2 Implications for Automated Sense Linking

To make use of automated means of sense linking in practice, the ability to quantify the quality of the linked data is paramount. Most approaches to automated sense linking rely heavily on general quality estimates for entire datasets. Such estimates are often unreliable, as they do not generalise well to other datasets. One example proving this notion comes from the stratification by part of speech. As shown in Figure 1. in the Method section, and likely due to high polysemy, *verb* sense links are much harder to predict than those of other parts of speech. Accordingly, quality estimates for all parts of speech are not representative of, for instance, the quality of a linked dataset consisting only of verbs.

Some automated sense linking approaches attempt to make more nuanced statements regarding link quality by introducing thresholds along the certainty axis. They may say, for instance, that sense pairs above one certainty threshold have one estimated precision, while sense pairs above another certainty threshold have different estimated precision. The issue here is the implicit assumption that the distribution of certainty scores remains the same across different datasets. Using the example of *verb* sense pairs again, we can see that this is a false assumption: *Adjective* sense pairs with a certainty score  $c > 0.25$ , for instance, have higher precision than *verb* sense pairs with a certainty score  $c > 0.25$  because there are many more *adjective* sense pairs with a certainty score  $c$  close to 1 than *verb* sense pairs.

In other words, general quality statements are not sufficient as they ignore the differences in distribution across different datasets. Probability estimates, on the other hand, circumvent this issue as they provide estimates for each sense pair independently, thereby being impartial to the distribution of the larger dataset. A further benefit of generating probability estimates for individual sense pairs is that quality estimation on the sense pair level opens up the opportunity of integrating automated sense linking in lexicographic processes without imprudently relinquishing the expertise of lexicographers. By linking sense pairs with high link-probability automatically while manually reviewing sense pairs with high uncertainty, it is possible to improve the quality of automatically generated sense links with limited editorial resources. Quality estimation, therefore, must not be seen merely as a validation mechanism for a given approach, but rather as a tool for making possible high-quality, (semi-)automated sense linking.

Probability estimates provide substantial benefits that help drive the implementation of automated sense linking in the applied, industrial context.

### 6.3 Implications for Lexicography

The work presented also has important implications for lexicography more generally. As mentioned above, the ability to quantify the probability that an automatically linked sense pair is indeed a link allows us to merge automated and manual sense linking. The sense pairs that can be automatically linked with high probability can be considered links without further inspection, while sense pairs automatically linked with low certainty (i.e. a probability near 0.5 for the binary task of sense linking) can be passed on to lexicographers for expert validation. This method can significantly speed up the process of linking lexical resources while retaining control over the quality of links. Precise probabilities also enable us to quantify the trade-off between the quality of the links and the amount of work needed from lexicographers. In that way, lexicographers can gain direct insights into how their work improves the quality of automatically linked lexical data, and base decision-making on the quality estimates.

Additionally, probabilistic automatic sense linking can point lexicographers to differences, incongruities, and omissions across lexical resources. For instance, when a sense in one resource has no corresponding sense in another resource (i.e. all relevant link probabilities are small), we may infer that in the latter resource the sense is (correctly or mistakenly) excluded or subsumed in another sense. Inversely, when a sense in one resource is linked with high probability to several senses in another resource, we have reason to believe that the latter resource applies a finer level of granularity. Investigating such cases can be useful for lexicographers when revising and expanding (linked) lexical resources.

Ultimately, lexicographers provide the gold standard for supervised machine learning based approaches to automated sense linking. As such, all linked data annotated by lexicographers is useful in improving the algorithms' outcomes by



providing more training data. Furthermore, inspecting sense pairs of interest may provide insights into how to improve a given algorithm or, in the very least, point out its shortcomings. Sense pairs of special interest may be false positive predictions that have a high probability score, reversely false negatives that have a low probability score, and sense pairs with high uncertainty.

## 7 Conclusion

This paper has described a method for obtaining probability estimates for predicted sense links between two dictionaries. These probabilities play an essential role not only in reliably estimating the quality of a given set of sense links but also in differentiating sense links with high certainty from those with low certainty. This differentiation makes possible the deliberate trade-off between the correctness and the completeness of generated links, as well as the optimal improvement of link quality through limited editorial input by lexicographers. Both reliable quality estimation and semi-automated sense linking are vital points in moving automated sense linking from a research interest to a content production tool. As part of the broader efforts of linking lexical content at Oxford University Press, the presented quality estimation algorithm achieves precisely that.

## 8 References

- Alper, M. (2017). Auto-generating Bilingual Dictionaries: Results of the TIAD-2017 Shared Task Baseline Algorithm. In *Proceedings of the LDK 2017 Workshops, co-located with the 1st Conference on Language, Data and Knowledge*, pp. 85–93.
- Arcan, M., Torregrosa, D., Ahmadi, S., & McCrae, J. P. (2019). Inferring translation candidates for multilingual dictionary generation with multi-way neural machine translation. Paper presented at the *Translation Inference Across Dictionaries Workshop (TIAD 2019)*, Leipzig, Germany, 20-23 May, doi:10.13025/S89K9J
- Donandt, K., Chiacos, C., & Ionov, M. (2017). Using Machine Learning for Translation Inference Across Dictionaries. In *Proceedings of the LDK 2017 Workshops*.
- Gollins, T., & Sanderson, M. (2001). Improving Cross Language Retrieval with Triangulated Translation. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 90–95. ACM.
- Gracia, J., Kabashi, B., & Kernerman, I. (2019) *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*. Co-located with the 2nd Language, Data and Knowledge Conference (LDK 2019) Leipzig, Germany, May 20, 2019. CEUR Workshop Proceedings, Vol. 2493. <http://ceur-ws.org/Vol-2493/>
- Kaji, H., Tamamura, S., & Erdenebat, D. (2008). Automatic Construction of a Japanese-Chinese Dictionary via English. In *LREC 2008*.
- Massó, G., Lambert, P., Rodríguez-Penagos, C., & Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation. In R. E. Banchs, F. Silvestri, T. Liu, M. Zhang, S. Gao, and J. Lang, editors, *Information Retrieval Technology*, pp. 263–271.
- Mausam, Soderland, S., Etzioni, O., Weld, D., Skinner, M., & Bilmes J. (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 262–270. ACL.
- Niculescu-Mizil, A., & Caruana, R. (2005). Obtaining calibrated probabilities from boosting. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI'05)*. AUAI Press, Arlington, Virginia, USA, pp. 413–420.
- Proisl, T., Heinrich, P., Evert, S., & Kabashi, B. (2017). Translation Inference across Dictionaries via a Combination of Graph-based Methods and Co-occurrence Stats. In *LDK Workshops*
- Saurí, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. *LDK*.
- Shezaf, D., Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*. Association for Computational Linguistics, USA, pp. 98–107.
- Tanaka, K., & Umemura, K. (1994). Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of COLING'94*, pp. 297–303, 1994.
- Varga, I., & Yokoyama, S. (2009). Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of EMNLP*, pages 862–870. URL: <http://www.aclweb.org/anthology/D09-1090>.
- Villegas, M., Melero, M., Bel, N. & Gracia, J. (2016). Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In *Proceedings of LREC 2016*, pp. 23–28.
- Wushouer, M., Lin, D., Ishida, T., & Hirayama, K. (2014). Pivot-Based Bilingual Dictionary Extraction from Multiple Dictionary Resources. In *PRICAI 2014: Trends in Artificial Intelligence*, pages 221–234, Cham, 2014. Springer International Publishing.



# Determining Differences of Granularity between Cross-Dictionary Linked Senses

Kouvara E., Gonzàlez M., Grosse J., Saurí R.

Oxford University Press, United Kingdom

## Abstract

Linking dictionaries at the sense level is highly beneficial because it facilitates the mutual enhancement of the linked datasets or the possibility of deriving new products from the combination of the two. However, one of the greatest challenges in cross-dictionary sense linking is that linked senses, although referring to the same meaning, may actually differ in their semantic extent due to dictionary distinctions of sense granularity. Not every pair of linked senses is therefore qualitatively the same. However, being able to identify and classify these differences is a crucial step towards enabling the comprehensive exploitation of sense-linked datasets. In this paper, we present a system to automatically identify the relation of sense links between a bilingual and a monolingual dictionary. Using sense granularity annotations by lexicographers as the gold standard, we trained a machine learning model to classify the relation between cross-dictionary linked senses as one of the following categories: *perfect*, where each sense fully covers the other sense; *wider/narrower*, where one sense fully encloses the other but not vice versa; *partial*, where each sense partially covers the other sense. Cross-validation shows the machine learning model to yield an overall accuracy of 86%, with a macro precision of 83% and a macro recall of 65% across the different classes. The model significantly outperforms a rule-based algorithm serving as the baseline.

**Keywords:** Sense granularity, word sense linking, word sense mapping, lexical resources, language data generation, multilingual data, data integration across languages

## 1 Introduction

This paper presents XD-SemOver (Cross-Dictionary Semantic Overlap Classifier), a machine learning-based tool for automatically predicting the type of the relationship between two senses from different dictionaries that have been linked as referring to the same meaning.

A sense link consists of a pair of senses, each from one of the linked dictionaries, belonging to the same lexeme and referring to the same meaning. Different dictionaries, however, may diverge in how they split the meaning of the words into different senses, that is, in the sense granularity criteria they apply. As a result, the two senses in a sense link may not be fully equivalent; rather, the following can hold:

- *Sense inclusion.* One sense has a wider (or narrower) semantic extent than the other. That is, one sense fully includes the other one.
- *Partial overlap.* The two senses overlap to some degree, but each of them includes a meaning component which is not covered by the other.
- *Perfect match.* The two senses refer exactly to the same meaning.

These distinctions are essential to ensure accuracy and completeness of information when inferring new multilingual data from chains of senses across languages. Otherwise, there is the risk of omitting important information that may be present in one of the senses, but not on the other. Assume, for example, the following chain of equivalent senses  $S_{tA} - S_{tB} - S_{tC}$ , as presented in the dictionaries for languages A, B and C. Each of these senses is associated with a term in its corresponding language:  $t_A$ ,  $t_B$ , and  $t_C$ . If sense  $S_{tA}$  for term  $t_A$  in language A has a narrower semantic extent than sense  $S_{tB}$ , and  $S_{tB}$  is also semantically wider than  $S_{tC}$ , then it is not possible to safely infer that term  $t_C$  in language C is a translation of term  $t_A$  in language A.

This paper presents a component for automatically classifying differences in *sense granularity* between two senses from different dictionaries that refer to the same meaning, thereby establishing how the senses diverge with regards to their *semantic extension*. It has been developed as a component in a suite of tools for rich sense linking across dictionaries, which also includes: the sense linking system described in Saurí et al. (2019), a quality estimator tool that generates certainty values for automatically linked pairs (Grosse & Saurí 2020), and an annotation tool supporting the creation of manually labelled data for developing those resources (Gonzàlez, Buxton & Saurí 2020). On top of these tools, we are currently developing a system for automatically inferring new bilingual content from cross-dictionary sense links.<sup>1</sup> Thus, our purpose is to take previous work on sense linking one step further, therefore contributing to a higher quality for new bilingual content generation.

The paper is structured as follows. Section 2 reviews previous work that relates to the one presented here. Then, section 3 presents the overall project methodology, which is further detailed in sections 4 to 6. Results are discussed in section 7. Section 8 closes with final remarks and suggestions for future work.

<sup>1</sup> This project as a whole is part of a wider programme aiming to create multilingual data by combining Linguistic Linked Open Data and language technologies (Prêt-à-LLOD: <https://www.pret-a-llod.eu/>).



## 2 Related Work

The goal of this project is to determine the semantic overlap between two dictionary senses that refer to the same meaning, which directly touches on the notion of word polysemy and one of its central issues, namely, sense granularity. The critical question is how to split the different uses of a polysemous word into discrete senses. Within the lexical semantics field, much discussion has revolved around the notion of the meaning of a word and its analysis into different senses (e.g. Apresjan 1973; Cruse 1986; Pustejovsky 1995; Hanks & Pustejovsky 2005; Hanks 2013). From the more applied view of lexicography, the criteria for splitting senses as part of the dictionary creation process has also been an important topic of debate. Among many others, see for instance, Shcherba 1940/1995; Stock 1984; Wierzbicka 1985; Geeraerts 1990; Sinclair 1991; Fillmore & Atkins 1994; Kilgariff 1997; Atkins & Rundell 2008.

At the computational arena, sense granularity has an impact in at least two areas of work: word sense disambiguation (WSD) on the one hand, and sense linking (or alignment) on the other. Within WSD, a task that in the past decades has been mostly tackled using machine learning-based approaches, sense granularity poses challenges when a word is split into excessively granular senses. The greater the granularity, the higher the complexity of the system's learning job. One solution comes from clustering senses to obtain more coarse-grained sense inventories (Navigli 2006; Navigli et al. 2007; Cinková, Holub & Križ 2012). More recently, a data-driven approach grounded in distributional semantics (e.g. Erk 2010) and which is now benefiting from recent advances in deep learning (e.g. Pilehvar & Camacho-Collados 2019; Breit et al. 2020), has been moving away from defining senses a priori, thus circumventing the problem of sense granularity.

Within the computational framework, sense granularity also poses challenges to the activity around automatic sense linking (or alignment), which is precisely the area of work of our project. The overall goal of this area is to connect lexical databases (including dictionaries) at the sense level. Doing so enriches lexical content and enables the creation of new resources. There has been much research on this topic over the past two decades. Some work has been applied to the linking of lexical knowledge bases such as WordNet (Miller 1998) or Wikipedia,<sup>2</sup> which as opposed to traditional dictionary datasets, organise their content in a graph-based structure representing lexical relations among senses (Gurevych, Eckle-Köhler & Matuschek 2016). Other sense linking approaches have focused on word overlap, i.e. the number of common words among sense definitions (Ponzetto & Navigli 2010), while others have taken advantage of similarity distance measures (Ruiz-Casado, Alfonseca & Castells 2005; Ahmadi, Arcan & McCrae 2019). More recently, other approaches have explored using machine learning techniques to align senses between two dictionaries (Saurí et al. 2019).

Sense linking (or alignment) has also taken place across lexical resources of different languages in order to support cross-lingual information retrieval tasks (e.g., Gollins & Sanderson 2001; Massó et al. 2013) or to facilitate the rapid creation of new bilingual or multilingual lexicons, thus touching on the area of lexical translation (Varga, Yokoyama & Hashimoto 2009; Mausam et al. 2009; Wushouer et al. 2014; Villegas et al. 2016; Ordan et al. 2017; Gracia et al. 2019, among others).

Most of those approaches to sense linking, however, do not take into account the issue of sense granularity; namely, the possibility that the senses from two different sources that have been aligned as corresponding to the same meaning differ in their semantic extension (i.e. one has a wider or narrower semantic reference than the other). As argued in the introduction, these distinctions are critical because when generating new content from the alignment of two senses, there is a risk of leaving out crucial semantic information that is only present in one of the two senses. Nevertheless, to the best of our knowledge, there is only one sense linking project that has modelled these distinctions as part of the knowledge to be learnt: the ELEXIS Monolingual Word Sense Alignment Shared Task.<sup>3</sup> Aware of the relevance of this information, our sense linking project also includes a component for automatically identifying these distinctions. To date, however, there is no information published on the ELEXIS task for us to compare approaches or results.

## 3 Methodology

We developed XD-SemOver from a supervised machine learning approach following the standard methodology:

1. Delimiting the problem, i.e. determining the relevant distinctions that need to be learnt by the classifier.
2. Developing the dataset to be used for training and testing the system. For this, lexicographers annotated sense links using the categories determined in the previous step. The annotation was carried out using XD-AT, a Cross-Dictionary Annotation Tool supporting the manual categorisation of differences in sense granularity between two linked senses (González, Buxton & Saurí 2020).
3. Training and evaluating a classifier. We experimented with several models, namely a boosted decision trees algorithm, a multi-layer perceptron, and a pre-trained BERT-model. We also evaluated different settings (i.e. data balancing and parameter tuning) to identify the best approach.

The following sections detail the development carried out towards each of these aspects.

## 4 Delimiting the Problem

We distinguished the different types of sense links based on two kinds of relationships that may hold:

<sup>2</sup> <https://www.wikipedia.org/>

<sup>3</sup> [https://competitions.codalab.org/competitions/22163#learn\\_the\\_details-overview](https://competitions.codalab.org/competitions/22163#learn_the_details-overview)



*Overlapping*: Determined by the extension of the meaning expressed by sense  $S_A$  that aligns with sense  $S_B$ . It can be:

- *Full*:  $S_A$  fully overlaps with  $S_B$  when  $S_A$  expresses the entire meaning of  $S_B$ .
- *Partial*:  $S_A$  partially overlaps with  $S_B$  when  $S_A$  only expresses part of the meaning of  $S_B$ .

*Enclosing*: Determined by whether  $S_A$  covers the entire extension of sense  $S_B$ . It can be:

- *True*:  $S_A$  encloses  $S_B$  if  $S_A$  covers the entire extension of  $S_B$ .
- *False*:  $S_A$  does not enclose  $S_B$  if  $S_A$  covers not all but only part of  $S_B$ .

These two levels of description can be orthogonally combined, generating a 4-fold distinction: *perfect*, *narrower-than*, *wider-than*, and *partial*. *Perfect* indicates that each sense aligns completely throughout the full extension of the other one. In other words, each sense fully covers the other. In contrast, *narrower-than* and *wider-than* account for sense pairs where a sense in one dictionary has a broader meaning than the sense in the other dictionary. It occurs when the meaning of one sense fully overlaps with the other one but does not fully enclose it (*narrower-than*), or the other way around (*wider-than*). Finally, *partial* denotes that each sense extends beyond the reference of the other. In this case, each sense includes a meaning that is not covered by the other. Table 1 illustrates the four relations that can occur between two linked senses. For the annotation task, we used these four sense link type classes.

The idea and the vocabulary for this distinction relate to the way the Simple Knowledge Organization System (SKOS) (Miles & Bechhofer 2009) expresses exact or fuzzy matching of concepts from one scheme to another using broader-narrower, or associative relationships. They also show strong similarities to the 5-fold categorisation in the ELEXIS Monolingual Word Sense Alignment Task (McCrae 2020), where the categories used are: *exact*, *broader*, *narrower*, *related*, or *none*. In our case, the *none* category is irrelevant, because we omitted from the classification process any sense pairs that are not linked.

		Perfect match	Different sense granularity		Different sense boundaries
Meaning alignment					
Grounding relationships					
	$S_A$ overlapping with $S_B$	fully	fully	partially	partially
	$S_A$ enclosing $S_B$	yes	no	yes	no
Sense link types		Perfect	Narrower-than	Wider-than	Partial
Symbol		=	<	>	~

Table 1: Types of sense links.

## 5 Gold Standard

### 5.1 Dataset Sampling

The dataset used to create the gold standard for developing the system includes the sense links between The Oxford Dictionary of English and three bilinguals published by Oxford University Press: English to Spanish, English to Russian, and English to Chinese (EN-ES, EN-RU, and EN-ZH). This set consists of 210,148 sense pairs that had been previously linked by human annotators.

Lexemes in this collection were split according to:

- 1) Their lexical category (noun, verb, adjective, adverb/preposition, or other). We could observe that each of these classes present a different behaviour in what refers to polysemy, and therefore pose different issues to sense link annotation.
- 2) Their polysemy degree (single-sense, small-size, medium-size, and large-size entry). Similar to above, senses in highly polysemous lexemes are more challenging to align than those in, e.g., monosemous entries.

With this classification, we aimed to help annotators focus on similar cases at a time, presumably with a similar degree of difficulty and similar features. Thus, for the manual annotation effort, we created batches of 100 sense links with the same lexical category and polysemy degree. Also, all sense links belonging to a lexeme were put together into the same batch.



Due to resource constraints, we could only annotate a subsample of the original dataset. For subsampling, we calculated the percentage of links for each combination of lexical category and polysemy degree and extracted 5% of the full collection of links. The sample sums up to 10,919 links (3,965 from EN-ES, 2,445 from EN-RU, and 4,403 from EN-ZH). Furthermore, we added additional batches that were annotated by several annotators (so-called *shared batches*) to compute inter-annotation agreement (IAA).

## 5.2 Manual Annotation Effort

Four annotators carried out the granularity classifications, and a fifth one acted as the judge to resolve disagreements in the *shared batches*, all of them expert lexicographers. Resolving disagreements was essential to obtain the labels conforming to the gold standard.

Sense links were classified according to the 4-fold distinction presented in section 4: *perfect match*, *wider-than*, *narrower-than*, and *partial match*. Additionally, annotators could use the tag *unlink*, if they considered that the sense pair did not correspond to a link, and *donotknow*, if they were uncertain about it.

The lexicographers annotated a total of 15,577 sense links, organised into 160 batches. Of these batches, 146 were annotated by one person (52, 40 and 54 batches from the EN-ES, EN-RU and EN-ZH dictionaries respectively). The remaining 14 batches were *shared batches* annotated by multiple annotators (5 EN-ES, 4 EN-RU and 5 EN-ZH), resulting in a total of 20,628 annotations.<sup>4</sup>

## 5.3 Inter-annotation Agreement

We used the annotations in the *shared batches* to analyse inter-annotator agreement as a proxy for the difficulty of the task. Since we are using the annotations as the gold standard for training the classifier, the inter-annotator agreement also represents the upper-bound for the classifier's performance. We measured the inter-annotator agreement using Fleiss' kappa metric (Fleiss 1971). The data consisted of the sense links annotated by all lexicographers involved, and included sense links labelled *donotknow* and *unlink*. Table 2 shows the results by dictionary, polysemy degree and lexical category.<sup>5</sup> Note that a kappa score of 0 signifies agreement that is fully explained by chance while a kappa score of 1 implies perfect agreement. According to Landis & Koch (1977), kappa scores between 0.41 and 0.60 denote moderate agreement and scores between 0.61 and 0.80 show substantial agreement.

		EN-ES			EN-RU			EN-ZH		
Polysemy degree:		S	M	L	S	M	L	S	M	L
lexical category	adjective	0.62			0.55			0.64		
	adverb/preposition	0.72			0.64			0.66		
	noun		0.67			0.59			0.59	
	verb			0.48			0.50			0.42
	other	0.61						0.60		

Table 2: Fleiss' kappa inter-annotator agreement by lexical category, polysemy degree and dictionary.

Batches containing lexemes with lower polysemy degree show higher agreement, which suggests that these sense links were easier to classify. Likewise, the disagreement increases as the polysemy degree gets higher in all three languages. In addition, all three languages obtain similar kappa scores on average, but we observe differences for lexical categories. Adverb/prepositions get the highest values across the 3 datasets. Kappa values for adjectives drop significantly in EN-ES and EN-RU compared to adverb/preposition, while they are similar in EN-ZH. Verbs seem to carry less complexity in EN-RU: they only lose 9.28 points with respect to the average, compared to 18.15 loss in EN-ZH and 15.39 in EN-ES.

## 6 Building the Classifier

### 6.1 Models Explored

To develop our classifier, we experimented with three different approaches: using an AdaBoost ensemble of decision trees (Hastie et al. 2009), building a neural network (Glorot & Bengio 2010), and fine-tuning a pre-trained BERT model

<sup>4</sup> Before clean-up, the set comprised 14,178 unique annotations, 1,397 quadruple annotations (5,588 in total), and 862 judge reviews.

<sup>5</sup> Note that shared batches did not include links from monosemous lexemes, but only those in the polysemy degree classes of small (S), medium (M), and large (L).



(Devlin et al. 2018).<sup>6</sup> We also implemented a baseline model to be used as a benchmark in the evaluation of results. The following paragraphs outline each of these models.

### 6.1.1 AdaBoost

The first classification model we trained was an AdaBoost ensemble of decision trees. The main advantages of decision trees are that they are highly interpretable, they allow both numerical and categorical input data, and they are computationally efficient.

The implementation of AdaBoost selected for our task makes use of the meta-estimator AdaBoostClassifier,<sup>7</sup> provided by *sklearn*. The AdaBoost classifier uses an ensemble of DecisionTreeClassifiers<sup>8</sup> as base estimators. We experimented with a variety of parameters using grid search (GridSearchCV<sup>9</sup>), as described in section 6.2.

### 6.1.2 Neural Network

The second classification model we trained was a feedforward neural network. Neural networks are known to perform well with high dimensionality data, and they can model complex, non-linear relations between input variables.

For the implementation, we chose the MLPClassifier from *sklearn*.<sup>10</sup> Again, we employed grid search to fine-tune the hyperparameters, as described in section 6.2. Due to the MLPClassifier's sensitivity to feature scaling, we further experimented with feature normalisation and standardisation.

### 6.1.3 Deep Learning with BERT

The third model was developed by fine-tuning a pre-trained BERT model for the sense granularity classification task. BERT generates different word embeddings for the same word depending on its context, and words in the same context generate similar word embeddings. For this reason, BERT is a viable candidate for predicting the granularity of sense links.

An essential part of BERT is Next Sentence Prediction (NSP), where the model learns to understand sentence relations by learning whether or not a given sentence follows another (Devlin et al. 2018). Accordingly, we treat the sense link granularity task as an NSP task. By representing each sense as a sentence made up of its lexical information (e.g. its definition), we train the model to classify sense links based on the chained sentences representing the two senses.

### 6.1.4 Baseline

Finally, we also created a baseline classification method against which to compare the above models. The baseline model is a rule-based algorithm employing the method described in Table 3, which takes into account the number of times each sense is linked to a sense in the other dictionary.



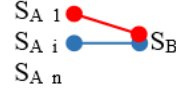
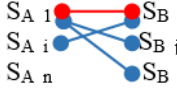
Situation	Graphical representation	Resulting link type
Neither $S_A$ nor $S_B$ are linked to any other sense in the other dictionary	$S_A$  $S_B$	<b>perfect (=)</b>
In addition to $S_{B_1}$ , $S_A$ is linked to one or more other senses (in blue) in the bilingual dictionary	$S_A$  $S_{B_1}$ $S_{B_i}$ $S_{B_n}$	<b>wider-than (&gt;)</b> (i.e. $S_A$ is wider than $S_{B_1}$ )
In addition to $S_{A_1}$ , $S_B$ is linked to one or more other senses (blue) in the monolingual dictionary	$S_{A_1}$  $S_B$ $S_{A_i}$ $S_{A_n}$	<b>narrower-than (&lt;)</b> (i.e. $S_{A_1}$ is narrower than $S_B$ )
Both $S_{A_i}$ and $S_{B_j}$ are linked to multiple senses in the other dictionary	$S_{A_1}$  $S_{B_1}$ $S_{A_i}$ $S_{B_i}$ $S_{A_n}$ $S_{B_n}$	<b>partial (~)</b>

Table 3: Baseline classification heuristics.

The algorithm works as follows: For senses  $S_A$  and  $S_B$ , it assigns the label: i) *perfect* if the number of links for both senses is equal to 1, ii) *wider-than* if the number of links is larger than 1 in  $S_A$  and equal to 1 in  $S_B$ , iii) *narrower-than* if the number of links is equal to 1 in  $S_A$  and larger than 1 in  $S_B$ , and iv) *partial* if the number of links for both senses is larger than 1.

<sup>6</sup> We used the Base, Uncased model, which can be downloaded from <https://github.com/google-research/bert>.

<sup>7</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

<sup>8</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<sup>9</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html)

<sup>10</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.neural\\_network.MLPClassifier.html](https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)



The rationale behind this method is that when a sense in one dictionary is linked to several senses in the other, it should have a wider semantic extension than any of those senses in the other dictionary. Thus, this one-to-many relation is labelled as *wider-than*. Inversely, many-to-one cases are labelled as *narrower-than*. One-to-one relations are instances where both senses should refer to the same extent of meaning, thus being labelled *perfect*. Finally, many-to-many relations suggest that both senses include a meaning component not present in the other, therefore receiving the label *partial*.

## 6.2 Evaluation Method

To find the best hyperparameters for the two machine learning models, AdaBoost and NN, we used sklearn's GridSearchCV. The choice of values to test, shown in Tables 4 and 5, results from research on the algorithms and includes the default plus neighbouring orders of magnitude. The experimental results were evaluated using 10-fold cross-validation and compared against the baseline model's performance.

AdaBoost parameters	values
<b>n_estimators</b>	10, 100, 1000
<b>learning_rate</b>	1, 0.1, 0.01, 0.001
<b>base_estimator__criterion</b>	gini, entropy
<b>base_estimator__max_depth</b>	1, 10, None
<b>random_state</b>	999

Table 4: Hyperparameters grid for AdaBoost and Decision tree estimators.

Neural Network MLP parameters	values
<b>hidden_layer_sizes</b>	(100,), (72, 36), (72, 36, 18)
<b>activation</b>	tanh, relu
<b>solver</b>	adam
<b>alpha</b>	0.001, 0.0001, 0.00001
<b>batch_size</b>	10, 100, 1000
<b>learning_rate_init</b>	0.01, 0.001, 0.0001
<b>early_stopping</b>	True
<b>random_state</b>	999

Table 5: Hyperparameters grid for Neural Network (MLP).

## 6.3 Experiments

The experiments were organised in four rounds of iterations with different goals: The first three incrementally built on top of each other and focused on the machine learning algorithms (AdaBoost and NN), while the fourth involved fine-tuning a pre-trained BERT model. The following is a summary of the main characteristics of each round.

### 6.3.1 Round 1: Ground Base

In our initial experiment, we used the same 42 dictionary-based features that had been employed for training XD-BaSeLink, the sense linking classifier, to gain initial insights into the feasibility of the classification task. The features used for this round appear listed in the appendix of Saurí et al. (2019).

### 6.3.2 Round 2: Optimising Features and Hyperparameters

In this iteration, we introduced 56 additional features. These features were both binary and categorical, and they relate to the sense order, domain, register, region, definition/indicators, and examples of the senses in each pair. Furthermore, we added as a feature the number of links for each sense in the sense pair. This feature is used in creating the baseline model (see section 6.1.4), and it is intuitively informative for the classification task. Consider, for example, a sense pair ( $S_{mono}, S_{bil}$ ) between a sense in the monolingual dictionary and another in the bilingual dictionary. Knowing that  $S_{mono}$  has



been linked to three senses in the bilingual dictionary and that  $S_{bil}$  has been linked twice intuitively makes it less likely that the sense link is *perfect*. Instead, and in agreement with the baseline algorithm, we may expect the link to fall into the *partial* category.

### 6.3.3 Round 3: Dataset Balancing

The dataset was highly imbalanced across the four classes, which can introduce a bias towards the majority classes. To counteract this bias, we experimented with several sampling techniques that balance out the classes.

*Oversampling.* The advantage of oversampling is that no observations (i.e. sense links) are lost. We used the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC), which is a variation of SMOTE (Chawla et al. 2002) tailored for datasets with continuous and categorical features. By synthetically generating more instances of the minority class, inductive learners, like decision trees, can broaden their decision regions for that class. We assessed two variations of oversampling: Oversampling the minority classes to the full number of data points in the majority class and oversampling the minority classes to half of the number of data points in the majority class.

*Undersampling.* Similarly, we also experimented with undersampling the dataset, reducing the size of the majority class. For this, we used the Random Under Sampling (RUS) technique, which randomly selects and removes data points from the majority classes, while leaving the minority classes' samples intact, until a balanced distribution is achieved.

*Hybrid resampling and boosting.* The results for the previous iterations did not show a significant change in performance for the AdaBoost classifier, and thus we decided to experiment with a further classification strategy. We used two different implementations that combine the SMOTE oversampling and the random undersampling with the boosting techniques, instead of resampling the dataset first and then using the AdaBoost classifier. The first algorithm used, called SMOTEBoost (Chawla et al. 2003), is a hybrid sampling/boosting algorithm that creates synthetic examples from the minority class, thus indirectly changing the updating weights and compensating for skewed distributions. The second classifier, called RUSBoost (Seiffert et al. 2009), is a combination of a boosting algorithm that uses RUS to randomly remove examples from the majority class, giving a clear advantage in training time, in comparison to SMOTEBoost.

### 6.3.4 Round 4: BERT Experiments

We fine-tuned the pre-trained BERT model to the classification task. This approach requires text as input rather than features; hence we defined 7 different combinations of text input to represent both senses in the link. We selected that approach after the promising results obtained by Breit et al. (2020) in identifying the sense of a word as used in context. In their experiments, they try to determine whether the meaning of a word used in a particular context matches the target sense represented by either its definition or its hypernyms. This setup is related to ours since we also represented senses using their definition. We put words in context using sense examples, and we further characterise senses with their collocates, domain labels, and more. Also, we consider BERT as a proper representation model since it is a bidirectional neural network that learns to distinguish the meaning of a word depending on its context, hence it represents words appearing in similar contexts through similar word embeddings vectors.

## 7 Results and Discussion

We evaluated the classifiers using the following evaluation metrics: *accuracy*, *precision*, *recall*, and *f1-score*. We further differentiated these scores between *macro* (equal weight per class, higher relevance to the results for small classes), and *weighted* (weight-adjusted by class, greater relevance to larger classes).<sup>11</sup> *Macro* and *weighted* averages respectively provide lower and upper bounds to the true average of these metrics. We also looked at Cohen's kappa score (Cohen 1960) to measure agreement between the human annotation and the predicted labels. The best performing model consists of an AdaBoost ensemble of decision trees, which clearly outperformed both the neural network and the fine-tuned BERT model.<sup>12</sup> We trained it with the features used in Saurí et al. (2019) and additional features encoding the lexical category of the sense link lexeme and the number of links for each sense in the sense link. Table 6 and Table 7 show the results of this model compared to the baseline model presented in section 6.1.4.

The AdaBoost classifier outperforms the baseline model in global terms, with a macro averaged f1 score of 71% compared to the baseline of 65%. It also yields a higher overall accuracy of 86% compared to 80% for the baseline. Unsurprisingly, the majority class *perfect* consistently scores highest, whereas the minority class *partial* scores lowest. This tendency is also represented by the weighted metrics, which receive higher scores than the macro averages, indicating an overrepresentation of the majority class in prediction.

<sup>11</sup> We used weighted instead of micro because the task is a multiclass classification problem, for which the micro average would yield the same values for precision, recall and subsequently f1. Refer to: <https://simonhessner.de/why-are-precision-recall-and-f1-score-equal-when-using-micro-averaging-in-a-multi-class-problem/> (05/2020)

<sup>12</sup> The best model, chosen based on weighted f1-score, was obtained with the following parameters: `base_estimator_criterion=entropy`, `base_estimator_max_depth=1`, `learning_rate=0.05`, `n_estimators=100`.



Precision			recall		f1 score		class size
	baseline	AdaBoost	Baseline	AdaBoost	baseline	AdaBoost	
Performance by class							
narrower-than	0.81	0.94	0.65	0.65	0.73	0.77	2979
partial	0.31	0.70	0.43	0.27	0.36	0.39	494
perfect	0.88	0.85	0.89	0.99	0.89	0.92	8826
wider-than	0.58	0.84	0.72	0.70	0.64	0.76	1416
Overall performance							
macro avg	0.65	0.83	0.67	0.65	0.65	0.71	N/A
weighted avg	0.82	0.87	0.80	0.86	0.81	0.85	N/A

Table 6: Overall and per class performance scores for baseline and AdaBoost models. The last column (GS class size) provides an indication of the biased nature of the gold standard (GS)

	Baseline	AdaBoost classifier
<b>accuracy</b>	0.80	<b>0.86</b>
<b>kappa</b>	0.63	<b>0.71</b>

Table 7: Overall accuracy and kappa scores for baseline and AdaBoost models.

The confusion matrix in Figure 1 visualises the data bias: most incorrect classifications had the label *wider-than* or *narrower-than*, but were predicted *perfect*. The bias stems from the skewed training data, with the majority class *perfect* having almost 18 times as many samples as the minority class *partial*. The classes *wider-than* and *narrower-than* perform a bit better than *partial*, but worse than *perfect*. We attempted to address the present bias by synthetically resampling the training data. However, doing so decreased the overall performance unjustifiably in terms of accuracy and macro averaged f-1 score, which is why we chose to stick to the original dataset.

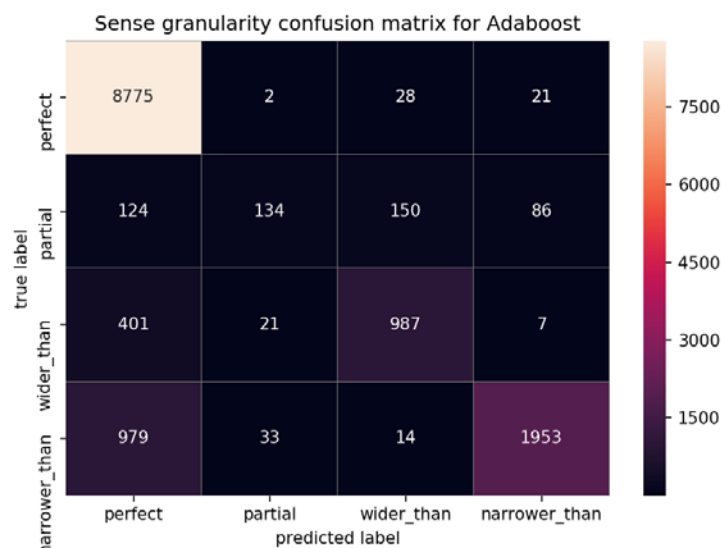


Figure 1: Heatmap showing the confusion matrix of predicted vs correct labels for the AdaBoost classifier.

As mentioned above, the training data is heavily imbalanced, posing the question whether the overall performance of our classifier could be improved by introducing additional (non-synthetic) data points, especially for the less represented classes. Alternatively, the algorithm may benefit from the introduction of additional features, which may aid in the identification of data points from the minority classes. Both of these approaches may prove viable alternatives to the sampling methods explored in this paper to decrease the algorithm's bias towards the majority class.

Overall, these results can be assessed as positive. We consider the *macro precision*, *weighted precision* and *weighted recall* values as quite good given that they are above 0.85 and reaching towards a 0.9, whereas the *macro recall* results are lower but still within a decent range. The lower *macro recall* was, however, expected, since the *macro average* metrics are very sensitive to class imbalance and assign greater prevalence to the smallest classes, for which we observed particularly low scores.

## 8 Conclusion

In this paper, we motivated the need for identifying distinctions of sense granularity between linked senses from different dictionaries to support automatic tasks in different areas, most significantly the creation and enhancement of multilingual lexical resources. As a solution, we proposed an automatic classifier that uses a standard supervised machine learning approach for multiclass classification. The resulting model performs convincingly, and can, therefore, be used for tagging linked senses with a reasonable degree of confidence. A limitation of the model is that it shows a bias towards predicting



the majority classes, due to the imbalanced nature of the training dataset. Future work can address this shortcoming in two ways: Firstly, by obtaining more manual annotations, especially for the minority classes, to have a balanced training dataset and thus improve the classifier's performance. And secondly, by enriching the training dataset with additional features that can help the classifier to distinguish among classes more accurately.

## 9 References

- Ahmadi, S., M. Arcan, J. McCrae (2019). Lexical Sense Alignment using Weighted Bipartite b-Matching. In *Proceedings of the Poster Track of LDK 2019*, pp. 12-16.
- Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 12(142), pp. 5-32.
- Atkins, B. T. S., M. Rundell (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Breit, A., Revenko, A., Rezaee, K., Pilehvar, M.T., Camacho-Collados, J. (2020). WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. URL: <https://arxiv.org/abs/2004.15016v1>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, pp. 321-357.
- Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* (pp. 107-119). Springer, Berlin, Heidelberg.
- Cinková, S., Holub, M., & Križ, V. (2012). Optimizing semantic granularity for NLP-report on a lexicographic experiment. In *Proceedings of the 15th EURALEX International Congress*, pp. 523-531.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psych. measurement*, 20(1), pp. 37-46.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Donandt, K., Chiarcos, C. (2019). Translation inference through multi-lingual word embedding similarity. In *Proc. of TIAD-2019 Shared Task Translation Inference Across Dictionaries, at 2nd Language Data and Knowledge (LDK) conference. CEUR-WS (May 2019)*.
- Fillmore, C. J., Atkins B. T. S. (1994). Starting Where the Dictionaries Stop: The Challenge for Computational Lexicography. In B. T. S. Atkins and A. Zampolli (eds.) *Computational Approaches to the Lexicon*. New York: Oxford University Press, pp. 349-393.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.
- Geeraerts, D. (1990). The Lexicographical Treatment of Prototypical Polysemy. In S. L. Tsohatzidis (ed.) *Meanings and Prototypes*. London: Routledge, pp. 195-210.
- Glorot, X., Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249-256.
- Gollins, T., M. Sanderson (2001). Improving Cross Language Retrieval with Triangulated Translation. In *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pp. 90-95.
- González, M., Buxton, C., Saurí, R. (2020). XD-AT: A Cross-Dictionary sense Alignment and mark-up Tool. In *Proceedings of the XIX EURALEX conference*. Alexandroupolis, Greece. To appear.
- Grosse, J., Saurí, R. (2020). Principled Quality Estimation for Dictionary Sense Linking. In *Proceedings of the XIX EURALEX conference*. Alexandroupolis, Greece. To appear.
- Gracia, J., B. Kabashi, I. Kernerman (2019). *Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries*, co-located with the Language, Data and Knowledge Conference (LDK). Leipzig, Germany, May 2019.
- Gurevych, I., J. Eckle-Kohler, & M. Matuschek (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool Publishers.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.
- Hanks, P., Pustejovsky, J. (2005). A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2), pp. 63-82.
- Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, 2(3), pp. 349-360.
- Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2), pp. 91-113.
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), pp. 159-174.
- Massó, G., P. Lambert, C. Rodríguez-Penagos, & R. Saurí (2013). Generating New LIWC Dictionaries by Triangulation. In R. E. Banchs, F. Silvestri, T. Liu, M. Zhang, S. Gao, J. Lang, (eds.) *Information Retrieval Technology*, pp. 263-271.
- Mausam, S. Soderland, O. Etzioni, D. Weld, M. Skinner, & J. Bilmes (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 262-270.
- McCrae, J. (2020, May). *ELEXIS Monolingual Word Sense Alignment Task*. Retrieved from CodaLab Competition: <https://competitions.codalab.org/competitions/22163>
- Miles, A., Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System Reference*. (W3C Recommendation). <http://www.w3.org/TR/skos-reference/>: World Wide Web Consortium.



- Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.
- Navigli, R. (2006). Meaningful clustering of senses helps boost WSD performance. In: *Proceedings of the 21st International Conference of Computational Linguistics and the 44th Meeting for the ACL*, pp. 105-112.
- Navigli, R., K.C., Litkowski, O. Hargraves (2007). Semeval-2007 Task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*: 30-35. ACL.
- Ordan, N., J. Gracia, M. Alper, I. Kernerman (2017). *Proceedings of TIAD-2017 Shared Task – Translation Inference Across Dictionaries*. Language, Data and Knowledge Conference, LDK 2017. Galway, Ireland, June 2017.
- Pilehvar, M. T., Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the NAACL-HLT Conference*. 2019.
- Ponzetto, S. P., Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1522-1531). ACL.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, Massachusetts: The MIT Press.
- Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *International Atlantic Web Intelligence Conference* (pp. 380-386). Springer, Berlin, Heidelberg.
- Saurí, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Leibniz-Zentrum für Informatik.
- Shcherba, L. V. (1940/1995). Towards a general theory of lexicography. *Inter. Journal of Lexicography*, 8.4, pp. 314–350
- Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1), pp.185–197.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stock, P. (1984). Polysemy. In R.K.K. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 131–140.
- Varga, I., S. Yokoyama, C. Hashimoto (2009). Dictionary generation for less-frequent language pairs using WordNet. In *Literary and Linguistic Computing*, Vol. 24, Issue 4, December 2009:449–466, <https://doi.org/10.1093/lc/fqp025>
- Villegas, M., M. Melero, N. Bel, & J. Gracia (2016). Leveraging RDF Graphs for Crossing Multiple Bilingual Dictionaries. In *Proceedings of the Language Resources and Evaluation Conference, LREC 2016*, pp. 23–28.
- Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.
- Wushouer, M., D. Lin, T. Ishida, & K. Hirayama (2014). Pivot-Based Bilingual Dictionary Extraction from Multiple Dictionary Resources. In *PRICAI 2014: Trends in Artificial Intelligence*. Springer International Publishing, pp. 221–234.

### Acknowledgements

This work has been funded by the H2020 project “Prêt-à-LLOD: Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors” under grant agreement No 825182. Also, we are very grateful to Charlotte Buxton, the expert lexicographer who contributed all the editorial knowledge we were lacking, and also helped with resolving conflicts in shared manual annotations. In addition, we would like to thank Eva Theodoridou and Anna Emberton for their support in managing and planning the work required for the project. The authors are responsible for any errors and problems.



# A Typology of Lexical Ambiforms in Estonian

Vainik E., Paulsen G., Lohk A.

*Institute of the Estonian Language*

## Abstract

The present study aims to elaborate an overall outline of the areas that give rise to PoS ambiguity in Estonian. The analysis is based on a database consisting of ca 3500 ambiguous units. Our goal is to map the problematic areas, analyse the processes behind the lexical versatility, and provide a typology of ambiguous forms (the ambiforms) for the lexicographic use. The proposed typology is based on bi- and unidirectional PoS combinations. As a result of the analysis, we show how the lexical confluence relations exhibit a network-like interaction of the traditional PoS categories. The typology of ambiforms is expected to have both theoretical and practical implications – from the perspective of the former, the topic of lexical ambiguity will be set in the modern linguistic and lexicographic frame, and from the angle of applicability, the results will support the lexicographers as the creators of lexicographic (root) databases and the developers of language technology systems analysing corpus data.

**Keywords:** parts of speech; lexical ambiguity; Estonian language

## 1 Introduction

The contemporary lexicographic processes are characterised by the tendency to data unification, which sets new demands on and urges rethinking of the part of speech (PoS) categorisation issues. Within Estonian lexicography, the procedure of part-of-speech tagging is considered unavoidable both in the compilation of new dictionaries and in the aggregation of the existing ones into lexical databases. The PoS categorisation is a part of the data model in Ekilex – the newest dictionary writing system and lexicographic root-database of The Institute of the Estonian Language. Explicit marking of word classes of every lexical entry is the ultimate goal in the output of the lexicographic resources gathered into Ekilex – the Combined Dictionary of Estonian (DicEst) 2020 and the Language portal Sõnaveeb<sup>1</sup> (“Wordweb 2019”; see Tavast et al. 2018; Koppel et al. 2019).

Generally, the word class categorisation is not a complicated task for the lexicographers, but in ambiguous cases, the decision making can be rather difficult (see Paulsen et al. 2019). In Estonian, as in many other languages, the boundaries between word classes are not always clear. One of the forces behind the word class ambiguity is conversational transposition, expressed in particular in the adjective-noun direction (see Vare 2006). As a morphologically rich language, Estonian also has a number of inflected forms that tend to move their basic lexical categorial status to another (see e.g. Grünthal 2003). The inflected word forms emerge as candidates for autonomous dictionary entries, often in a different PoS category than the base word. There are, continuously, plenty of word forms in a transition stage (Erelt et al. 2017). For these kinds of words, i.e. words or word forms that may be interpreted as belonging to more than one word class, we use the general term *ambiform*. The ambiforms, if not handled properly, cause problems while integrating different lexicographic resources into larger databases and provide inaccurate results in the corpus processing systems. Therefore, there is a need to discover the types of potential ambiforms and to create, if possible, standardised procedures to handle them in lexicographic databases as well as for disambiguation of corpus data.

The aim of the present study is to elaborate an overall outline of the areas that give rise to PoS ambiguity in Estonian, based on the existing lexical databases and the data collected during lexicographic work and in the metalexicographic study carried out among Estonian lexicographers (Paulsen et al. 2019; Paulsen et al. 2020). The quantitative and qualitative analysis is based on a database consisting of ca 3500 ambiguous units. In this study, we map the problematic areas, analyse the processes behind the linguistic changes, and provide a typology of ambiguous forms for the lexicographic use. Validation of this typology on corpus data will be the next step in our research.

We expect the results of this study to be useful not only for Estonian lexicographers but also for other languages with rich morphology, especially from the point of view of the objective to move towards integrated lexicographic resources and harmonized standards in the lexicographic description in Europe (Pedersen et al. 2018). Another field of applicability of the typology of ambiforms are the automatic word sense disambiguation and morphological segmentation systems. Today, there is a strive for exploitation of the data of rich lexical databases for the needs of diverse language technological applications.

The article is organized as follows: the theoretical background and a basic outline of the prototypical Estonian PoS categories are presented in Section 2. Section 3 gives an overview of the content and organisation of the database; it also explains the methodological solutions in the analysis of the ambiforms. The typology of ambiforms, based on bi- and unidirectional PoS combinations, is proposed in Section 4. In section 5, the results are summarised and discussed.

<sup>1</sup> <https://sonaveeb.ee/>



## 2 Background

### 2.1 Lexical Categories and Applied Linguistics

Substantially, there are two types of approaches to linguistic categories: the classes of natural language can, by nature, be seen as discrete (classical categories) or prototype-based, graded ones.<sup>2</sup> The definitions of word class categories typically combine semantic, morphological, syntactic, and sometimes even pragmatic information; the categorisation being in essence language-specific. There are hence no universal parameters to specify the boundaries of PoS and the actual criteria depend largely on the properties of the language under consideration and the aims of the classification. Word classes are often also divided into closed and open classes, depending on the ability to admit new members and to convey independent meanings. The organisation of lexemes is affected by homonymy, polysemy,<sup>3</sup> and derivational relations; due to the dynamic processes shaping language, words or word forms move from one class to another. The category change in language is not always easily distinguishable, however, the reuse of a linguistic unit in different functions can be seen as an intrinsic feature of language, favouring many-to-many relationships.

There are many linguists that have questioned the sufficiency of lexical categories to capture the grammatical behaviour of words (e.g. Culicover 1999; Croft 2001; Taylor 2012; Smith 2015). However, PoS as a categorial frame is not significant only from the theoretical point of view, these concepts are fundamental in applied linguistics as lexicography and language technology; the theoretical problems related to lexical categories are even more exigent in part-of-speech tagging and word sense disambiguation procedures. There is, for example, a set of universal tags for PoS (UPOS) developed in order to enable cross-linguistically consistent treebank annotation (see e.g. de Marnfette et al. 2014). Expanding flexibly the coverage of the limited number of UPOS labels and using an additional XPOS tag for the language specific categories seems to be a strategy to handle the less prototypic or ambiguous cases.<sup>4</sup> Such a strategy works well in the case of corpus tokens, which are surrounded by an immediate context and are, in principle, disambiguable. Several methods – either rule-based, probabilistic or neural – have been invented for morphological disambiguation (see e.g. Quecedo 2019 for further references). However, attributing a POS label to a decontextualized dictionary entry is a different task, which, inevitably, lies on a generalisation made over the correct analyses of tokens in the corpora.

When creating corpora and lexical databases over Estonian, the crucial information for distinguishing between different realisations of ambiforms is word class affiliation; incorrect or absent PoS tagging yields incorrect automatic analysis (see Koppel 2020: 62). The fluidity of the PoS-boundaries in Estonian has been explored to some extent from the lexicographic point of view. Based on the experience with compilation of the Explanatory Dictionary of Estonian (EKSS), Karelson (2005) estimates the sets of ambiguous cases to cluster around the (dominating) classes – noun, adjective, adverb, interjection, numeral, verb. Habicht et al. (2011) discuss the challenge associated with PoS subdivision by the example of adverbs (verbal particles, modal adverbs, proadverbs) in the analysis of the PoS annotation problems of Old Written Estonian lexis. A metalexicographic survey mapping lexicographers' experiences concerning PoS categorisation (Paulsen et al. 2019: 327–329; Paulsen et al. 2020) points to the following three most difficult pairs of PoS: noun-adjective, noun-adverb and noun-adposition. Hence, there are numerous ambiguous forms in Estonian with the possibility of more than one interpretation.

### 2.2 Estonian Parts of Speech in a Nutshell

Estonian is a language with rich morphology; it has both nominal and verbal inflection and derivation, and in addition productive compounding. Estonian words can be divided into four main classes by their morphological behaviour: (1) words that inflect for mood, time and person (verbs), (2) words that inflect for case and number including for grammatical cases, i.e. nominative, genitive and partitive (nominals, i.e. nouns, adjectives, numerals, and pronouns), (3) words that inflect in (some) semantic cases, but have no grammatical case forms (some adverb types and some adpositions), (4) words that have no inflectional forms (some adverb types and adpositions, conjunctions, interjections) (Viitso 2003: 32). From the point of view of PoS border areas, the exceptions of this division are of most interest – for instance, most adjectives have in addition to nominal inflection also forms for degrees of comparison, some atypical adjectives do not inflect, and some adverbs and adpositions may come in (series of) semantic case forms.

The morphological form of a word conveys information about its syntactic and semantic characteristics. In Estonian, the nominals decline in 14 cases, in singular and plural: three grammatical (nominative, genitive, partitive) and eleven semantic (illative, inessive, elative, allative, adessive, ablative, translative, terminative, essive, abessive, comitative) cases. In the group of nominals, nouns and adjectives form noun phrases that function as arguments of the predicate (subject, object, predicative). Nouns are the heads of noun phrases, and adjectives modify the noun, agreeing with its head in case and number (*suur-te-st tera-de-st* [big-PL-ELA grain-PL-ELA] “from big grains”). However, there are some systematic exceptions to this rule: adjective in the attributive position agrees only in number in four cases, terminative, essive, abessive and comitative, being marked with the genitive case (*suur-te tera-de-ga* [big-PL.GEN grain-PL-COM]

<sup>2</sup> An example of the classical categorization would be e.g. Chomsky's (1974) feature-based approach to lexical categories using a set of internal features (+/-N, +/-V); the category “verb” can be explained by the absence of the property “noun”: [-N, +V]. The prototype-based categorization can be illustrated by the famous example of Rosch (1978): there are differences in how exactly different kinds of birds correspond to the concept of “bird” (the sparrow is in certain respect “birdier” than the penguin).

<sup>3</sup> Polysems are the elements with the same form and etymology but different meanings, the units with different etymology are homonyms.

<sup>4</sup> <https://universaldependencies.org/u/pos/> [23.07.2020]



“with big grains”).

The specific property of numerals (e.g. *kaks* “two”, *teine* “second”, *neljandik* “a quarter”) is that they refer to the numeral quantity and are typically used as the head of quantifying phrase with nominal complement in partitive case (*kaks last* [two kid-PART] “two kids”). Pronouns in Estonian share the syntactic properties of nouns, adjectives, and numerals, with emptier semantics.

Adverbs modify verb phrases, adjectives, or whole sentences. The adpositional phrases (both pre- and postpositions) are often used (parallelly) with nominal cases. Adpositions are a syntactically dependent word class, they are grammatical heads and determine the position and case form of the nominals participating in adpositional phrases. Conjunctions in turn play no role in the main clause structure and serve a bridging function of construals, connecting words, phrases, or clauses.

The interjection is an exceptional PoS in the Estonian lexical system. Due to their independence of the main clause structure, interjections are sometimes treated as a type of a sentence rather than a word (Erelt 2017: 61). A special case of interjections are the expressive or ideophonic (involving both onomatopoeic and descriptive) words – the irregular/abnormal category as opposed to the “neutral vocabulary”, constituting a noticeable part of the Estonian vocabulary that is difficult to describe and categorise (Mikone 2002; 2001: 223).

The verb in Estonian has finite forms (occurring as predicates or auxiliary components of complex predicates) and non-finite forms. The non-finite forms occur in complex predicates with a finite form (past participles); in less verb-like functions, the non-finite forms appear also as subjects and objects (infinitive), as attributes and predicatives (participles), and as adverbials (supines and gerund). There is one infinitive (*luge-da* “to read”) and one gerund – the inessive case form of the infinitive (*luge-des* “while reading”). Participles inflect for voice and tense, present participles also for case and number (*luge-va-te-ga* [read-PTC-PL-COM] “with the ones that read”). Supines are inflected for voice and case, the personal supine is inflected for five cases but not for number (e.g. *luge-ma-st* [read-SUP-ELA] “from reading”). Estonian has certain verbal nouns close to non-finite forms: agent nouns (*luge-ja* “reader”; *luge-nu* “one who read”), patient nouns (*loe-tu* “something that was read”), and action nouns (*luge-mine* “reading”) (Viitso 2003: 52).

### 3 The Database of Ambiforms

The analysis of PoS border areas grounds on the database of ca 3500 ambiforms. The different sources of data are presented in Table 1. Most of the data derive from the morphological database of Estonian (MAB).<sup>5</sup> Another systematic source of the ambiguous forms is the database of The Dictionary of Estonian (ES2019), where the items specified as subheadwords<sup>6</sup> (and, thus, missing a PoS label) were retrieved.<sup>7</sup> One of the largest sources is also the file of notes taken by Geda Paulsen in the course of compiling the Estonian collocation dictionary (ECD). No duplicates of ambiforms were allowed.

Source	N	%
Morphological database	2385,00	68,07%
Subheadwords from the dictionary ES2019	494,00	14,10%
Excerpts from the collocation dictionary (ECD)	447,00	12,76%
Metalexigraphic study	124,00	3,54%
Literature	42,00	1,20%
Other	12,00	0,34%
<b>Total:</b>	<b>3504</b>	<b>100%</b>

Table 1: Constitution of the database.

The database (MS Access) comprises related tables of linguistic information. One of them records the ambiforms, i.e. the linguistic expressions, which’s categorization is not straightforward and may cause PoS disambiguation problems (e.g. *asjata* “needless(ly)”). The central table records the different interpretations of ambiforms in terms of PoS (e.g. *asjata* – adjective; *asjata* – adverb). Yet the contexts giving rise to those different interpretations are stored in a separate but related table.

Further descriptors can be added to every table (containing the ambiforms, the interpretations, the contexts, or labels) targeting the aspects that are specific to or relevant for that particular object of description. For example, we suggest that the very specific PoS labels should be categorised into more general groups that would represent the basic level PoS categories (e.g. both prepositions and postpositions can be subsumed under the label adpositions). Another example of further descriptors would be a typology of contexts. It is possible to create a set of tags (pointing e.g. to the syntactic parameters) that would characterise the contexts in general terms and reveal, thus, their commonalities. Yet another further classification could be the typology of ambiforms themselves. The purpose of the present study is to propose such a typology.

<sup>5</sup> The database unifies the morphological info of different dictionaries compiled by different authors at the Institute of the Estonian Language. At least part of the multiple markings is due to the fact that different sources have alternative markings. Importantly, the material of MAB is mostly based on paper dictionaries that have excluded a large amount of word (forms).

<sup>6</sup> In the ES2019, the inflected forms detaching from their base forms (situating on an intermediate stage in their respective grammaticalization-lexicalisation processes) are tagged as subheadwords instead of separate independent headwords; the PoS tag of the subheadword is, in this case, unmarked (see Langemets et al. 2018: 948–950).

<sup>7</sup> We thank Ülle Viks for both excerpts.



## 4 Results

In broad terms, the material demonstrates ambiguity in two respects: i) interpretability of some forms as belonging to different parts of speech, and ii) ambiguity in respect of whether and in which conditions a lexical unit should be treated as a proper headword in the dictionary in its own rights (see e.g. Karelson 2005; Blensenius & Martens 2019). The typology reported here generalises information about the interpretability of the ambiforms in terms of their PoS categorisation, only. The aspect of the entrenchment of the forms as potential new headwords of a dictionary will be tackled in another study focusing on the distributions of the ambiforms as compared to the behaviour of the ordinary, non-entrenched distribution of case forms.

The essence of present typology comprises combinations of PoS categorisations that can occur as the interpretations of an ambiform. The data about combinations in or analysis originates in two sources. First, the data imported from the MAB (N=2385) was provided with the labels of combining PoS categories. These non-coincidental markings were further inherited from the numerous aggregated dictionaries. Second, our total list of 3504 ambiforms was subjected to automatic morphological analysis<sup>8</sup> and the interpretations including different PoS labels were marked as potential PoS combinations. We decided to adopt the same PoS categories and labels as in the morphological database and used by the automatic morphological (morph-) analysis.<sup>9</sup>

Altogether 33 ambiform combinations by two, 21 by three, and 5 by four tags occurred, the majority (94%) of these are biforms by nature (i.e. the forms with two interpretations in terms of PoS categorisation). Table 2 presents the most prominent combinations of PoS by two, which are, basically, the types of ambiforms that will be described more closely in the typology below.

Directionality	No	Notification	Donor	Target	CMC	Morph-analyser (Total = 3504)	Morphological database (Total = 2385)
Bidirectional	1.1	A <> S	adjective/noun	noun/adjective		40,20%	49,10%
	1.2	D <> A	adverb/adjective	adjective/adverb	yes	8,34%	14,21%
	1.3	D <> S	adverb/noun	noun/adverb	yes	8,06%	3,86%
	1.4	I <> D	interjection/adverb	adverb/interjection		1,51%	6,54%
	1.5	N <> P	numeral/pronoun	pronoun/numeral		0,29%	0,50%
Unidirectional	2.1	V > S	verb	noun		12%	0,13%
	2.2	V > A	verb	adjective		7,77%	0,21%
	2.3	S > K	noun	adposition	yes	3,57%	0,34%
	2.4	K > D	adposition	adverb		3,11%	6,29%
	2.5	V > D	verb	adverb	yes	2,71%	0,13%
	2.6	I > S	interjection	noun	yes	2,57%	3,69%
	2.7	P > S	pronoun	noun		1,06%	0,63%
	2.8	N > S	numeral	noun		0,80%	1,30%
	2.9	P > A	pronoun	adjective		0,37%	0,63%
	2.10	J > D	conjunction	adverb		0,29%	0,50%
[...]		Others				7,57%	12,02%
						100%	100%

*Note:* CMC – change of morphological class

Table 2: Combinations of PoS and directions of biforms.

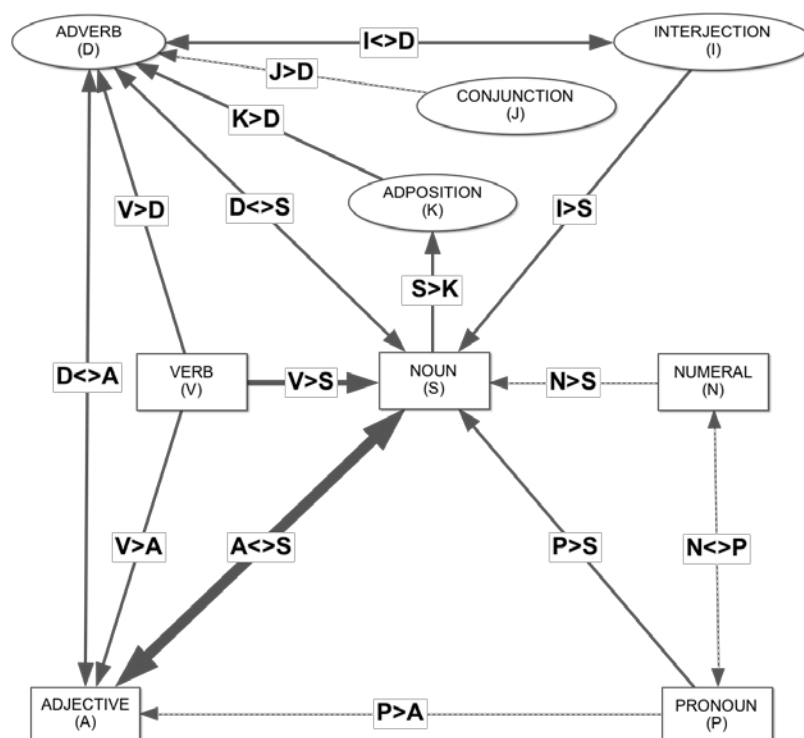
One can observe that some PoS categories occur in multiple biforms and are more prone to alter their interpretation than others. For example, nouns participate in 8 combinations, adverbs in 6, adjectives in 4, etc. Figure 1 presents the set of combinations as a network of PoS categories where the biforms occur as the connections. We explain the occurrence of some biforms as a shift in PoS categorisation. The shift can be seen as having a direction from a “donor” (i.e. original or dominant) PoS category to a “target” PoS category (i.e. the alternative interpretation). The incoming arrows can be described as the processes of adjectivisation, adverbisation, nominalisation, etc, respectively. Figure 1 demonstrates that some categories occur purely in the role of donors (interjections, conjunctions, verbs) while most of them can take both the roles of a donor and a target. According to the donor-target directionality, the types of biforms were further classified into bidirectional and unidirectional ones. This classification, as well as the roles of donor and target, are also presented in Table 2 together with the marking (“yes”) of cases where a shift across the boundary of the main morphological class (of inflected vs non-inflected words) occurs.

The descriptions of ambiform types below involve brief observations about the morphological and syntactic aspects explaining the mechanisms of the emergence of biforms. In addition, the cases of systematic polysemy and the productivity of the patterns are commented, and some suggestions are given about handling the biforms in a dictionary. For lack of space, we do not comment nor present examples of the cases where homonymy causes the multiple interpretations.

<sup>8</sup> [https://estnltk.github.io/estnltk/1.2/tutorials/morf\\_tables.html#postag-table](https://estnltk.github.io/estnltk/1.2/tutorials/morf_tables.html#postag-table) [19.05.2020].

<sup>9</sup> A – Adjective; D – Adverb; G – Genitive attribute (indeclinable adjective); H – proper noun; I – Interjection; J – Conjunction; K – Adposition; N – Cardinal numeral; P – Pronoun; S – Noun; V – Verb (see the link in the previous footnote).





Legend: Rectangles – inflected; Ovals – uninflected; Bold – proportion of cases<sup>10</sup>; Arrow – direction.

Figure 1: Network of PoS categories as connected by biforms.

## 5 The Typology of Biforms

The numeration of the biforms in the typology below follows a) directionality; b) prominence (see Table 2); the subpartition in the following analysis of biform types adheres to the type numbers as presented in Table 2.

### 5.1 Bidirectional Types

#### 1.1 A<>S [adjective<>noun]

The A<>S biform displays productive patterns of cross-using the two PoS categories: every adjective can employ the same syntactic functions as nouns (i.e. occur as a subject, object, or predicative), due to ellipsis, and some nouns can be used as a modifier. This type represents a two-way relation: the adjectives undergo nominalisation and the nouns undergo adjectivisation. Some of A<>S biforms are entrenched to the extent of inseparability on the scale pan between adjective and noun and can occasionally be tagged with both classes in DicEst.<sup>11</sup> As such, the A>S biform represents a case of systematic polysemy (Langemets 2010) (typically, QUALITY – CARRIER OF THE QUALITY), based on metonymy, and the sense menu of the dictionary reflects both senses. Also, the S>A direction reflects a metonymic relationship, where a property stands for its carrier.

(1)	A>S	<i>kallim</i>	comparative form of “dear; expensive”	“the loved one”
		<i>rase</i>	“pregnant”	“a gravid woman”
		<i>loll</i>	“stupid”	“fool”
		<i>sinine</i>	“blue”	“blue colour”
		<i>vaimulik</i>	“cleric”	“clergyman”
	S>A	<i>lemmik</i>	“favourite thing”	“favourite, dearest”
		<i>koer</i>	“dog”	“naughty, frisky”
		<i>pull</i>	“bull”	“cool, terrific”
		<i>räbal</i>	“rug”	“cheesy”

<sup>10</sup> The proportion is given according to the results of morph-analyser, see Table 2.

<sup>11</sup> In the analyses below, we compare our ambiform types to PoS tags in the most recent and comprehensive dictionary (regarding also the PoS marking), DicEst.



## 1.2 D<A [adverb<adjective]

This biform subsumes a subclass of atypical indeclinable adjectives that tends to occur in contexts typical for both adjectives and adverbs. It is possible to use the same D<A biform as a universal modifier – for instance, a such biform in the position of predicative is indistinguishable from a modifier of state/position in a sentence like *Ta õlad on lāngus* “His shoulders are dropped”. The D<A biforms are mostly used in the informal register; they convey expressive semantics and their forms are often ideophonic, i.e. phonologically motivated (see Mikone 2001, Kasik 2015: 77). There are specific suffixes deriving D<A biforms (-s -kil, -li, -il, -vel, -vil, see examples 2–7) and a suffixoid (-võitu). Another kind of ideophonic word formation is (partial) reduplication (see 8–9).

- |     |                       |                      |
|-----|-----------------------|----------------------|
| (2) | <i>kiivas</i>         | “catawampus, aslant” |
| (3) | <i>krussis</i>        | “curly”              |
| (4) | <i>laokil</i>         | “uncared for”        |
| (5) | <i>purjus</i>         | “drunken”            |
| (6) | <i>lõmmis</i>         | “crumpled”           |
| (7) | <i>lõntis</i>         | “saggy”              |
| (8) | <i>tippentoppen</i>   | “tiptop”             |
| (9) | <i>triksistraksis</i> | “ready to go”        |

Some D<A examples could be analysed also as (locative) case forms of existent nouns: *āhmi-s* “excited” [flap-INE], *küüru-s* “crooked” [hump-INE]. In some cases, a base noun (e.g. *küür* “hump”) is detectable, and a triform S>D>A emerges. A phonological feature distinguishes the nominal uses of such words from adverbial and adjectival instances: the nominal reading displays long and the adverbial/adjectival overlong<sup>12</sup> quantity (see Tiits 1982). The prolonged quantity may reveal the emancipation of the form as well as emphasis. Some, but not all such examples have orthographic distinction: *aukus* “hollow” pro *augus* “in the hole”, *harkis* “spreaded” pro *hargis* “in the fork”, *mõlkis* “dented” pro *mõlgis* “in the dent”. Importantly, the static locative semantics (inessive and adessive cases) contributes to the adjective interpretation. The directional (illative/elative; allative/ablative) forms of the same words (*mõlki*, *mõlgist*; *laokile*, *laokilt*) can be interpreted either as adverbs or the respective case forms of nouns but not as adjectives (i.e. as a type S<D according to our typology).

A subtype of D<A biform uses an adjective (by its origin) as a modifier of another adjective (see 10–11). Using (emotive) adjectives as intensifiers brings forth expressivity and new biforms may emerge on that ground. These cases are handled as a pair of homonymous entries in dictionaries. As adjectives, they inflect and agree with the head noun of the phrase; as adverbs, they stay uninflected.

- |      |                    |                    |                       |                |
|------|--------------------|--------------------|-----------------------|----------------|
| (10) | <i>hirmus lugu</i> | “dreadful affair”  | ~ <i>hirmus armas</i> | “awfully cute” |
| (11) | <i>kaunis aed</i>  | “beautiful garden” | ~ <i>kaunis külm</i>  | “pretty cold”  |

## 1.3 D<S [adverb<noun]

There are basically two ways for intersection of adverbs and nouns. First, genuine modifiers are used in the position and function of nouns, occasionally. It appears that adverbs of manner are especially prone to such a shift. One subtype of D<S biforms comprises words of foreign origin that are opaque in respect of their original meaning and are interpreted as denoting things or persons characterised by associable manners (cf. examples (12–13). A pattern of systematic polysemy, MANNER – THE PERSON/BEHAVIOUR IN THAT MANNER, appears. Another group of words reflecting a similar shift are the expressive descriptors of manner (ideophonic stems that can also be (fully or partially) reduplicated, cf. 14–15).

The second D<S subtype comprises expressive words (often derivatives, e.g. with the diminutive suffix -ke as *kübeke* (see 16), *sutike*, *natukey*, *tsipake*, all meaning roughly “a tiny thing” and “a little X”) that can be used to modify adjectives or adverbs. These words function as heads of a quantifying phrase (a noun) and they are subjected to declination in the same way as the numerals in the same position (*kübeke aega* [speck time-PART] “tiny bit of time”). They are interpreted as adverbs of measure (*kübeke pikem* [speck longer] “a tiny bit longer”); cf. the nominal use: *üks kübeke* [one speck] “one tiny bit”. Another subtype comprises case forms of nouns functioning as modifiers and situated at diverse grammaticalization stages between noun and adverb (see 17–18). The PoS categorisation depends on the level of emancipation of the forms in these specific functions and meanings. Defining this level presumes case studies and individual examination.

- |      |     |                  |            |                       |                                   |
|------|-----|------------------|------------|-----------------------|-----------------------------------|
| (12) | D>S | <i>allegro</i>   |            | “quickly”             | “a quick and lively composition”  |
| (13) |     | <i>inkognito</i> |            | “unrecognizably”      | “appearance with hidden identity” |
| (14) |     | <i>jõnks</i>     |            | “abruptly”            | “jounce”                          |
| (15) |     | <i>liga-loga</i> |            | “not taken care of”   | “trash, rubbish”                  |
| (16) | S>D | <i>kübe-ke</i>   | speck-DIM  | “tiny grain or speck” | “tiny bit of”                     |
| (17) |     | <i>ideaali-s</i> | ideal-INE  | “in ideal”            | “ideally”                         |
| (18) |     | <i>kahju-ks</i>  | damage-TRA | “to damage/harm”      | “unfortunately”                   |

<sup>12</sup> Estonian has a three-way quantity system in disyllabic feet (Lehiste 1997; Krull & Traunmüller 2000): short (quantity 1), long (quantity 2), and overlong (quantity 3).



### 1.4 I<D [interjection<adverb]

The intersection of adverbs and interjections happens when a specific type of genuine interjections is used as an expressive modifier in the role of an adverb. The cases in our database indicate that the I<D relation reflects the expressions of manner. The interjections involved in this type are of a kind that imitate a movement and often also a sound accompanying that movement. The expressive implications of I<D biforms are reflected in their ideophonic phonological form (see 19–27); there are often reduplicative patterns (21–23) and these biforms may also contain specific suffixes implicating manner (-*ti/-di*, -*ki*, cf. 24–27). The PoS shifting from interjection to adverbs hence seem to comprise a pattern of systematic polysemy, SOUND – MANNER OF MOVEMENT, and, as such, should be represented systematically in a dictionary.

(19)	<i>siuh</i>	“noise accompanying a fast strike or friction”
(20)	<i>vups</i>	“jump out of”
(21)	<i>vutt-vutt</i>	“[child’s movements] quickly, with short steps”
(22)	<i>sulla-sulla</i>	“[child’s movements in water, e.g. in bath] splash, paddle”
(23)	<i>kippadi-kappadi</i>	“clip-clop; gallop, prance with clacking noise”
(24)	<i>klõmdi</i>	“bang, plump; tiff”
(25)	<i>müraki</i>	“bang”
(26)	<i>siuhti</i>	“whizz (off)”
(27)	<i>vupsti</i>	“jump out of, slip”

Another subtype of I<D biforms comprises adverbs used as affirmative interjections: *hästi*, “well” *just* “exactly”, *justament* “exactly” (in humorous register) or the other way around: affirmative interjections that can be used as adverbs, e.g. *okei* “OK”, *oolrait* “all right”. The I<D biforms can be used in the position and function of nouns, too, which leads to an emergence of a triform I<D<S. Such a triform occurred 74 times in the data retrieved from the morphological database (e.g. *plärts* “splash”, *prõks* “crack”).

### 1.5 N<P [numeral<pronoun]

This biform constitutes a closed set of word forms as both numerals and pronouns are closed word classes. In Estonian, pronouns can substitute for numerals (yielding sc. pronominals as *mitmendik* “which part”; cf. also Section 2). The N<P biforms are compounds of a pronoun and a numeral (see 28–29). The word forms are interpretable as numerals since they occur in quantifying constructions and they are classified as pronouns due to their semantic emptiness and deictic nature. The practical solution regarding the N<P biforms’ presentation in DicEst is to tag them with both labels. A special case of N<P is the use of some numerals (e.g. *üks* “one”) as determiners marking definiteness/indefiniteness and accompanying noun phrases to indicate that its referent is identifiable (a tendency in a language lacking grammatical articles, see e.g. Dryer 2013a–b), also Hint, Nahkola, Pajusalu 2017: 66–67).

(28)	<i>mõnisada</i>	some + hundred	“a couple of hundred”
(29)	<i>paarkümmend</i>	couple + -teen	“about twenty”

## 5.2 Unidirectional Types

### 2.1 V>S [verb>noun]

The double interpretations of V>S ambiforms in our data occur due to homonymy of form, unexceptionally (e.g. *mõistes* [concept-INE] and [understand-GER]). The V>S shift comprises potentially a large set of ambiforms due to the productive and regular patterns of nominalisations (the action nouns derived with the suffix -*mine* and agent nouns derived with -*ja*) applicable to the verb stems. Such nominalisations obtain all the syntactic functions of nouns. They are analysed as nouns by automatic morphological analysis and create no disambiguation problems. The regular nominalisations are presented in DicEst only occasionally, e.g. *võimlema* “to work out” > *võimlemine* “gymnastics” > *võimleja* “gymnast”. The nominalisations will probably find their way into DicEst more often as there is no need to save the space in the era of electronic dictionaries and the goal is as detailed coverage of the vocabulary as possible. It would be useful to explicate the derivational link to the respective verbs while presenting them and add the regular derivational morphology to the block presenting conjugation.

### 2.2 V>A [verb>adjective]

The V>A biforms comprise the non-finite forms of verbs functioning as attributes – participles and supines – occurring both in verb phrases and, as modifiers, in noun phrases. Distinguishing these two types of usages is a huge problem for the automatic morphological analysis – 97% of the non-disambiguable word forms belong to this type.<sup>13</sup> The V>A biforms are also problematic for the lexicographers because it is not obvious when a verb form has emancipated enough to be handled as an autonomous dictionary entry rather than a regular conjugational verb form. There are examples of past participles in our database (see 30–34), present participles (35–37), and abessive supine forms expressing a

<sup>13</sup> The statistics originates in our analysis of the Corpus of the Estonian Web (etTenTen), containing 270 million words from 686 000 web pages, to be published.



non-performed obligatory action (see Viitso 2003: 64; examples 38–39). In some cases, the adjectival forms of verbs can be further subjected to nominalisation, in which case a triform V>A>S emerges (40–42).

(30)	<i>armunud</i>	“fallen in love”	
(31)	<i>joobnud</i>	“drunken”	
(32)	<i>surenud</i>	“dead”	
(33)	<i>austatud</i>	“honorable”	
(34)	<i>suletud</i>	“closed”	
(35)	<i>siduv</i>	“binding”	
(36)	<i>lööv</i>	“striking”	
(37)	<i>hävitav</i>	“destroying”	
(38)	<i>rääkimata</i>	“untold”	
(39)	<i>värvimata</i>	“unpainted”	
(40)	<i>alluv</i>	“subordinating”	“subordinate (person)”
(41)	<i>liidetav</i>	“adding”	“addend”
(42)	<i>tagaotsitav</i>	“wanted”	“persona non grata”

### 2.3 S>K [noun>adposition]

The S>K biforms represent certain forms of nouns (typically in locative or other semantic cases) that are (at least nearly) entrenched to the extent of independent lexical units. Such a shift reflects the process of grammaticalization. The meaning of the emancipated item has undergone bleaching but has not yet lost its semantic content fully (see (43–45)). The word forms are accompanied by a noun in genitive case form in most of the cases (forming a head-complement relation, like adpositions). The process of evolving adpositions (and adverbs, see type 1.3, the adverbs emerging from case forms of nouns) from nouns in locative cases is in Estonian ongoing one (see e.g. Grünthal 2003: 26; EKG II: 38). The lexicographic presentation of such forms would depend on the level of their entrenchment, and the syntactic and semantic analysis of the forms in context.<sup>14</sup>

(43)	<i>aja-l</i>	time-ADE	“in the time of, during”
(44)	<i>aluse-l</i>	base-ADE	“on the basis of”
(45)	<i>andme-te-l</i>	data-PL-ADE	“according to”

### 2.4 K>D [adposition>adverb]

K>D biform is a pattern that employs a relational word either in an adpositional or adverbial function, referring often to spatial relations, for instance *juurde* “hither” and *pihta* “targeted, at”. The biform is considered as syntactically dependent in the adpositional phrase, where it functions as a head of a complement (typically a noun phrase), and syntactically independent when occurring in an adverb phrase modifying a verb or an adjective. The shift from adposition to adverb can happen by skipping the nominal complement by ellipse. However, this is only one of the possible analyses, reflecting the synchronic view; diachronically, most of these ambiforms are grammaticalized forms of nouns and according to the (historical) interpretation these are tri-forms (S>D>K;<sup>15</sup> cf. also the D<>S subtype 1.3). Series of locative case forms (the lative, locative and separative ones, see 46–47) occur among those ambiforms. This is an instance of the case of the (prototypically) non-inflected words that some of them have (1–3) inflected forms: directional (illative or allative), static (inessive or adessive), and separative (elative or ablative) forms, depending on the verb’s semantics (see Viitso 2003: 66).

(46)	<i>äär</i>	“edge”	<i>äär-de</i>	ILL	“to the edge”	<i>ääre-s</i>	INE	“at the edge”	<i>ääre-st</i>	ELA	“from the edge”
(47)	<i>kand</i>	“heel”	<i>kannu-le</i>	ALL	“to behind”	<i>kannu-l</i>	ADE	“behind”	<i>kannu-lt</i>	ABE	“from behind”

The K>D type comprises also compounds as the words with the final component *-poole* ‘towards’ which are exceptional<sup>16</sup> also because both constituents are subjected to partial case inflection (see 48–50):

(48)	<i>allapoole</i>	“downwards”	<i>alla</i> “down” + <i>poole</i> “towards”
(49)	<i>allpool</i>	“lower down”	<i>all</i> “down” + <i>pool</i> “at about”
(50)	<i>altpoolt</i>	“from below”	<i>alt</i> “from down” + <i>poolt</i> “from about”

### 2.5 V>D [verb>adverb]

The V>D biform comprises non-finite verb forms converbs, i.e. supines (e.g. the abessive supine *äraarvamata* “unbelievably”), and gerunds (e.g. *mängeldes* “easily”, lit. “by playing”) in adverbial function. The converbal biforms

<sup>14</sup> The procedures are currently under development.

<sup>15</sup> Habicht & Penjam (2006: 57) argue that the direction of grammaticalization in case of Estonian adpositions is generally the following: lexical form (noun+case ending) > adverb > adposition. Based on our data, we would not generalize this direction to all cases; at this point we confine ourselves to the recognition that this topic should be investigated further.

<sup>16</sup> Generally, only the final lexeme of a compound is subjected to declination in Estonian.



are rare in our database; because of their regularity they are included to DicEst as keywords only in the case they have adapted a deviant meaning. Another case of the V>D biform is the use of the supine in the role of modifier of the main verb, e.g. *läks minema* “went away” lit “went to go (inf)” and *tuleb tulema* “he/she leaves” lit. “he/she comes to come (inf)”.

## 2.6 I>S [interjection>noun]

The I>S biform represents a productive pattern where prototypical (non-inflectional) interjections function as nouns. In this case, interjective word forms refer to the acts of interjecting, the activity itself would be expressed with verbal (finite) morphological forms that cannot be mixed with nominal patterns (and therefore this is not a I>S>V pattern in Estonian as it would be in e.g. English). Part of the I>S ambiforms expresses sounds of birds and animals and have ideophonic-imitative phonological form (see 51–54). Another I>S subtype are certain exclamatives, both loanwords and native words (55–57). The current lexicographic practice is to present the biforms of the interjections and respective nouns as a pair of homonyms, which is not the case, but relies on a practical principle that lexemes with different inflection would obtain individual entries. The I>S biform is very regular and could be described by a pattern of systematic polysemy SOUND – THE ACT OF SOUND-MAKING.

- |      |               |           |
|------|---------------|-----------|
| (51) | <i>kraaks</i> | “croak”   |
| (52) | <i>nurr</i>   | “whisker” |
| (53) | <i>prääk</i>  | “quak”    |
| (54) | <i>urr</i>    | “growl”   |
| (55) | <i>aamen</i>  | “amen”    |
| (56) | <i>braavo</i> | “bravo”   |
| (57) | <i>aitäh</i>  | “thanks”  |

## 2.7 P>S [pronoun>noun]

This is a closed group as much as the class of pronouns is closed by nature. Typical P>S biform is a pronoun that has acquired a specific (conceptually richer) meaning: *mina* “I” and “self, ego”; *eikeegi* “no-one” and “person with no value”. The P>S biforms tend to be presented as pairs of homonyms in dictionaries, which might not be an optimal solution because of their semantic relatedness. Incorporating the alternative interpretations as nouns into the sense menu of pronouns could be considered.

## 2.8 N>S [numeral>noun]

The N>S biform is a closed class as the class of numerals is closed. The numerals can be systematically used as nouns referring to the signifier of a number or a mark in a school system. Thus, a pattern of systematic polysemy NUMBER – SIGNIFIER-MARK can be postulated. Lexicographers should keep this in mind while compiling the entries for numerals. Another subtype of N>S biforms are group nouns, exploiting the numeral stems in compounds with figurative quantifiers as result (*kuradi+tosin* “the devil’s dozen, 13”, *must+miljon* “black million”).

## 2.9 P>A [pronoun>adjective]

The P>A biform is a closed group as the class of pronouns is closed by nature. The word forms can be interpreted either as adjectives because they function as attributes or pronouns (in the case they are used as substituting nouns in a clause). A special PoS tag – adjectival pronoun – is coined in some dictionaries (e.g. EKSS). A typical biform comprises the root of a pronoun and an adjectival suffix (*niisugune* “such”, lit. “this-like”; *samane* “same”, lit. same-like; *teistsugune* “different”, lit. “other-like”).

## 2.10 J>D [conjunction>adverb]

This biform is a closed set of word forms as the class of conjunctions is closed. The intersection of adverbs and conjunctions can happen when the conjunctions are used in adverbial functions. They appear, as modifiers, typically, emphasising some constituent or a whole sentence (*ega* “nor”, *justkui* “as if”, *nagu* “like”). Another case of J>D biforms are the compound conjunctions that consist of a conjunction and an adverb (*niihästi ... kui (ka)* “both ... as”).

# 6 Conclusion and Discussion

The main contribution of this study is outlining the typology of ambiforms and explaining the PoS border areas as resulting from the network-like interaction of the traditional PoS categories. Each and one of the types described above deserves a more elaborated analysis. The described network presents our current understanding based on the database of ca 3500 records; an analysis of corpora may reveal new types of connections missing from our current outline – for instance, some types described here as unidirectional may turn out to be bidirectional in nature. Therefore, we foresee that the tentative numeration presented here (reflecting the Donor-Target direction and prominence of ambiforms) could change in the later stages of the studies. Also, more specific subtypes could be distinguished and described in the future. One of the general observations arising from the descriptions above is that the noun has a special position among the interacting PoS categories. On the one hand, we found it “feeding” the syntactically dependent PoS categories as a Donor. A reason for this are the ongoing processes of grammaticalization utilising the means of nominal morphology. This



finding is in line with the study of Karelson (2005) who also addressed several nominal pairs (noun-adjective, adjective-proper noun, noun-adverb, interjection-noun) and focused separately on the phenomenon characteristic to the Estonian language, i.e. the continual supplementation of adverbial and adpositional classes by (typically locative) case forms of nouns. The phenomenon was seen as one of the biggest challenges also in the analysis of Habicht et al. (2011) concerning the problems arising in connection with PoS annotation of the corpus of Old Written Estonian. The noun-centred pairs mentioned in their study were (noun-adjective, noun-adposition, noun-adverb).

The finding in the present study is that the noun is also a popular Target: the categories with lower syntactic/semantic status could be “upgraded” to nouns while there was a need to use them in roles of subject, object or predicative. Subjecting the uninflected words occasionally to case inflection is also a possibility of morphologically rich language. Another aspect of the noun-centeredness is that in numerous cases we admitted the emergence of triforms by accepting an alternative extra analysis of the ambiform as a noun – either by nominalisation (e.g. I<D>S, V>A>S) or interpreting the form as a certain semantic case form (e.g. S>D>A).

Adverb occurred as the second attractive interacting PoS category. The analysis of bi-forms revealed that adverb emerged mostly as the Target (see Figure 1) with two exceptions (adverb>adjective and adverb>noun). Three adverbial pairs (adjective-adverb, interjection-adverb, adposition-adverb) are described also by Karelson (2005); we were able to introduce two additional pairs (verb>adverb and conjunction>adverb).

The interactions of the adjective (with noun, verb, and adverb) have also been described in the previous studies (Karelson 2005; Habicht et al. 2011; Paulsen et al. 2019). We were able to explicate the major role of the A<S> biform (40–50%) in the pool of ambiforms (see Table 2). This finding repeated the result of the previous metalexicographic study where lexicographers mentioned the noun-adjective ambiform as the most problematic area of PoS tagging (Paulsen et al. 2019). The adjective-noun conversion in Estonian is neither a clear case of the notion of syntagmatic category mixing (the syntax and semantics of one class are mixed with the morphological properties of another class) nor paradigmatic category mixing (a word has morphological properties of two categories, see Nikolaeva & Spencer 2019: 42), and the deadjectival person noun can have both a generic and referential interpretation.

A striking result of the ambiform typology is the weighty role of expressive vocabulary (onomatopoeic and descriptive words) in PoS ambiguity overall, explainable with the richness of ideophonic language in Estonian. Ambiguity with respect to lexical categorisation is a characteristic feature of the descriptive words in particular: as the results of Mikone’s (2002: 154) study show, “there is an adjective element in descriptive substantives and an adverbial element in descriptive verbs”. Another point of view on the categorisation of ideophonic expressions, traditionally also involving interjections, in Balto-Finnic (including Finnish and Estonian, among other) languages is to treat these words as a special class of ideophonic verbs, substantives, and particles (Mikone 2001: 225). We observed that the expressive/emphatic function might be a reason also for the phonological change of some ambiforms while shifting from inflected forms of nouns into the uninflected class of adverbs and atypical uninflected adjectives. The strive for increased expressivity was suggested also as a cause for certain adjectives to turn into intensifying adverbs and for a minor subtype of figurative quantifiers to occur.

Systematic polysemy was found as a predictive force behind the emergence of ambiforms. Its role in lexicology and lexicography has been described mostly from the perspective of sense alterations (Langemets 2010), its PoS altering potential is not fully discovered, yet. Surprisingly, the typical word-class altering device in Estonian, suffixal derivation (see, e.g. Vare 2006: 199) did not appear to be particularly problematic (e.g. nominalisations); neither did adjectival uses of verbal forms (participles, infinitives) – which complicate automatic PoS tagging of text corpora – emerge as problematic in the morphological database nor the metalexicographical study (Paulsen et al. 2019). This may be due to the unsteady status of these forms or even the general principle of exclusion of more or less regular phenomena from the (paper) dictionaries.

Lexicographers strive to order the lexemes correctly. The typology of ambiforms presented in this study would hopefully help the lexicographers to raise their awareness of the potential alternative interpretations. Besides, the knowledge about the proneness of certain PoS to combine with others either in bi- or unidirectional manner can be built directly into the dictionary writing system, which, then, would guide the lexicographer to check for the most probable alternative uses (and categorisations) of the headword. Awareness of the role of expressivity as a potential force bringing forth PoS ambiguity can be put to work in lexicographic work, too. It can be done, for example, by creating a module of the lexicographers’ tool that reminds them to check for alternative uses of the word whenever the PoS label “interjection” has been entered. Knowledge about productive patterns like systematic polysemy is useful in lexicographic work as it facilitates cross-checking of the meanings (and PoS categories) involved. An understanding of lexical categorisation benefits of the awareness of border areas, which, at bottom, reflects the essence of language.

## 7 References

- Blensenius, K., von Martens, M. (2019). Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies. *Proceedings of eLex 2019 conference*. 1–3 October 2019. Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 660–675.
- Chomsky, N. (1974). *The Amherst Lectures. Lectures given at the 1974 Linguistic Institute*, University of Massachusetts, Amherst; Université de Paris VII.
- CombiDic = The Combined Dictionary of Estonian (2020). I. Hein, J. Kallas, O. Kiisla, K. Koppel, M. Langemets, T. Leemets, M. Melts, S. Mäearu, T. Paet, P. Päll, M. Raadik, M. Tiits, K. Tsepelina, M. Tuulik, U. Uiibo, T. Valdre, Ü. Viks & P. Voll (eds.). Eesti Keele Instituut. Sõnaveeb 2020. Accessed at: <https://sonaveeb.ee> [14.2.2020].



- Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Culicover, P. W. (1999). *Syntactic Nuts*. Oxford University Press: Oxford.
- Dryer, M. S. (2013a). Indefinite articles. In S. Matthew Dryer, M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Accessed at: <http://wals.info/chapter/38> [14.2.2020].
- Dryer, M. S. (2013b). Definite Articles. In S. Matthew Dryer, M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Accessed at: <http://wals.info/chapter/37> [14.2.2020].
- ECD = Eesti keele naabersõnad [The Estonian Collocations Dictionary]. (2019). Kallas, J., Koppel, K., Paulsen G. & Tuulik, M., Institute of the Estonian Language. Accessed at: <http://www.sonaveeb.ee>. [14.2.2020].
- EKG I = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K. & Vare, S. (1995). *Eesti keele grammatika I. Morfoloogia*. Sõnamoodustus. Tallinn: Eesti TA Eesti Keele Instituut.
- EKG II = Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K. & Vare, S. (1993). *Eesti keele grammatika II. Süntaks*. Lisa: kiri. Tallinn: ETA Keele ja kirjanduse instituut.
- EKSS = Eesti keele seletav sõnaraamat I–VI [The Explanatory Dictionary of Estonian]. (2009). M. Langemets, M. Tiits, T. Valdre, L. Veski, Ü. Viks, P. Voll (eds.). Institute of the Estonian Language. Tallinn: Eesti Keele Sihtasutus. Accessed at: <http://www.eki.ee/dict/ekss/> [14.2.2020].
- ES2019 = Eesti keele sõnaraamat [The Dictionary of Estonian]. (2019). M. Langemets, M. Tiits, U. Uiibo, T. Valdre & P. Voll, (eds.); Institute of the Estonian Language. Accessed at: <http://www.sonaveeb.ee>. [14.2.2020].
- Grünthal, R. (2003). *Finnic Adpositions and Cases in Change*. Suomalais-Ugrilaisen Seuran toimituksia 244. Helsinki: Finno-Ugrian Society.
- Erelt, M. (2017). Sissejuhatus süntaksisse. In M. Erelt, H. Metslang (eds.) *Eesti keele süntaks*. Tartu: Tartu Ülikooli Kirjastus, pp. 537–564.
- Habicht, K., Penjam, P. (2006). Kaassõna keeleuurija ja -kasutaja käsituses [Adpositions as viewed by a linguist and by a language user]. *Emakeele Seltsi aastaraamat* 52. Tallinn, pp. 51–68.
- Habicht, K., Penjam, P., Prillop, K. (2011). Sõnaliik kui rakenduslik probleem: sõnaliikide märgendamise vana kirjakeele korpuses [‘Parts of speech as a functional and linguistic problem: annotation of parts of speech in the corpus of Old Written Estonian’]. *Estonian Papers in Applied Linguistics*, 7, pp. 19–41. Accessed at: <https://doi.org/10.5128/ERYa7.02> [14.2.2020].
- Hint, H., Nahkola, T., Pajusalu, R. (2017). With or without articles? A comparison of article-like determiners in Estonian and Finnish. *Lähivõrdlusi. Lähivertailuja*, 27, pp. 65–106.
- Karelson, R. (2005). Taas probleemidest sõnaliigi määramisel [Once more on the issues of determining parts of speech]. *Estonian Papers in Applied Linguistics*, 1, pp. 53–70.
- Kasik, R. (2015). Sõnamoodustus [Estonian word-formation]. Eesti keele varamu I. Tartu: Tartu Ülikooli kirjastus.
- Koppel, K. (2020). Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. PhD thesis. Tartu: Tartu Ülikooli Kirjastus.
- Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnavveeb: issues with and without a solution. In I. Kosem, Z. Kuhn, T. Correia, M. Ferreria, J. P. Jansen, M. Pereira, J. Kallas, M. Jakubiček, S. Krek & C. Tiberius (eds.). *Proceedings of the eLex 2019 conference. 1–3 October 2019*, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 434–452.
- Krull, D., Traunmüller, H. (2000). Perception of quantity in Estonian. *Proceedings of fonetik 2000*, pp. 85–88.
- Langemets, M. (2010). Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras [Systematic polysemy of nouns in Estonian and its lexicographic treatment in Estonian language resources]. PhD thesis. Tallinn: Eesti Keele Sihtasutus.
- Langemets, M., Uiibo, U., Tiits, M., Valdre, T. & Voll, P. (2018). Eesti keel uues kuues. Eesti keele sõnaraamat 2018 [Estonian lexis revisited: The Dictionary of Estonian 2018]. *Keel ja Kirjandus*, 12, pp. 942–958.
- Lehiste, I. (1997). Search for phonetic correlates in Estonian prosody. In I. Lehiste, J. Ross (eds.) *Estonian prosody: papers from a symposium*. Tallinn: Institute of Estonian Language, pp. 11–35.
- MAB = Eesti Keele Instituudi eesti keele morfoloogiline andmebaas (2019). Ü. Viks, I. Hein, & K. Tsepelina (Koost.). Eesti Keele Instituut. Sõnavveeb. Accessed at: <https://sonaveeb.ee> [14.02.2019].
- de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. & Manning C. D. (2014). Universal Stanford Dependencies: A cross-linguistic typology. In: *Proceedings of LREC*.
- Mikone, E. (2001). Ideophones in the Balto-Finnic Languages. In F. K. Erhard Voeltz & C. Kilian-Hatz (eds.) *Ideophones*. Amsterdam: John Benjamins, pp. 223–233.
- Mikone, E. (2002). *Deskriptiiviset sanat: määritelmät, muoto ja merkitys* [Descriptive words: definitions, form and meaning]. Helsinki: SKS.
- Nikolaeva, I., Spencer, A. (2019). *Mixed Categories. The Morphosyntax of Noun Modifications*. Cambridge University Press.
- Paulsen, G., Vainik, E., Tuulik, M. & Lohk, A. (2019). The lexicographer’s voice: word classes in the digital era. *Proceedings of eLex 2019 conference. 1–3 October 2019*, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., pp. 319–337.
- Paulsen, G., Vainik, E. & Tuulik, M. (2020). Sõnaliik leksikograafi töölaual: uuring sõnaliikide rollist tänapäeva leksikograafias [On word classes in contemporary lexicography. The lexicographers’ view]. *Estonian Papers in Applied Linguistics*, 16.



- Pedersen, B. S., McCrae, J., Tiberius, C. & Krek, S. (2018). ELEXIS – a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In *Proceedings of GlobalWordNet Conference 2018*. Singapore.
- Quecedo, J. M. H. (2019). Neural models for unsupervised disambiguation in morphologically rich languages. Master Thesis in the University of Helsinki.
- Rosch, E. (1978). Principles of categorization. In *Cognition and categorization* (27–48), E. Rosch & B. B. Lloyd (eds.). Hillsdale, Lawrence Erlbaum, New York.
- Smith, M. C. (2015). Word categories. In J. R. Taylor (ed.) *The Oxford Handbook of the Word*. OUP Oxford: Kindle Edition.
- Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana*. J. Čibej, V. Gorjanc, I. Kosem, S. Krek, (eds.). Ljubljana University Press, Faculty of Arts, pp. 749–761.
- Taylor, J. R. (2012). *The Mental Corpus: How Language is Represented in the Mind*. Oxford: Oxford University Press.
- Tiits, M. (1982). Seisundiadverbidest [On state adverbs]. *Keel ja Kirjandus* 1, pp. 17–21.
- Vare, S. (2006). Adjektiivide substantivatsioonist ühe tähendusrühma näitel. [On substantivisation of adjectives: Analysing a semantic group] *E. Niit. Keele ehe*. Tartu: Tartu Ülikool. Tartu Ülikooli eesti keele õppetooli toimetised; 30, pp. 205–222.
- Viitso, T.-R. (2003). Structure of the Estonian language: Phonology, morphology and word formation. In M. Erelt (ed.) *Estonian language*. Tallinn: Estonian Academy Publishers, pp. 1–9.

### Acknowledgements

This work was supported by the Estonian Research Council grant PSG227.

**Abbreviations:** ABE = abessive; ADE = adessive; ALL = allative case; DIM = diminutive; COM = comitative; COMP = comparative; ELA = elative; GEN = genitive case; GER = gerund; ILL = illative case; INE = inessive; PART = partitive case; PTC = participle; PL = plural; SUP = supine; TERM = terminative; TRA = translative.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicography and Corpus Linguistics**







# By the Way, do Dictionaries Deal with Online Communication? On the Use of Meta-Communicative Connectors in CMC Communication and their Representation in Lexicographic Resources for German

Abel A.

Eurac Research, Italy

## Abstract

Nowadays, people write more than ever and online writing (CMC) is a driving force in this development. A large part of everyday writing is embedded in written dialogues. This leads to a series of specific writing conventions having developed in recent years. CMC communication becomes also relevant in literacy teaching and its didactics, in particular at school. Two relevant questions arise: How can teachers evaluate the quality of digital communication? Which language resources can teachers but also students rely on in case of doubts?

One of the main functions of lexicography is to record actual language use. In its evolution, lexicography is increasingly considering oral language next to written language use. In our paper, we investigate the question whether and how interaction-oriented online writing is represented in dictionaries for German. Our analysis focuses on two meta-communicative connectors.

**Keywords:** CMC Communication, Connectors, General Dictionaries, Specialized Dictionaries

## 1 Introduction and Research Question

Online writing (CMC) is steadily increasing<sup>1</sup>. Some specific writing conventions are developing. CMC communication plays a growing role in educational contexts, too. It is even used in high-stakes tests, such as school leaving examinations<sup>2</sup>. Therefore, the question arises which language resources teachers but also students can rely on in case of doubts.

Mapping actual language use is one of the main functions of lexicography. Lexicography is evolving and increasingly considers oral language next to written language use (cf. Davies, 2017). Traditionally, lexicography has always been oriented towards the written, formal standard. Today, oral standard variants are becoming more and more important for lexicographic codification (cf. Davies, 2017), but not only, also grammatical codification is affected by the same trend (Fiehler, 2015). This development can be seen in the enhanced usage oriented and corpus based approach in lexicography and grammaticography. In addition, the emergence of an everyday standard language, including also oral language, may equally play a role (cf. Eichinger, 2005). There is, however, no research particularly focusing in the representation of CMC communication in (German) lexicography (cf. Abel & Glaznieks, 2020a). In our paper, we investigate the question whether and how interaction-oriented online writing (Storrer, 2013) is represented in selected dictionaries.

## 2 Data and Method

The investigation started with a larger study<sup>3</sup> on the use of German connectors in traditional vs. online writing (cf. Abel & Glaznieks, 2020b). In the study, selected connectors (relying on Breindl et al. 2014) were analyzed comparing different corpora. To ensure the comparability of our data we built subcorpora of about the same size (cf. table 1).

	corpus	tokens
corpora of interaction-oriented online writing	Wikipedia article discussions	376,478
	Wikipedia user discussions	377,373
	Facebook	373,383
corpora of text-oriented writing	newspaper texts	376,378
	student texts	376,184

Table 1: Corpus data overview

In this paper we will address the research question in form of a case study, focusing on two meta-communicative connectors, *übrigens* (*by the way*) and *das heißt* (*that is to say, i.e.*), considering the Facebook corpus only. This corpus is

<sup>1</sup> see <https://de.statista.com/statistik/daten/studie/475072/umfrage/taegliche-nutzungsdauer-von-sozialen-medien/> (03.07.2020)

<sup>2</sup> see e.g. <https://www.srdp.at/> (03.07.2020)

<sup>3</sup> Project „MIT.Qualität“, see <https://mitqualitaet.com/> (03.07.2020)



the *DiDi*-corpus<sup>45</sup> that includes Facebook wall posts, comments on wall posts and private messages. The texts represent the communication of different persons in the year 2013 (cf. Frey, Glaznieks, & Stemle, 2016).

We checked the usages of the connectors and their descriptions in two online general German dictionaries: the *DUDEN Online* (Duden Online, n.d.) and the *DWDS* (Digital Dictionary of the German Language, DWDS, n.d.). In addition, we consulted the specialized dictionaries of German particles by Métrich et al. (2009) and Helbig (1994).<sup>6</sup>

### 3 Some Results

This section reports some results of the study. In the first subsection, we introduce the main functions of the selected connectors as described in Breindl et al. (2014), then we present the different usages found in our Facebook corpus *DiDi* (for a more detailed description of the usages see Abel & Glaznieks, 2020a). In the second subsection, we discuss the representations of the particular usages in the dictionaries mentioned before.

#### 3.1 Usages of Connectors in a Facebook Corpus

In the following, we will first describe the connector *übrigens* and then the connector *das heißt*:

Example 1: *übrigens*

*Übrigens* is used for discourse organization. Usually it refers to side information, often given in form of a parenthesis. It is connected with an (even abrupt) change of subject. There are no restrictions with regard to the syntactic position (Breindl et al., 2014).

What is striking in the *DiDi* Facebook corpus is that *übrigens* is used relatively often – although altogether rarely – in pre-prefield position (for a definition see Pasch, Brauße, Breindl, & Waßner, 2003), especially in comparison with our newspaper and student texts<sup>7</sup>. In our *DiDi* corpus we detected particular usages, i.e. functions, in the pre-prefield position:

*übrigens* can signal a change of subject. The change of subject is less important than the distance between the connector and the action of reference. Thus, the action of reference can be located in a distant part of the interaction or even outside the specific interaction. A necessary prerequisite for mutual understanding is always the activation of shared previous knowledge. This is shown by example (1). In this case, it is hardly plausible that *übrigens* may introduce a side note and that the main topic be resumed later. Instead, *übrigens* suddenly introduces a completely new topic and implicitly includes the information that the writer is referring to shared knowledge acquired in another occasion.

(1) post (ID:56973\_4864870953340):

PERSON\_NAME\_1, schicksch du des bitte dem PERSON\_NAME\_2? Er isch net mit mir befreundet ... Übrigens: hosch es Maskottchen für die Expo in STADT\_NN schun mocht?<sup>8</sup>

(“PERSON\_NAME\_1, Could you please send it to PERSON\_NAME\_2? We are not friends ... By the way: have you already done the mascot for the Expo in CITY\_NN?”)

*Übrigens* can also signal an attempt to steer the topic of a communication back to a previous one. By doing so, the user tries to bridge a parenthesis, while the connector normally serves to insert a parenthesis in a discourse (see e.g. the definition in Breindl et al., 2014). Example (2) may illustrate this kind of usage. In this case, the dialogue concerns the particular topic of a ski lift project. Then, more people join the discussion and the topic is continued on a much more general level, focusing on questions of progress and economy. At a certain point, the initiator of the interaction remembers that the initial topic was another one and should be resumed. At the same time, the writer uses the chance to state his or her position concerning another (controversial) infrastructure project.

(2) comment (ID:54635\_6766129):

Post (ID\_54635):

ich bekomme sooo einen hals, wenn ich an das neue liftprojekt in ORT\_1 denke. neue lifte als einzige antwort auf die tourismuskrisis sind ausdruck armseliger kreativitätslosigkeit. arme natur, arme menschen.

Comments:

1 PERSON\_E: KOMMENTAR\_1 (wortloser Kommentar aus Iteration von Buchstaben und Interpunktionszeichen)

2 PERSON\_E: KOMMENTAR\_2 (Aufforderung zu einer Diskussion zum Thema)

3 ID\_54635: können wir gerne liebe PERSON\_E ☺

<sup>4</sup> To be precise, it is a subcorpus of the *DiDi*-corpus, reduced to the size necessary for the project. For practical reasons we will call this subcorpus *DiDi*-corpus in this paper.

<sup>5</sup> The *DiDi*-corpus can be accessed either by querying it through an ANNIS-interface (<https://commul.eurac.edu/annis/didi>) or by downloading it from the Eurac Research CLARIN-repository (<http://hdl.handle.net/20.500.12124/7>) (03.07.2020).

<sup>6</sup> On the difficulties of classifying the part of speech (connector, particle, discourse marker, adverb) see e.g. Breindl et al. 2014

<sup>7</sup> *übrigens* in pre-prefield position: newspaper corpus: 5 occurrences out of a total of 41, student corpus: 1 out of 3, Wikipedia AD corpus: 5 out of 118, Wikipedia UD corpus: 28 out of 121, Facebook corpus: 13 out of 55

<sup>8</sup> Written in a dialectal variant of the German language.



4 PERSON\_E: KOMMENTAR\_4 (Konkretisierungsvorschlag für Diskussion)

5 PERSON\_F: KOMMENTAR\_5 (Kritik an Fortschrittsgläubigkeit und immerwährendem Wirtschaftswachstum allgemein)

6 ID\_54635: viele haben noch nicht begriffen, dass die party zu ende ist. daher gilt es zumindest das zu schützen, was unsere einzige ressource für die zukunft ist.

7 PERSON\_G: KOMMENTAR\_7 (Zurückweisung der verallgemeinernden negativen Szenarien)

8 ID\_54635: hallo PERSON\_G! schön von dir zu hören!! lass mich halt mal ein bisschen frust loswerden. vielleicht reagier ich so, weil es die landschaft meiner kindheit betrifft. und zum thema party: ich bin mir sicher, dass das was jetzt kommen wird, nennen wir es die zeit der kleineren brötchen, uns nicht unglücklicher machen wird. im gegenteil. und deshalb schlaf ich jetzt mit einem lächeln ein :-)

9 ID\_54635: Übrigens: ich bon FÜR den ORT\_1 flughafen. nur, damit es nicht dogmatisch wird.

("it makes me sooo angry when I think of the new ski lift project in PLACE\_1. new lifts as the only answer to the tourism crisis are a sign of lack of creativity. poor nature, poor humans.

Comments:

1 PERSON\_E: COMMENT\_1 (a nonverbal comment consisting of an iteration of letters and punctuation marks)

2 PERSON\_E: COMMENT\_2 (invitation to discuss the issue)

3 ID\_54635: we can do so, dear PERSON\_E ☺

4 PERSON\_E: COMMENT\_4 (suggestion for concretizing the discussion)

5 PERSON\_F: COMMENT\_5 (criticism of the faith in progress and everlasting economic growth in general)

6 ID\_54635: many haven't understood yet that the party is over, thus, we have to protect at least what is our only resource for the future.

7 PERSON\_G: COMMENT\_7 (rejection of the generalizing negative scenarios)

8 ID\_54635: hi PERSON\_G! nice to hear from you!! let me just vent some frustration. maybe I react like this because it affects the landscape of my childhood. and regarding the issue of the party: I am sure that what is coming, let's call it the time of smaller things, won't make us less happy. quite the contrary. and that's the reason why I am now falling asleep with a smile :-)

9 ID\_54635: By the way: I am IN FAVOUR of the PLACE\_1 airport. only to avoid becoming dogmatic.")

Furthermore, *übrigens* also signals the start of a conversation in an initial post, as in oral conversations (for oral conversations cf. Duden, 2016). Example (3) shows a wall post beginning with *übrigens*, which indicates that an interaction can start without any kind of introduction.

(3) post (ID:54625\_10201088400924653):

Übrigens: Die Übertragung des Urteils #Mediaset bzw. #Berlusconi wird in Italien dem Privatfernsehen überlassen #öffentlichrechtlich

("By the way: The transmission of the #Mediaset or rather #Berlusconi judgment in Italy is left to private TV #public")

Summing up, the following functions of *übrigens* are attested in our Facebook corpus:

- Signaling a change of subject with the action of reference being quite distant
- Signaling an attempt to return to a previous topic
- Signaling the start of a conversation.

Example 2: *das heißt*

The main function of *das heißt* is to provide a reformulation of the so called "external connect", i.e. of a linguistic expression that does not immediately follow the connector but is linked to it. In addition, it can be used to specify an expression (*genauer gesagt* – more precisely), to generalize an expression (*allgemeiner gesagt* – more generally) or to correct an expression (*besser gesagt* – or rather).

Again, we found particular functions in our Facebook corpus:

*das heißt* can be used to establish interactional coherence when an external reformulation is used to ensure understanding. In this case it is not the writer who reformulates his or her own expression but the interlocutor (in the sense of *i.e./so you are telling me that ...*) as in example (4). The connects of *das heißt* are not adjacent, i.e. *SUUUPER, [...]* is not the external connect as one may expect. Instead, the external connect is in the previous statement by the interlocutor (line 4). In line 5, the writer rephrases the statement. By doing so he or she tries to prove that he or she understood (cf. Deppermann & Schmidt, 2014, p. 13).

(4) message (ID:56150\_1376649580660):

1 ID\_56150: hoila, PERSON\_H!

2 ID\_56150: jetzt seh i, du bist mitn handy online...macht nix, i schick dir a Einladung ;) nach Graz für Samstag, woasch eh, zum fußboll in Wimblon LINK\_1

3 ID\_56150: jetz obo winsch i no a guate Nocht aich olle O:)

4 PERSON\_H: LINK\_2 (Link auf einen Beitrag in einer Online-Zeitung, kommentarlos)

5 ID\_56150: SUUUPER, PERSON\_H.... des hoast, am Sonntag spilet PERSON\_NN sein 1. offizielles Spiel mit DEG, mir werdn Daumen druckn :) glg ID\_56150<sup>9</sup>

<sup>9</sup> Written in a dialectal variant of the German language.



(“1 ID\_56150: hi, PERSON\_H!  
 2 ID\_56150: now I realize, you are online with your smartphone...it doesn't matter, I am sending you an invitation;) to Graz for Saturday, you know, for football in Wimblon LINK\_1  
 3 ID\_56150: but now I wish you all a good night O:)  
 4 PERSON\_H: LINK\_2 (Link to an article in an online-newspaper, without any comment)  
 5 ID\_56150: SUUPER, PERSON\_H.... i.e./so you are telling me, on Sunday PERSON\_NN will have his first official game with DEG, we will keep fingers crossed :) lol ID\_56150”)

Finally, we will report on another usage of *das heißt*, namely a self-initiated self-correction after a slip of the pen, that we know from conversational linguistics. In this case, however, the correction is referred to a slip of the tongue (cf. Pfeiffer, 2015). Again, we will illustrate the usage by means of an authentic example from our Facebook corpus (example 5). At first sight, *das heißt* seems to function as a correction of an expression according to Breindl/Volodina/Waßner (2014). However, the sense mentioned there (i.e. “or rather”) does not match the context. Instead, in our example the writer corrects a slip of the pen.

(5) message (ID:57279\_1388449371177):

1 ID\_57279: Ich finde einige wie der PERSON\_NN1 oder die neue PERSON\_NN2 sind sehr gut und ich bin indessen froh keinen Zeitdruck mehr zu haben. Es geht mir sehr gut' danke ! Ich Manns oft noch gar nicht fassen  
 2 ID\_57279: D.h Ich Manns nicht fassen  
 3 ID\_57279: Schon wieder: och kann es nicht fassen.  
 (“1 ID\_57279: I think that some, such as PERSON\_NN1 or the new PERSON\_NN2, are very good, and, so, I am happy not to be under time pressure anymore. I am very well, thank you! I still man't believe it.  
 2 ID\_57279: I.e. I man't believe it  
 3 ID\_57279: Again: O can't believe it.”)

Summarizing, we found the following particular functions of *das heißt*:

- Establishing interactional coherence when an external reformulation is used to ensure understanding
- Self-initiating a self-correction after a slip of the pen.

### 3.2 Representations of Online-USages of Connectors in Dictionaries

In this part we will detail how *übrigens* and *das heißt* are described in the dictionaries selected for the study. We will keep the same order as in the previous subsection.

Example 1: *übrigens*

We start with the *Duden Online* dictionary. The lemma<sup>10</sup> *übrigens* is introduced as an adverb. The meaning explanation addresses exclusively the meaning “on a side note”; no further meanings or functions are mentioned. Two lexicographic examples illustrate the usage of the lemma: *du könntest mir übrigens einen Gefallen tun* (“by the way, you could do me a favor”); *übrigens, hast du schon davon gehört?* (“by the way, have you already heard about it?”). These examples look as if they were extracts taken from oral interaction or if they would reflect oral interaction. However, the *Duden Online* shares no detailed information on the source materials used for its compilation. It is important to note that no hints on any particularities or differences neither at a diaphasic or diamesic level nor on a syntactic level are given. Furthermore, as the extracts are quite short, it is not fully clear whether *übrigens* can be paraphrased with “by the way” – i.e. whether the lemma serves to introduce some side information – or whether also other meanings, as they have been presented above, would be plausible.

The second dictionary examined is the *DWDS*<sup>11</sup>. Again, the dictionary indicates “adverb” as part of speech. Also, the meaning explanation is restricted to the sense “on a side note”. The lexicographic examples show the usages of the lemma in different syntactic positions, e. g. *übrigens könntest du mir einen Gefallen tun* (“by the way, you could do me a favor”); *ich habe übrigens ganz vergessen, dir zu danken* (“by the way, I quite forgot to thank you”); *habe ich dir übrigens schon gesagt, dass ...* (“by the way, have I already told you that ...”); *übrigens (= apropos), habe ich dir schon gesagt, dass ...?* (“by the way (= apropos) have I already told you that ...?”). These examples seem to be taken from oral communication, too. In this case, the source of the meaning description and examples is indicated. It is the retrodigitized version of the “Wörterbuch der deutschen Gegenwartssprache” (“Dictionary of Contemporary German”), 1976 edition. It is actually not possible to state whether the information items of said lemma have undergone any (corpus based) re-elaboration since the 1970s. Just like the *Duden Online*, the *DWDS* does not provide any comments on particular usages.

Thus, we can state that the two large online dictionaries convey a central meaning of the lemma that is the focus also in Breindl et al. (ibid.). We do not get any hints as to particular functions for discourse organization. Those seeking for more detailed information have to refer to specialized dictionaries. For the German language, these are the dictionaries of German particles by Métrich et al. (2009) and Helbig (1994).

In Métrich et al. (2009) the article structure is quite complex: Next to the meaning or rather function description, the article for *übrigens* contains diatopic, diastratic and diaphasic information items as well as indications of the context of use, the syntactic position etc. A special feature in the article structure is the comparison with lemmas with similar meanings/functions. In the case of *übrigens*, this is a comparison with *im Übrigen*, elsewhere treated as synonymous: [...]

<sup>10</sup> <https://www.duden.de/rechtschreibung/uebrigens> (03.07.2020)

<sup>11</sup> <https://www.dwds.de/wb/uebrigens> (03.07.2020)



“Sag mir wenigstens, wie das Spiel ausgegangen ist.“ „Welches Spiel?“ „Becker gegen ...“ „Ach das ... hab's nicht zu Ende gesehen. ~, ein Typ hat angerufen. 'n Name wie Baum.“ (“Tell me at least how the game finished.” “Which game?” “Becker against ...” “Ah ... I didn't watch it to the end. ~, a guy called. Had a name like Baum”) (ibid., p. 887). According to the dictionary, in this function, i.e. in the sense of “I just remembered that” or “something I wanted to add”, *übrigens* cannot be replaced by *im Übrigen* (ibid.). Even though we found this usage also in our *DIDI*-corpus (see 3.1), the dictionary entry gives no indication on the possible role of the syntactic position or the distance of the action of reference.

In the *DIDI*-corpus, *übrigens* is also used to signal the start of a conversation. Interestingly, in the dictionary by Métrich et al. (2009) this use is – albeit not described explicitly – illustrated implicitly within the lexicographic examples used to illustrate the main sense of the lemma, i.e. adding some side information. In one of the examples *übrigens* is shown in pre-prefield position and seems to function as a starting signal: (*Aus einem Kindermärchen:*) *Einen Tag später erzählte der Löwenbruder beim Mittagessen: “~, Herr Ulster war heute direkt menschlich. Er hat meine Hausaufgaben angesehen und mich gelobt. Und als mein Freund etwas nicht verstanden hat, hat er es ganz ausführlich noch mal erklärt”*. (“(from a fairy tale:) One day later, at lunch, the lion brother reported: ‘~, today, Mister Ulster was really human. He had a look at my homework and praised me. And when my friend didn't understand something, he explained it once again in detail.’”). The dictionary, being elaborated on a corpus based approach (ibid., p. XXIII), lists both examples from written and oral language among its sources, with an explicit focus on “everyday prose” (ibid.) and oral language. However, a closer look at the bibliography reveals that the vast majority of source texts are written. Furthermore, interview data primarily stem from the 1980s and were originally produced as audio cassettes for foreign language teaching. Given the age of the data, we cannot expect that written dialogical communication was considered.

The specialized dictionary by Helbig (1994) similarly offers a detailed description of the possible functions of *übrigens* as well as a series of lexicographic examples; in this case the examples are constructed. They mainly illustrate oral language actions as the authors assume a higher frequency of particles in oral, particularly colloquial language (ibid.). This also applies to the article on *übrigens*. The fact that CMC communication does not play a role is not surprising, considering that the dictionary was written in the early 1990s. Despite the conscious decision to allow many examples to exceed the length of a single sentence and to include turn-takings (ibid.), this does not apply to *übrigens*. The example sentences reflect all the possible syntactic positions for *übrigens*. The pre-prefield position is mentioned in a specific comment which also includes examples. In this case as well, it is not perfectly clear in what way the context-free single sentence examples actually reflect those functions that are presented in the corresponding descriptions. Overall, however, it can be said that the descriptions are more detailed in comparison with the resources mentioned so far, as in the following comment: “Marks a mitigation and signals that, with regard to the main topic, the following is of minor importance, and, that the change of subject is to be understood as a digression (and justified as such), and, that a return to the previous topic or to the main topic is intended.”<sup>12</sup> (ibid.). Despite the comprehensive information, the range of all possible functions of *übrigens* we found in the *DiDi*-corpus is not covered.

#### Example 2: *das heißt*

In the *Duden Online* dictionary, *das heißt* is an independent lemma entry. There is no meaning description, only a list of synonyms (*also, beziehungsweise, nämlich, oder, respektive*, “well, or rather, namely, or, respectively”). The abbreviated form *d. h.*, a commonly used abbreviation in German, is also mentioned within the item class on orthography, next to examples illustrating the correct spelling. There is no further information. The abbreviated form *d. h.* is directly linked to the corresponding lemma entry, which contains even less information. Conversely, from the entry for *d. h.* there is no link to the lemma entry for the full form. None of the entries reference to the lemma *heißen*<sup>13</sup> (“to mean”), even though the lemma for *heißen* contains information on *das heißt*. Within the sense “to correspond to an utterance or the like in another context, to a word in another language or the like; to mean, to say, to express the same”<sup>14</sup> *das heißt* is listed among the example sentences and includes an explanatory remark: (*als Erläuterung oder Einschränkung von etwas vorher Gesagtem:*) *ich komme morgen, das heißt, nur wenn es nicht regnet; Abkürzung: d. h.* (“(as an explanation or limitation of something said before:) I will come tomorrow, i.e./to be more precise only if it is not raining”). From a user perspective a much more consistent description of *das heißt* would be desirable. Also, no indications on diasystematic particularities are mentioned anywhere.

The *DWDS* has not recorded *das heißt* as an own entry in any of its dictionary resources. There is, however, a so-called *Minimalartikel* (“minimal article”) for *d. h.* as a multiword expression<sup>15</sup>. The meaning is explained simply by giving the full form *das heißt*. Again, *das heißt* is treated within the lemma *heißen* (taken over from the “Dictionary of Contemporary German” from 1969), more precisely within the sense “to have a particular sense, to mean something”, graphically detached below the example sentences illustrating the sense. The particular function is presented as a grammar comment: *Grammatik: als Einleitung eines erläuternden Zusatzes oder einer Einschränkung des vorher Gesagten* (“Grammar: to introduce an explanatory addition or a limitation of what has been said before”). The following example illustrates the usage: *meine Bekannten wohnen in Berlin, das heißt in einem Vorort von Berlin* (“my acquaintances live in Berlin, i.e. in a suburb of Berlin”). The entry reports no particular usages nor the possibility of an abbreviated spelling.

Thus, both reference works do not record the occurrences and functions of *das heißt* we detected in our online data. As

<sup>12</sup> Own translation from the German original.

<sup>13</sup> infinitive of *heißen*

<sup>14</sup> Own translation from the German original.

<sup>15</sup> <https://www.dwds.de/wb/d.%20h>. (03.07.2020).



*das heißt* is not to be considered a particle, we cannot find it as a lemma neither in the dictionary on German particles by Métrich et al. (2009) nor in the one by Helbig (1994).

#### 4 Conclusion and Outlook

To answer the research question, we can summarize the results as follows:

- Differences at a diaphasic and diamesic level are not consistently considered in the dictionaries selected for the study. None of the dictionaries mention online writing at all. Thus, particular functions of CMC communication as illustrated in this article are not represented in reference works. The DUDEN online and the DWDS exclusively present the well established functions (see Breindl et al., 2014). The two specialized dictionaries contain much more detailed descriptions (including e.g. references to oral vs. written language) and lexicographic examples. With regard to CMC communication we have to keep in mind that the two specialized dictionaries – or rather, their sources – originated in the 1980s and 1990s.
- Differences related to the syntactic position of the connectors are not considered in any of the dictionaries. Thus, there is no clue about e.g. the particular function that *übrigens* may have in the pre-prefield position (for an analysis of its role in the middle field of a sentence in oral conversations see Egbert, 2003).

In our study we detected particular uses of two connectors in a Facebook corpus that differ from the descriptions in well established reference works. On one hand, we can notice that standard reference works as well as specialist literature tend to consider mainly traditional monological (written) texts when presenting the functions of connectors. Such texts usually aim to answer an explicit or implicit *quaestio* (for an example see e. g. Breindl et al., 2014). This, however, seems to be less important in interaction-oriented dialogical online texts. In fact, in everyday communication often quick reactions, funny jokes and fast subject changes play a much more important role (cf. Abel & Glaznieks, 2020a; Storrer, 2013). On the other hand, discourse studies mainly focus on oral interactions and the functions of discourse markers<sup>16</sup> (cf. Egbert, 2003; Imo, 2017), while the attempt to consider both interactional and monological written and oral language usages still seems to be rare (an exception is e.g. Imo, 2016). This should be taken into account more in future considering the changing contexts and habits in which (written) language is used. Practical lexicography could benefit from such a synergy.

Although the findings of our study have shown to be quite promising, we would need a larger data base to verify whether our findings represent individual or peculiar cases in our corpus or whether such usages have already become part of everyday language, and, thus, are worth being considered in dictionaries. More generally, large social media corpora for the German language covering different CMC genres and including relevant metadata as well as complete interactions (cf. Imo, 2017) would be a great asset not only for practical lexicography but also for research in applied linguistics.

#### 5 References

- Abel, A., & Glaznieks, A. (2020a). Kohärenz digital: Zum Konnektorengebrauch in der Online-Kommunikation und dessen Repräsentation in Sprachressourcen. *Deutsche Sprache. Themenheft: Textqualität Im Digitalen Zeitalter*. Hrsg. v. Abel, Andrea / Glaznieks, Aivars / Müller-Spitzer, Carolin / Storrer, Angelika, 2, 146–173. Retrieved from [https://www.dsdigital.de/download/\\_sid/NASJ-031526-Kke1/152949/ds\\_20200205.pdf](https://www.dsdigital.de/download/_sid/NASJ-031526-Kke1/152949/ds_20200205.pdf)
- Abel, A., & Glaznieks, A. (2020b). Textqualität in sozialen Medien. In K. Marx, H. Lobin, & A. Schmidt (Eds.), *Deutsch in sozialen Medien. Interaktiv – multimodal – vielfältig*. Berlin/Boston: de Gruyter.
- Breindl, E., Volodina, A., & Waßner, U. H. (2014). *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfen* (Vol. 13). Berlin/Boston: de Gruyter.
- Davies, W. (2017). Gymnasiallehrkräfte in Nordrhein-Westfalen als SprachnormvermittlerInnen und Sprachnormautoritäten. In W. Davies, A. Häcki-Buhofer, R. Schmidlin, M. Wagner, & E. Wyss (Eds.), *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke* (pp. 123–146). Tübingen: Narr Francke Attempto.
- Deppermann, A., & Schmidt, T. (2014). Gesprächsdatenbanken als methodisches Instrument der Interaktionalen Linguistik - Eine exemplarische Untersuchung auf Basis des Korpus FOLK in der Datenbank für Gesprochenes Deutsch (DGD2). *Mitteilungen Des Deutschen Germanistenverbandes*, 61(1), 4–17. Retrieved from [https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/2222/file/Deppermann\\_Schmidt\\_Gesprächsdatenbanken\\_al\\_s\\_methodisches\\_Instrument\\_2014.pdf](https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/2222/file/Deppermann_Schmidt_Gesprächsdatenbanken_al_s_methodisches_Instrument_2014.pdf)
- Duden. (2016). *Die Grammatik: unentbehrlich für richtiges Deutsch*. Berlin: Dudenverlag.
- Duden Online. (n.d.). Duden. Berlin: Bibliographisches Institut/Dudenverlag. Retrieved from [www.duden.de](http://www.duden.de)
- DWDS. (n.d.). Das Digitale Wörterbuch der deutschen Sprache. *Berlin-Brandenburgische Akademie Der Wissenschaften*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. Retrieved from [www.dwds.de](http://www.dwds.de)
- Egbert, M. (2003). Die interaktionelle Relevanz einer gemeinsamen Vorgeschichte: Zur Bedeutung und Funktion von “übrigens” in deutschen Alltagsgesprächen. *Zeitschrift Für Sprachwissenschaft*, 22(2), 189–212.
- Eichinger, L. M. (2005). Standardnorm, Sprachkultur und die Veränderung der normativen Erwartungen. In L. M. Eichinger & W. Kallmeyer (Eds.), *Standardvariation: Wie viel Variation verträgt die deutsche Sprache?* (pp.

<sup>16</sup> partly used synonymously with „connector“ (see footnote above)



- 363–381). Berlin/Boston: de Gruyter. <https://doi.org/10.1515/9783110193985.363>
- Fiehler, R. (2015). Grammatikschreibung für gesprochene Sprache. *Sprachtheorie Und Germanistische Linguistik*, 25(1), 3–20.
- Frey, J.-C., Glaznieks, A., & Stemle, E. W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In A. Corazza, S. Montemagni, & G. Seneraro (Eds.), *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), 5-6 December 2016, Napoli* (pp. 157–161). Torino: Accademia University Press. Retrieved from [www.aAccademia.it/CLIC\\_2016](http://www.aAccademia.it/CLIC_2016)
- Helbig, G. (1994). *Lexikon deutscher Partikeln* (3., durchg.). Leipzig: Langenscheidt Verl. Enzyklopädie.
- Imo, W. (2016). *Diskursmarker: grammatischer Status - Funktionen in monologischen und dialogischen Kontexten - historische Kontinuität* (Arbeitspapiere Sprache Interaktion No. Nr. 65 (06/2016)). Retrieved from <http://arbeitspapiere.sprache-interaktion.de>
- Imo, W. (2017). Interaktionale Linguistik und die qualitative Erforschung computervermittelter Kommunikation. In M. Beißwenger (Ed.), *Empirische Erforschung internetbasierter Kommunikation* (pp. 81–108). Berlin/Boston: de Gruyter. <https://doi.org/10.1515/9783110567786-004>
- Métrich, R., & Faucher, E. (2009). *Wörterbuch deutscher Partikeln: Unter Berücksichtigung ihrer französischen Äquivalente*. Berlin/Boston: De Gruyter. Retrieved from <http://ebookcentral.proquest.com/lib/unibz/detail.action?docID=533636>
- Pasch, R., Brauße, U., Breindl, E., & Waßner, U. H. (2003). *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfers*. Berlin: De Gruyter.
- Pfeiffer, M. (2015). *Selbstreparaturen im Deutschen: Syntaktische und Interaktionale Analysen*. Berlin/Boston: de Gruyter. Retrieved from <http://ebookcentral.proquest.com/lib/unibz/detail.action?docID=4054136>
- Storrer, A. (2013). Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia. In H. Feilke, J. Köster, & M. Steinmetz (Eds.), *Textkompetenzen für die Sekundarstufe II* (pp. 277–304). Stuttgart: Fillibach bei Klett.







# Δημιουργία ηλεκτρονικής λεξικογραφικής βάσης για το περιθωριακό λεξιλόγιο της ΝΕ: αρχικός σχεδιασμός

Χριστοπούλου Κ.<sup>1,3</sup>, Ξυδόπουλος Ι. Γ.<sup>2,3</sup>

<sup>1</sup> Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Ελλάδα

<sup>2</sup> Πανεπιστήμιο Πατρών, Ελλάδα

<sup>3</sup> Ελληνικό Ανοικτό Πανεπιστήμιο, Ελλάδα

## Περίληψη

Στην εργασία αυτή, αφού αρχικά αναφερθούμε στους λόγους που μας οδήγησαν στη δημιουργία μίας ηλεκτρονικής λεξικογραφικής βάσης για στοιχεία του περιθωριακού λεξιλογίου της νέας ελληνικής (βλ. εν. 1), παρουσιάζουμε τα λεξιλόγια που θα φιλοξενοούνται στη βάση (βλ. εν. 1.1) και τα τεχνικά της χαρακτηριστικά (βλ. εν. 1.2). Εκτενώς θα αναλύσουμε τη μεθοδολογία που ακολουθούμε στο αρχικό στάδιο της έρευνας αλλά και τους λόγους που επιλέξαμε να σχεδιάσουμε τη βάση σε εφαρμογή ιστού βλ. (εν. 2.1). Ειδικότερα, εστιάζουμε στη μακροδομή της βάσης και τη μικροδομή των λημμάτων (βλ. εν. 2.2 & 2.3). Παρουσιάζουμε μία πρώτη μορφή της μικροδομής των λημμάτων, αναλύοντας τις πληροφορίες που θα εμφανίζονται (π.χ. μορφολογικές, φωνητικές, σημασιολογικές, παραδείγματα χρήσης κ.ά.). Στην ενότητα 2.4 καταγράφουμε τους λόγους που θα αξιοποιήσουμε σώματα κειμένων για την τεκμηρίωση των πληροφοριών της μικροδομής. Τέλος, στην ενότητα 3 αναφερόμαστε στις καινοτομίες που θα παρουσιάζει η ηλεκτρονική λεξικογραφική βάση για τα στοιχεία του περιθωριακού λεξιλογίου.

**Έννοιες-Κλειδιά:** περιθωριακό λεξιλόγιο νέας ελληνικής· ηλεκτρονική λεξικογραφική βάση· εφαρμογή ιστού· σχεδιασμός· μικροδομή· σώματα κειμένων· Sketch Engine

## 1 Λόγοι δημιουργίας ηλεκτρονικής λεξικογραφικής βάσης

Τα περιθωριακά λεξιλόγια της νέας ελληνικής, σύγχρονα και μη, αν και εμφανίζονται συχνά στον προφορικό λόγο και σε γλωσσάρια με λαογραφικό κυρίως χαρακτήρα, σπάνια καταγράφονται/καταχωρίζονται σε ειδικά λεξικά που ακολουθούν τις λεξικογραφικές πρακτικές. Στην προσπάθειά μας να καταγράψουμε σημαντικό μέρος του λεξικού αποθέματος αυτών των λεξιλογίων, στο ευρύτερο πλαίσιο διαφύλαξης της άυλης πολιτιστικής μας κληρονομιάς (Hoffman 2009; Forrest 2011), σκοπεύουμε να συγκεντρώσουμε μεγάλο μέρος του βασικού κορμού του περιθωριακού λεξιλογίου σε μία ηλεκτρονική λεξικογραφική βάση, που θα ικανοποιεί τις αρχές της σύγχρονης λεξικογραφικής ανάλυσης (Burke 2003; Oppentocht & Schutz 2003; Fellbaum 2014).

Με τη δημιουργία της συγκεκριμένης ηλεκτρονικής λεξικογραφικής βάσης για το περιθωριακό λεξιλόγιο στοχεύουμε αφενός στη διάσωση του ειδικού λεξιλογίου περιθωριοποιημένων κοινωνικών ομάδων, σύγχρονων και μη, και αφετέρου στην ύπαρξη ενός έργου αναφοράς, που θα φιλοξενεί συγκεντρωμένο τον πλούτο του περιθωριακού λεξιλογίου, θα καλύπτει το λεξικογραφικό κενό που παρατηρείται στο συγκεκριμένο τμήμα του λεξικού αποθέματος της νέας ελληνικής και θα ανανεώνεται διαρκώς με νέα στοιχεία. Στόχος μας είναι η καταγραφή, η συγκεντρωση, η διάσωση και η λεξικογραφική τεκμηρίωση, παρωχημένων και μη, στοιχείων του περιθωριακού λόγου που εντοπίζονται σε σύγχρονα λεξικά της νέας ελληνικής, σε ειδικά γλωσσάρια και σε έντυπες ή ψηφιακές πηγές με περιθωριακό λόγο (βλ. εν. 2.2.1).

Επιλέξαμε να αναπτύξουμε μία ηλεκτρονική λεξικογραφική βάση για τα στοιχεία του περιθωριακού λεξιλογίου και όχι έντυπο λεξικό, αναγνωρίζοντας τα σημαντικά πλεονεκτήματα που παρέχουν στον χρήστη τα ηλεκτρονικά μέσα, συγκριτικά με τα έντυπα. Με την ηλεκτρονική λεξικογραφική βάση αντιμετωπίζουμε προβλήματα ή/και αδυναμίες, όπως ο περιορισμός χώρου, η περιορισμένη μικροδομή των λημμάτων και η αλφαβητική ταξινόμηση και αναζήτηση από τον χρήστη (Burke 2003: 242-6; Oppentocht & Schutz 2003: 215).

Πιο συγκεκριμένα, η ηλεκτρονική βάση δίνει τη δυνατότητα αναπτυγμένης μικροδομής χωρίς δυσνόητες συντομογραφίες και δυσανάγνωστη σελιδοποίηση, που δυσχεραίνει τον χρήστη. Ο απεριόριστος αποθηκευτικός χώρος μας επιτρέπει να καταχωρίσουμε ολοκληρωμένα, και όχι αποσπασματικά, πληροφορίες για το κάθε λήμμα, καθώς και παραδείγματα χρήσης από αυθεντικό λόγο με παραπομπή στις πηγές και με παράλληλη χρήση οπτικοακουστικού υλικού (Dodd 1989: 88; Burke 2003: 246-249) (βλ. εν. 2.3). Αναφορικά με την αναζήτηση λημμάτων και πληροφοριών, οι ηλεκτρονικές λεξικογραφικές βάσεις καταργούν τα υποταγμένα λήμματα και την επανάληψη των πληροφοριών (διπλοεγγραφές), αφού η μακροδομή δεν είναι στατική αλλά δυναμική. Επίσης, πολύ σημαντικό είναι ότι τα ηλεκτρονικά μέσα δίνουν τη δυνατότητα να ανανεώνονται συνεχώς οι πληροφορίες, να επικαιροποιούνται με νέες και να διορθώνονται τυχόν λάθη ή/και παραλείψεις. Τέλος, χάρη στη σύγχρονη τεχνολογία, οι ηλεκτρονικές βάσεις θεωρούνται το πλέον φιλικό και χρηστοκεντρικό εργαλείο τόσο για το ευρύ κοινό όσο και τους ερευνητές, που εμπλέκονται στη δημιουργία και ανάπτυξη τους.



## 1.1 Σύσταση λημμάτων/λεξιλογίου

Στόχος μας είναι να συγκεντρώσουμε, να καταγράψουμε και να τεκμηριώσουμε λεξικογραφικά περίπου 5.000 λήμματα του περιθωριακού λεξιλογίου. Όλα τα λήμματα που θα ενταχθούν στη βάση θα ανήκουν σε σύγχρονα και παλαιότερα περιθωριακά λεξιλόγια της νέας ελληνικής (βλ. Χριστοπούλου 2016: 82-125). Ειδικότερα, αναμένεται να ενταχθούν σε αυτή στοιχεία από το λεξιλόγιο της πιάτσας, τα καλιαρντά, των ρεμπέτηδων και από νεότερα λεξιλόγια, όπως το λεξιλόγιο των νέων, των τοξικομανών, των φυλακισμένων, των φαντάρων και των φιλάθλων. Σκοπός μας είναι, επίσης, να συγκεντρώσουμε στοιχεία και από άλλα περιθωριακά λεξιλόγια, όπως το λεξιλόγιο των χαρτοπαικτών, των μηχανόβιων, των ιεροδούλων και στοιχεία από το λεξιλόγιο των games, ανάλογα με τις διαθέσιμες πηγές (βλ. εν. 2.2.1). Στοιχεία που στα γενικά λεξικά χαρακτηρίζονται ως λέξεις ταμπού, υβριστικές, προσβλητικές ή/και χυδαίες στη βάση θα ενταχθούν στην ευρύτερη κατηγορία «άσεμνο λεξιλόγιο».

## 1.2 Χαρακτηριστικά ηλεκτρονικής βάσης

Η ηλεκτρονική βάση, που αναπτύσσεται σε εφαρμογή ιστού, θα περιέχει υπολεξικά και υπερλεξικά στοιχεία από παλαιότερα και σύγχρονα λεξιλόγια (βλ. εν. 1.1). Το υλικό θα συλλεχθεί από υπάρχουσες, έντυπες και ηλεκτρονικές, πηγές και η μικροδομή κάθε λήμματος, λόγω της ηλεκτρονικής φύσης της βάσης, θα είναι αναπτυγμένη, χωρίς συντομογραφίες και σύμβολα. Τα λήμματα θα συνδέονται μεταξύ τους με διαφορές μέσω υπερσυνδέσμων, όπου αυτό κρίνεται αναγκαίο (π.χ. συνώνυμα, αντίθετα, παράγωγες λέξεις κ.λπ.). Βασική καινοτομία της ηλεκτρονικής βάσης θα είναι, μεταξύ άλλων, η χρήση σωμάτων κειμένων (Baker 2006; Γούτσος & Φραγκάκη 2015; Heuberger 2016) (βλ. εν. 3). Στο πλαίσιο αυτό αξιοποιούμε το *Sketch Engine*, ένα διαχειριστή σωμάτων κειμένων και ανάλυσης λογισμικού που περιλαμβάνει εκτενές σώμα κειμένων (*web based*) σε δέκα γλώσσες, μεταξύ αυτών, και τα ελληνικά (βλ. εν. 2.4). Με τη βοήθεια του *Sketch Engine* θα έχουμε τη δυνατότητα να ελέγχουμε τη συχνότητα εμφάνισης κάθε λήμματος και να αναζητήσουμε τις διαφορετικές σημασίες και αυθεντικά παραδείγματα χρήσης για κάθε σημασία.

## 2 Μεθοδολογία

### 2.1 Επιλογή εργαλείου

Η ηλεκτρονική λεξικογραφική βάση αναπτύσσεται σε εφαρμογή ιστού, δηλαδή μια διαδικτυακή εφαρμογή, που βασίζεται στο πρωτόκολλο HTTP και οι αλληλεπιδράσεις της επιδέχονται μηχανικής επεξεργασίας. Επιλέξαμε το συγκεκριμένο εργαλείο και όχι κάποιο από τα προσφερόμενα λογισμικά/προγράμματα ανάπτυξης (π.χ. TLex, Lexique Pro, SDL MultiTerm 2014), καθώς καλύπτει πλήρως τις προδιαγραφές που θέσαμε για το τελικό προϊόν. Η εφαρμογή ιστού δίνει τη δυνατότητα στον ερευνητή να σχεδιάσει και να αναπτύξει μια ηλεκτρονική λεξικογραφική βάση εξ' ολοκλήρου σε ψηφιακή μορφή (Pautasso, Zimmermann & Leymann 2008: 806). Πιο συγκεκριμένα, μία εφαρμογή ιστού, που βασίζεται στην υπάρχουσα υποδομή του παγκόσμιου ιστού, λειτουργεί ανεξαρτήτως λειτουργικού συστήματος και γλώσσας προγραμματισμού, χωρίς την εγκατάσταση κάποιου λογισμικού στον υπολογιστή του δημιουργού και του χρήστη. Βασικό πλεονέκτημα είναι ότι μπορεί να λειτουργήσει σε οποιαδήποτε συσκευή, που διαθέτει σύνδεση με το διαδίκτυο (Rodriguez 2008). Η αποθήκευση των δεδομένων γίνεται αυτόματα, χωρίς κάποια επεξεργασία ή περαιτέρω κίνηση από τον δημιουργό. Οποιαδήποτε αλλαγή ή διόρθωση στο περιεχόμενο των πληροφοριών μπορεί να γίνει άμεσα και με ευκολία, με τον χρήστη να έχει πρόσβαση στο τελικό αποτέλεσμα. Επιπλέον, η εφαρμογή ιστού δίνει τη δυνατότητα για άμεση επικαιροποίηση/βελτίωση των ορισμών και ανανέωση του ληματολογίου, όποτε αυτή κριθεί αναγκαία, ενώ δίνει και τη δυνατότητα χρήσης υπερσυνδέσμων και πολυμεσικού υλικού για την τεκμηρίωση των πληροφοριών της μικροδομής. Προβλήματα ασυμβατότητας με το ελληνικό αλφάβητο δεν έχουν παρατηρηθεί.

### 2.2 Οργάνωση μακροδομής

Ως προς τα τυπολογικά χαρακτηριστικά, η ηλεκτρονική λεξικογραφική βάση εντάσσεται στην κατηγορία μονόγλωσσα, συγχρονικά, περιορισμένα, έργα αναφοράς, ειδικού λεξιλογίου (Landau 2001: 12-22, 25-27, 28-42; Swanepoel 2003: 58-67). Ως προς τη δομή της, η βάση αναμένεται να περιέχει πληροφορίες για την πολιτική της (π.χ. λόγοι δημιουργίας, σύσταση ληματολογίου, πληροφορίες μικροδομής κ.ά.) και το κύριο σώμα των λημμάτων, στο οποίο ο χρήστης θα έχει πρόσβαση στην μικροδομή κάθε λήμματος μέσω αναζήτησης συγκεκριμένων πληροφοριών. Ειδικότερα, η βάση θα επιτρέπει στους χρήστες να εφαρμόζουν πολλαπλά κριτήρια αναζήτησης, π.χ. γραμματική κατηγορία, λεξιλόγιο εμφάνισης, μορφολογική διαδικασία σχηματισμού κ.ά., προκειμένου να εντοπίσουν τα στοιχεία που αναζητούν. Λόγω της ηλεκτρονικής μορφής της βάσης, θα υπάρχει η δυνατότητα ο κατάλογος των λημμάτων να εμφανίζεται με πολλαπλές μακροδομές. Οι χρήστες θα έχουν τη δυνατότητα να ανατρέχουν σε άλλα λήμματα που σχετίζονται με το κύριο λήμμα, όπως σύνθετα, παράγωγα, συνώνυμα, αντίθετα, «πατώντας» στο λήμμα και έχοντας πρόσβαση στη μικροδομή (βλ. εν. 2.3). Επομένως, οι λέξεις αυτές θα καταχωρίζονται και ως ξεχωριστά/ανεξάρτητα λήμματα στη βάση με όλες τις παρεχόμενες πληροφορίες για τη μικροδομή. Στην αρχική σελίδα, αναμένεται να εμφανίζονται οι πιο πρόσφατες καταχωρίσεις ή/και λήμματα με την υψηλότερη συχνότητα εμφάνισης.

Η μακροδομή της ηλεκτρονικής λεξικογραφικής βάσης θα είναι δυναμική, εφόσον σκοπεύουμε να εμπλουτίζεται και να επικαιροποιείται με νέα στοιχεία και δεδομένα τόσο από ερευνητές όσο και από χρήστες. Επίσης, στόχος μας είναι να υπάρχει σύνδεση με άλλες ψηφιακές βάσεις, που περιέχουν πληροφορίες για στοιχεία του περιθωριακού λεξιλογίου,

1 Η πλατφόρμα είναι διαθέσιμη στη διεύθυνση: <https://www.sketchengine.co.uk/>.



δίνοντας στον χρήστη τη δυνατότητα να διασταυρώνει, να επαληθεύει και να εμπλουτίζει τις πληροφορίες που αντλεί, αναδεικνύοντας έτσι την περιγραφική και ελαχιστοποιώντας τη ρυθμιστική λειτουργία της βάσης.

### 2.2.1 Επιλογή λημμάτων

Αναφορικά με την επιλογή των λημμάτων που θα φιλοξενοούνται στη βάση, αξίζει να αναφερθεί πως η επιλογή τους είναι ιδιαίτερα δύσκολη. Αρχικά, σκοπεύουμε να συγκροτήσουμε κατάλογο λημμάτων, αξιοποιώντας τις διαθέσιμες έντυπες και οπτικοακουστικές πηγές που περιέχουν περιθωριακό λεξιλόγιο. Για τη συγκρότηση του ληματολογίου θα αναζητήσουμε στοιχεία από το λόγο περιθωριοποιημένων κοινωνικών ομάδων σε:

(α) κάποια από τα υπάρχοντα συγχρονικά, μεσαίου μεγέθους, λεξικά της ΝΕ (π.χ. *Λεξικό του Ιδρύματος Τριανταφυλλίδη* 1998; *Λεξικό της Νέας Ελληνικής Γλώσσας*, Μπαμπινιώτης 2002; *Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας*, Χαράλαμπάκης 2014), στα οποία καταχωρίζονται κυρίως στοιχεία που ανήκουν στο λεγόμενο «άσεμνο λεξιλόγιο» ή «ταμπού» ή «χυδαίο», όπως το χαρακτηρίζουν και τα ίδια τα λεξικά·

(β) ειδικά λεξικά, γλωσσάρια, ευρετήρια από δοκίμια περιθωριακού ή αγοραίου λεξιλογίου, παλαιότερα και σύγχρονα (π.χ. *Καλιαρντά*, Πετρόπουλος 1971; *Το λεξικό της ντάγκλας*, Χρηστάκης & Επάρατος 1995; *Το λεξικό της πιάτσας*, Ζάχος (Παπαζαχαρίου) 1999; *Το λεξικό της αργκό*, Μακροδήμος & Ακριτίδης 2007; *Λεξικό λαϊκής και περιθωριακής γλώσσας*, Κάτος 2016)·

(γ) ψηφιακές πλατφόρμες, portals, ψηφιακά λεξικά (π.χ. [www.slang.gr](http://www.slang.gr),<sup>2</sup> <http://lexislang.neurolingo.gr/>, <https://www.cyslang.com/>) που δημιουργήθηκαν για καταγραφή ή/και μελέτη του περιθωριακού λεξιλογίου·

(δ) λήμματα που έχουν αποτελέσει αντικείμενο έρευνας σε σχετικές μελέτες (π.χ. Κεχαγιά 1997; Κατσίκης & Σπυρόπουλος 1999; Ανδρουτσόπουλος 1998; 2001; 2011; Σηλιώτης & Φραγκιαδάκης 2007; Ανδρουτσόπουλος & Κακριδή 2010; Χρήστου & Στάμου 2013; Καμηλάκη, Κατσούδα & Βραχιονίδου 2015)·

(ε) στοιχεία περιθωριακού λεξιλογίου που έχουν περιληφθεί ως αντικείμενο μελέτης της διδακτορικής διατριβής της Χριστοπούλου (2016) (κατάλογος περίπου 2.000 λέξεων/φράσεων)·

(στ) διαθέσιμο οπτικοακουστικό υλικό, ταινίες και τραγούδια από ή για ομάδες του περιθωρίου (π.χ. το ντοκιμαντέρ της Π. Ρεβενιώτη «*Καλιαρντά*», το τραγούδι «*Το λεξικό του μάγκα*» του Π. Κυριακού<sup>3</sup> κ.ά.).

Συγκεντρώνοντας το υλικό από όλες τις παραπάνω διαθέσιμες πηγές ο κατάλογος των λημμάτων αναμένεται να έχει μεγάλο μέγεθος και αρκετές αλληλεπικαλύψεις, καθώς αρκετά στοιχεία εμφανίζονται σε παραπάνω από μία πηγές. Αξίζει να αναφερθεί πως τα διαθέσιμα λήμματα στην ψηφιακή πλατφόρμα [www.slang.gr](http://www.slang.gr) αγγίζουν τις 24.000, ενώ στον ιστότοπο <http://lexislang.neurolingo.gr/> ξεπερνούν τα 550. Στο ψηφιοποιημένο λεξικό του Κάτου (2016) τα λήμματα ανέρχονται στις 50.000, στο λεξικό των Μακροδήμου & Ακριτίδη (2007) εμπεριέχονται 800 εκφράσεις με στοιχεία περιθωριακού λόγου, ενώ σε λεξικά γενικής χρήσης μπορεί να εντοπίσει κανείς περίπου 200 λήμματα χαρακτηρισμένα ως άσεμνα, ταμπού, χυδαία ή υβριστικά.

Όπως είναι αναμενόμενο, στο πλαίσιο της συγκεκριμένης έρευνας δεν είναι εφικτό να συγκεντρώσουμε, να καταγράψουμε και να τεκμηριώσουμε λεξικογραφικά το σύνολο των λημμάτων που εμπεριέχονται στις διαθέσιμες πηγές. Για αυτό το λόγο θέσαμε ως βασικό κριτήριο την συχνότητα εμφάνισης των λημμάτων. Συγκεκριμένα, όσα από τα στοιχεία δεν εμφανίζονται τουλάχιστον σε πέντε αναφορές σε σώματα κειμένων ή μέσω αναζήτησης στον παγκόσμιο ιστό (google) δεν θα καταχωρίζονται στη βάση (π.χ. *αιδοισέλ*, *αϊποντιά*, *αδικοκούτι*, *σουχλοφάητα*, *στρογγυλός ντοσιές*, *μαγκουρογαμόσαυρος*, *μουνιαγάρας*, *μουσαντογλειψού*). Το κριτήριο αυτό είναι πιθανό να δημιουργήσει προβλήματα στην καταχώριση και ανάλυση στοιχείων από παλαιότερα περιθωριακά λεξιλόγια. Για παράδειγμα, λέξεις από τα καλιαρντά που εντοπίζονται και καταχωρίζονται στο ομότιτλο λεξικό του Πετρόπουλου (1971) εμφανίζονται σε σώματα κειμένων ή μέσω αναζήτησης στο διαδίκτυο ελάχιστες φορές ή/και καμία. Ας πάρουμε, για παράδειγμα, τη λέξη «*σταρομολότσαρδο*»<sup>4</sup> η οποία καταχωρίζεται στο λεξικό του Πετρόπουλου αλλά δεν εμφανίζεται σε κανένα αποτέλεσμα μέσω αναζήτησης στο διαδίκτυο και σε σώματα κειμένων. Τέτοιες περιπτώσεις, όπως το «*σταρομολότσαρδο*», σκοπεύουμε να καταχωριστούν στη βάση σε δεύτερο στάδιο, στην προσπάθειά μας να διαφυλάξουμε τα παλαιότερα περιθωριακά λεξιλόγια που δεν χρησιμοποιούνται στις μέρες μας και να τα γνωστοποιήσουμε στο ευρύ κοινό. Όπως είναι αναμενόμενο, οι πληροφορίες που θα δίνονται για τα συγκεκριμένα λήμματα θα περιορίζονται στις διαθέσιμες από τις υπάρχουσες πηγές.

Αναφορικά με τις αλληλοεπικαλύψεις στοιχείων του περιθωριακού λεξιλογίου, στοιχεία δηλαδή που εμφανίζονται σε περισσότερες από μία πηγές, στη βάση θα εμφανίζονται μόνο μία φορά, αντλώντας πληροφορίες για τη μικροδομή από όλες τις πηγές. Περιπτώσεις λέξεων που εντοπίζονται σε διαφορετικά λεξιλόγια και με διαφορετικές σημασίες (πολύσημα λήμματα) θα εμφανίζονται ως ένα λήμμα και στη μικροδομή θα δηλώνονται με χρονολογική σειρά τα διαφορετικά λεξιλόγια εμφάνισης. Για παράδειγμα, το λήμμα *τάπα* εντοπίζεται στο λεξιλόγιο της πιάτσας με διαφορετικές σημασίες: (α) ως χαρακτηρισμός για κοντό άνθρωπος, (β) ως χαρακτηρισμός για μεθυσμένο<sup>5</sup> και (γ) για να δηλώσει μια

<sup>2</sup> Στον ιστότοπο [www.slang.gr](http://www.slang.gr) καταγράφονται λέξεις και φράσεις από παλαιότερα και σύγχρονα περιθωριακά λεξιλόγια με ερμηνεία της σημασίας τους και κατασκευασμένα παραδείγματα, αλλά χωρίς λεξικογραφική τεκμηρίωση. Τα ετυμολογικά, μορφολογικά κ.ά. σχόλια για κάθε λέξη/φράση ανήκουν στους εγγεγραμμένους χρήστες, συνήθως, μη γνώστες (βλ. και Xydopoulos, Iordanidou & Efthymiou 2011).

<sup>3</sup> Διαθέσιμο στο: <https://www.youtube.com/watch?v=sJqpyb1pBoI>.

<sup>4</sup> «Η μπυραρία» (Πετρόπουλος 1971).

<sup>5</sup> Κάτος (2016).

<sup>6</sup> Ζάχος (1999: 471).



αποστομωτική απάντηση κάποιου στον συνομιλητή του<sup>7</sup>. Η ίδια λέξη στο λεξιλόγιο των νέων χρησιμοποιείται και (δ) για να δηλώσει είδος ερωτικού παιχνιδιού. Η ίδια λέξη εμφανίζεται και στο λεξιλόγιο του μπάσκετ με διαφορετική σημασία. Συγκεκριμένα, αναφέρεται (ε) στο κόψιμο της ανοδικής πορείας της μπάλας προς το καλάθι από έναν αμυντικό παίκτη. Στη βάση θα αναφέρονται και τα τρία λεξιλόγια εμφάνισης του λήμματος κατά χρονολογική σειρά, όπου αυτό είναι εφικτό να εντοπιστεί από τις διαθέσιμες πηγές (βλ. εν. 2.3, δ). Επίσης, περιπτώσεις ομωνυμίας, όπως το ακρωνύμιο *ΤΑΙΙΑ < Του Αγίου Πούτσου Ανήμερα*, από το λεξιλόγιο των φαντάρων, που δηλώνει την καθυστερημένη μέρα απόλυσης από τον στρατό, θα καταχωρίζονται ως διαφορετικά λήμματα, χωρίς κάποιο διακριτικό σημάδι στη μικροδομή.

### 2.3 Οργάνωση μικροδομής

Η μικροδομή του κάθε λήμματος αναμένεται να περιέχει τις ακόλουθες πληροφορίες: (α) λέξη-κεφαλή/λήμμα, (β) γραμματική κατηγορία, (γ) φωνητικό τύπο, (δ) λεξιλόγιο εμφάνισης, (ε) τρόπος σχηματισμού, (στ) εναλλακτικοί τύποι, (ζ) ειδικοί γραμματικοί τύποι, (η) σημασία, (θ) προέλευση, (ι) συχνότητα εμφάνισης, (ια) παρωχημένο ή μη, (ιβ) παράγωγες λέξεις, (ιγ) σύνθετες λέξεις, (ιδ) φράσεις/ιδιωματισμοί και (ιε) σχόλια.

Πιο αναλυτικά, τα παραπάνω στοιχεία της μικροδομής αναμένεται να έχουν την ακόλουθη δομή. Ενδεικτικά και για την καλύτερη κατανόηση της μικροδομής, στις εικόνες 1α-γ παραθέτουμε τις πληροφορίες της μικροδομής του λήμματος «μαλάκας», όπως αναμένεται να εμφανίζονται στον χρήστη<sup>8</sup>.

(α) Η λέξη-κεφαλή/λήμμα θα δίνεται με έντονα, μικρά γράμματα στον ουδέτερο τύπο.

(β) Θα δηλώνεται η γραμματική κατηγορία, π.χ. ουσιαστικό, ρήμα, επίθετο, επίρρημα κ.ά., χωρίς τη χρήση συντομογραφιών<sup>9</sup> για λήμματα που ανήκουν σε περισσότερες από μία κατηγορίες θα γίνεται αναφορά σε όλες.

(γ) Ο φωνητικός τύπος/προφορά θα δίνεται ανάμεσα σε αγκύλες με φωνητική μεταγραφή του λήμματος, σύμφωνα με το Διεθνές Φωνητικό Αλφάβητο. Επάνω από το τονιζόμενο φωνήεν, θα υπάρχει τονικό σημάδι (βλ. εικόνα 1)<sup>10</sup>.

(δ) Θα υπάρχει η πληροφορία για το λεξιλόγιο εμφάνισης, ένα ή περισσότερα, ανάλογα με τις πληροφορίες από τις διαθέσιμες πηγές. Αν ένα λήμμα εμφανίζεται σε περισσότερα από ένα λεξιλόγια αυτά θα δίνονται με σειρά εμφάνισης.

(ε) Ο τρόπος σχηματισμού του λήμματος θα δηλώνεται με αναφορά στη μορφολογική διαδικασία σχηματισμού, π.χ. παραγωγή, σύνθεση, σύμφυση, περικοπή, φράση κ.ά..

(στ) Το πεδίο «εναλλακτικοί τύποι» θα αξιοποιείται μόνο για τα λεξήματα που εμφανίζονται με εναλλακτική ορθογραφία ή σε εναλλακτικό τύπο, π.χ. *γκολκίπερ*, *γκολκήπερ* ή *γλίτσας*, *γλίτσης*, *γλίτζας*.

(ζ) Στο πεδίο «ειδικοί γραμματικοί τύποι» θα εμφανίζονται εκείνοι οι γραμματικοί τύποι των λημμάτων που παρουσιάζουν ιδιαιτερότητες/αποκλίσεις ως προς τον σχηματισμό τους<sup>11</sup> τύποι, δηλαδή, που δεν προβλέπονται από τα κλιτικά παραδείγματα.

λήμμα	μαλάκας
γραμματική κατηγορία	ουσιαστικό
φωνητικός τύπος	[malákas]
λεξιλόγιο εμφάνισης	άσεμνο λεξιλόγιο
τρόπος σχηματισμού	-
εναλλακτικοί τύποι	-
ειδικοί γραμματικοί τύποι	μαλάκηδες, μαλάκηδων

Εικόνα 1α: Παράδειγμα μικροδομής.

(η) Η σημασία ή οι διαφορετικές σημασίες των λημμάτων θα δίνονται με αρίθμηση και για κάθε σημασία θα υπάρχει ακριβώς από κάτω παράδειγμα είτε από σώματα κειμένων είτε από το διαδίκτυο. Στα δεξιά θα εμφανίζεται η πηγή του κάθε παραδείγματος, στην οποία θα έχει πρόσβαση ο χρήστης «πατώντας» στον υπερσύνδεσμο (βλ. εικ. 1β). Επίσης, θα γίνεται αναφορά πριν τη σημασία στον σημασιολογικό μηχανισμό που εμπλέκεται (μεταφορά, μετωνυμία κ.λπ.). Κάτω

<sup>7</sup> Ό.π..


<sup>8</sup> Ενδέχεται να υπάρχουν αλλαγές ως προς την δομή και οργάνωση των πληροφοριών.

<sup>9</sup> Για παράδειγμα, στην ονομαστική ενικού για τα ουσιαστικά, στο α' ενικό της οριστικής του ενεστώτα στην ενεργητική φωνή για τα ρήματα, στα τρία γένη για τα επίθετα κ.ό.κ..

<sup>10</sup> Για τις μονοσύλλαβες λέξεις που στη γραφή τους μορφή δεν έχουν τόνο, θα εμφανίζεται τονικό σημάδι στη συλλαβή που ακούγεται το δυνάμωμα της έντασης στη φωνή, π.χ. /kál/, ακολουθώντας τη λογική του ΑΚΝ.



από κάθε σημασία θα δίνονται τα συνώνυμα<sup>11</sup> και αντίθετα που συνδέονται με την εκάστοτε σημασία του λήμματος. Σε ξεχωριστό πεδίο θα καταγράφονται τα διαθέσιμα διαλεκτικά συνώνυμα. Στην ίδια λογική, κάτω από κάθε σημασία, θα δίνεται πολυμεσικό υλικό<sup>12</sup> που θα αναδεικνύει την εκάστοτε σημασία του λήμματος και δίπλα ο σύνδεσμος που θα οδηγεί τον χρήστη στην πηγή. Μέσα στο πεδίο της σημασίας αναμένεται να υπάρχει και πεδίο για τον βαθμό προσβλητικότητας της κάθε σημασίας. Η συγκεκριμένη πληροφορία θα είναι διαθέσιμη για κάθε σημασία, εφόσον, ως γνωστό, οι διαφορετικές σημασίες των λέξεων αποδίδουν και διαφορετικό βαθμό προσβλητικότητας (π.χ. *μαλάκας*, *σκατά*)<sup>13</sup>. Το συγκεκριμένο πεδίο, που δεν αναμένεται να αξιοποιηθεί στην πρώτη φάση της έρευνας, θα αναφέρεται στο αν και πόσο ένα στοιχείο κρίνεται ότι προσβάλλει τον χρήστη, μέσα από μία κλίμακα που θα διαμορφωθεί από τους εγγεγραμμένους χρήστες στη βάση ή/και έρευνα που θα διεξάγουμε με τη χρήση ερωτηματολογίου.

σημασία	1. (κυριολεκτικά) πρόσωπο που συνανιέται
	<i>Μαλάκας</i> ονομάζεται ο ανίκανος να κάνει έρωτα με γυναίκα λόγω ατολμίας και γι' αυτό καταφεύγει στην αυτοϊκανοποίηση. <span style="float: right;">Πηγή</span>
	<b>ΣΥΝΩΝΥΜΑ</b> ανανιστής, ξεφλουδάς, πουλοπαίχτης, πουλοπλέρ, τρομπέο, μοναχικός καβαλάρης, τρομπαδούρος
	<b>ΔΙΑΛΕΚΤΙΚΑ ΣΥΝΩΝΥΜΑ</b> μινάρας, γρόθος, τρόμπας
	<b>ΑΝΤΙΘΕΤΑ</b> -
	<b>ΒΑΘΜΟΣ ΠΡΟΣΒΛΗΤΙΚΟΤΗΤΑΣ</b> -
	<b>ΠΟΛΥΜΕΣΙΚΟ ΥΛΙΚΟ</b> -
	2. (μεταφορά) χαρακτηρισμός για κάποιον που πέφτει θύμα (χρησιμοποιείται για άνδρα και γυναίκα)
	α. <i>Είμαι μαλάκας</i> να πληρώνω φόρους για να κάθεται εσύ σπίτι με το μωρό; <span style="float: right;">Πηγή</span>
	β. <i>Μόνη μου εντυπωσιάζομαι ο μαλάκας</i> . Και δεν φτάνει αυτό. Όσοχι. Τη λέω και στους άλλους. <span style="float: right;">Πηγή</span>
	<b>ΣΥΝΩΝΥΜΑ</b> βλάκας, χαζός, κορόιδο
	<b>ΔΙΑΛΕΚΤΙΚΑ ΣΥΝΩΝΥΜΑ</b> μινάρας, γρόθος, τρόμπας
	<b>ΑΝΤΙΘΕΤΑ</b> -
	<b>ΒΑΘΜΟΣ ΠΡΟΣΒΛΗΤΙΚΟΤΗΤΑΣ</b> -
	<b>ΠΟΛΥΜΕΣΙΚΟ ΥΛΙΚΟ</b>
	 <span style="float: right;">Πηγή</span>
	3. υβριστικός χαρακτηρισμός για κάποιον που λέει ή κάνει ανοησίες
	<i>Είμαι σίγουρος ότι όλοι οι Ελληνοαμερικάνοι ακούνε αυτή την περίοδο πολλά κακόγουστα αστεία για την Ελλάδα.</i> <span style="float: right;">Πηγή</span>
	<i>Σε κάτι τέτοιους μαλάκες αναγκάζομαι να υπενθυμίσω δυο πραγματάκια για το τι εστί δημοκρατία και πολιτισμός.</i>
	<b>ΣΥΝΩΝΥΜΑ</b> ηλίθιος
	<b>ΔΙΑΛΕΚΤΙΚΑ ΣΥΝΩΝΥΜΑ</b> μινάρας, γρόθος, τρόμπας
	<b>ΑΝΤΙΘΕΤΑ</b> -
	<b>ΒΑΘΜΟΣ ΠΡΟΣΒΛΗΤΙΚΟΤΗΤΑΣ</b> -
	<b>ΠΟΛΥΜΕΣΙΚΟ ΥΛΙΚΟ</b> <a href="#">Απαράδεκτοι «Δεν είμαι μαλάκας»</a>
	4. οικεία προσφώνηση
	<i>Κι αφού είδαν και απέιδαν ότι με ανθρώπους όλων των ηλικιών με θλιμμένα πρόσωπα και σφηνισμένη ψυχή</i> <span style="float: right;">Πηγή</span>
	<i>δε γινόταν πάρτυ της προκοπής, κάποιος απ' τη φρουρά έριξε την ιδέα ρε μαλάκες δεν πάμε να αράζουμε Εξάρχεια;</i>
	<b>ΣΥΝΩΝΥΜΑ</b> φίλος
	<b>ΔΙΑΛΕΚΤΙΚΑ ΣΥΝΩΝΥΜΑ</b> μινάρας, γρόθος, τρόμπας
	<b>ΑΝΤΙΘΕΤΑ</b> -
	<b>ΒΑΘΜΟΣ ΠΡΟΣΒΛΗΤΙΚΟΤΗΤΑΣ</b> -
	<b>ΠΟΛΥΜΕΣΙΚΟ ΥΛΙΚΟ</b> <a href="#">Η Ψυχή Στο Στάμι - 2006</a>

Εικόνα 1β: Παράδειγμα μικροδομής: σημασία.

(θ) Στο πεδίο «προέλευση» αναμένεται να ενσωματωθούν πληροφορίες για κάθε λήμμα από διαθέσιμες πηγές και λεξικά που περιέχουν κάποια από τα λήμματα του περιθωριακού λεξιλογίου και κυρίως του άσεμνου λεξιλογίου. Σε αυτή την περίπτωση θα γίνεται σύνδεση με την αντίστοιχη βάση, εφόσον διατίθεται σε ψηφιακή μορφή, με τη χρήση υπερσυνδέσμων, π.χ. *Λεξικό του Ιδρύματος Τριανταφυλλίδη* (1998), *Λεξικό λαϊκής και περιθωριακής γλώσσας*, Κάτος (2016) ή αναφορά στην έντυπη πηγή. Επίσης, για τα λήμματα που είναι δάνεια και υπάρχει η αντίστοιχη πληροφορία στις διαθέσιμες πηγές, αυτή θα δηλώνεται με παράλληλη αναφορά στην πηγή.

(ι) Οι πληροφορίες για τη συχνότητα εμφάνισης θα προέρχονται από τα σώματα κειμένων που υπάρχουν διαθέσιμα στην πλατφόρμα Sketch Engine και από το διαδίκτυο με αναφορά σε αριθμό εμφάνισης αποτελεσμάτων για το κάθε λήμμα και αναφορά στην ημερομηνία που έγινε η μέτρηση, καθώς η συχνότητα εμφάνισης είναι πιθανό να διαφέρει από μέρα σε

<sup>11</sup> Πρόκειται για λογοκυπτοασιακά συνώνυμα ή πλησιώνυμα (Cruse 1986: 265-288; 2004: 154-157; Ξυδόπουλος 2008: 127-133).

<sup>12</sup> Το πεδίο για το πολυμεσικό υλικό αναμένεται να μην αξιοποιηθεί σε αρκετές περιπτώσεις, λόγω της περιορισμένης εμφάνισης του συγκεκριμένου λεξιλογίου σε διαθέσιμο οπτικοακουστικό υλικό.

<sup>13</sup> βλ. σχετικά Christopoulou, Xydopoulos & Tsangalidis (2017); Ξυδόπουλος & Χριστοπούλου (υ.έ.).



μέρα, από εβδομάδα σε εβδομάδα, κ.ο.κ.. Για τις εμφανίσεις στο διαδίκτυο θα δίνεται ο ακριβής αριθμός εμφανίσεων, ενώ για την πλατφόρμα Sketch Engine, από την οποία θα αξιοποιηθεί το σώμα κειμένων *eITenTen* (2014), θα υπάρχει αναφορά στον αριθμό που αντιπροσωπεύει τη συχνότητα εμφάνισης ανά εκατομμύριο και στον αριθμό εμφάνισης κάθε λήμματος. Για τη μέτρηση της συχνότητας θα χρησιμοποιείται ο ουδέτερος τύπος του λήμματος<sup>14</sup>. Επίσης, άλλοι γραμματικοί τύποι, εναλλακτικοί ορθογραφικοί τύποι ή περιπτώσεις γραμμένες με *greeklish* δεν θα λαμβάνονται υπόψη στη μέτρηση της συχνότητας.

(ια) Το πεδίο «παρωχημένο ή μη» θα αναφέρεται στη χρήση του λήμματος και όχι σε κάθε επιμέρους σημασία. Όπως είναι αναμενόμενο, τα στοιχεία που συγκροτούν τα παλαιότερα περιθωριακά λεξιλόγια θα χαρακτηριστούν ως παρωχημένα. Ενδιαφέρον θα έχει να ελέγξουμε την εικόνα που παρουσιάζουν τα στοιχεία από τα σύγχρονα περιθωριακά λεξιλόγια. Το συγκεκριμένο πεδίο αναμένεται να συμπληρωθεί σε δεύτερο στάδιο, αφού θα έχουν συγκεντρωθεί τα ποσοτικά στοιχεία για τη συχνότητα εμφάνισης των λημμάτων και θα έχουμε αποκτήσει μία συνολική εικόνα που θα μας επιτρέψει να εντοπίσουμε στοιχεία που είναι παρωχημένα ή μη.

(ιβ, ιγ, ιδ) Θα δίνονται παράγωγες, σύνθετες λέξεις και φράσεις/ιδιωματισμοί με τη λέξη-κεφαλή/λήμμα, οι οποίες θα υπάρχουν και ως ανεξάρτητα λήμματα. Ο χρήστης θα έχει τη δυνατότητα να ανατρέξει σε αυτά με τη χρήση υπερσυνδέσμου. Ειδικότερα, στο πεδίο «φράσεις/ιδιωματισμοί»<sup>15</sup> θα εντάξουμε και εκείνα τα φραστικά σύνθετα, που περιέχουν το λήμμα και εμφανίζουν υψηλή συχνότητα. Κάθε φράση θα αναλύεται και ως ανεξάρτητο λήμμα με όλες οι διαθέσιμες πληροφορίες της μικροδομής. Ο χρήστης θα έχει τη δυνατότητα να μεταβεί στο λήμμα της φράσης «πατώντας» τον υπερσυνδέσμο. Αν μία φράση περιέχει μία ή και περισσότερες λέξεις του περιθωριακού λεξιλογίου θα γίνεται αναφορά σε όλες (π.χ. η ιδιωματική φράση *πίπα κώλος* θα υπάρχει στο πεδίο «φράσεις/ιδιωματισμοί» του λήμματος *πίπα* και του λήμματος *κώλος*).

(ιε) Στο τελευταίο πεδίο θα φιλοξενοούνται τα σχόλια των εγγεγραμμένων χρηστών, που θα έχουν, μεταξύ άλλων, τη δυνατότητα να προσθέτουν πληροφορίες, να προτείνουν την προσθήκη νέων λημμάτων και να συμμετέχουν σε έρευνες.

προέλευση	(μαλάκ(α) η 'μαλάκωση' -ας < ελνστ. <i>μαλακ(ός)</i> 'παθητικός ομοφυλόφιλος' -α (αναδρ. σχημ.), με αλλ. της σημ. κατά το <i>μαλακία</i> : <i>μαλάκ(ας)</i> -ούλης) Πηγή: <i>AKN (Ιδρυμα Μανόλη Τριανταφυλλίδη, 1998)</i>
συχνότητα εμφάνισης	Σώματα Κειμένων: 9,13 ανά εκατομμύριο Εμφανίσεις: 17.911 Google: 422.000 Ημερομηνία ελέγχου: 29/1/2020
παρωχημένο ή μη	-
παράγωγες λέξεις	<i>μαλακία, μαλακίζομαι, μαλακισμένος, μαλακιστήρι, μαλακάκος, μαλάκαρος, μαλάκοβιτς</i>
σύνθετες λέξεις	<i>ψύομαλάκας, αρχιμαλάκας, μαλακοβιόλα, μαλακοκυρίδες, μαλακόβιος, μαλακογάμης, μαλακοδείχνο, μαλακοκαύλης, μαλακοκεφτές, μαλακοπούνκομα, μαλακοπίτουρας, ηθικομαλάκας, χοντρομαλάκας, ντερνετομαλάκας, μαλακόπουστας, μαλακοπουτοσγλειφτρα, μαλακόπτηνος, μαλακόφατσα, μαλακοφένω</i>
φράσεις/ιδιωματισμοί	<i>κάνω τον μαλάκα, μαλάκας με περκεφαλαία, μασίφ μαλάκας</i>
σχόλια χρηστών	

Εικόνα 1γ: Παράδειγμα μικροδομής.

## 2.4 Σώματα Κειμένων με τη χρήση του Sketch Engine

Όπως αναφέραμε και στην προηγούμενη ενότητα (βλ. εν. 2.3), για την τεκμηρίωση συγκεκριμένων πληροφοριών της μικροδομής των λημμάτων θα αξιοποιήσουμε σώματα κειμένων, τα οποία προσφέρουν μια πληρέστερη αλλά και ασφαλέστερη εικόνα για το πώς χρησιμοποιείται κάθε λέξη μέσα από αυθεντικά κείμενα (Γούτσος & Φραγκάκη 2015: 185). Τα σώματα κειμένων που θα αξιοποιήσουμε προέρχονται από την εξειδικευμένη ψηφιακή πλατφόρμα *Sketch Engine* και συγκεκριμένα το σώμα κειμένων της ελληνικής, *eITenTen* (2014), που περιέχει περισσότερες από 1,5 δισεκατομμύριο λέξεις από διαφορετικά περιβάλλοντα χρήσης, επίσημου και ανεπίσημου λόγου (Kilgarriff & Grefenstette 2003; Kilgarriff *et al.* 2014).

Το ιδιαίτερο πλεονέκτημα του *Sketch Engine* είναι ότι δίνει τη δυνατότητα για την εξαρχής δημιουργία σωμάτων κειμένων με υλικό από το διαδίκτυο, χωρίς περιορισμούς. Το υλικό αυτό ανακτάται με λέξεις-κλειδιά ή συγκεκριμένα URL ή από ψηφιοποιημένα κείμενα. Στην αρχική φάση της έρευνάς μας θα αξιοποιηθούν τα διαθέσιμα σώματα κειμένων από το λογισμικό *Sketch Engine*, καθώς υπάρχουν και ανεπίσημα, μη τυπικού ύφους, κείμενα, που καλύπτουν τις ανάγκες μας. Σε μεταγενέστερη φάση, ενδεχομένως να συγκροτηθεί παραμετροποιημένο σώμα κειμένων (*taylor-made*), που θα δημιουργηθεί από κείμενα ανεπίσημου λόγου, από εφημερίδες, περιοδικά, *blogs*, ιστότοπους, αναρτήσεις σε μέσα κοινωνικής δικτύωσης, διαλόγους από περιθωριακές ομάδες ποικίλης θεματολογίας (π.χ. αθλητισμός, μόδα, *lifestyle*), αλλά και κείμενα διαφορετικών χρονικών περιόδων. Τα παραμετροποιημένα σώματα κειμένων θα είναι διαθέσιμα και εκτός πλαισίου της συγκεκριμένης έρευνας.

<sup>14</sup> Βλ. υποσ. 9.

<sup>15</sup> Δεν αναμένεται να ενταχθούν στο πεδίο αυτό και παροιμίες με στοιχεία του περιθωριακού λεξιλογίου.



[illegible][illegible]

### 3 Καινοτομίες της βάσης

[www.euralex2020.gr](http://www.euralex2020.gr)



απευθύνεται και σε ερευνητές που μελετούν και αναλύουν τη γλώσσα, δίνοντάς τους τη δυνατότητα να εξάγουν σημαντικά συμπεράσματα αν αξιοποιήσουν συνδυαστικά πληροφορίες της μικροδομής (π.χ. γραμματική κατηγορία, λεξιλόγιο εμφάνισης, μορφολογική διαδικασία σχηματισμού κ.ά.).

## Βιβλιογραφία

- Ανδρουτσόπουλος, Γ. (1998). Γλώσσα των νέων και γλωσσική αγορά της νεανικής κουλτούρας. Στο *Μελέτες για την ελληνική γλώσσα. Πρακτικά της 18ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του Α.Π.Θ.*, 41-53. Θεσσαλονίκη.
- Ανδρουτσόπουλος, Γ. (2001). *Γλώσσα των νέων*. Εγκυκλοπαιδικός Οδηγός για τη Γλώσσα (επιμ. Α.-Φ. Χριστίδης, σε συνεργασία με Μ. Θεοδωροπούλου). Θεσσαλονίκη: Κέντρο Ελληνικής Γλώσσας, 108-113.
- Ανδρουτσόπουλος, Γ. (2011). Στην... Οξφόρδη η αργκό του Ίντερνέτ: Το κορυφαίο λεξικό ενέταξε λέξεις από την «γλώσσα» των νέων. *Τα Νέα*, 28/3/2011.
- Ανδρουτσόπουλος, Γ., Κακριδή, Μ. (2010). Γλώσσα των νέων. Αναγνώριση, αποδοχή και κριτική των νεανικών ιδιωμάτων. Στο Β. Βαμβακάς, Π. Παναγιωτόπουλος (επιμ.), *Η Ελλάδα στη δεκαετία του '80. Κοινωνικό, πολιτικό και πολιτισμικό λεξικό*. Αθήνα: Το πέρασα, 86-89.
- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London/New York: Continuum.
- Burke, S.M. (2003). The design of online lexicons. In Sterkenburg, P. van (ed.) *A practical guide to lexicography*. Amsterdam: John Benjamins, 240-249.
- Christopoulou, K., Xydopoulos, G.J. & Tsangalidis, A. (2017). Grammatical gender and offensiveness in Modern Greek slang vocabulary. In *Proceedings of the 12th International Conference on Greek Linguistics Vol. 1 (ICGL 12)*, Βερολίνο, 291-305.
- Γούτσος, Δ., Φραγκάκη, Γ. (2015). *Εισαγωγή στη γλωσσολογία σωμάτων κειμένων*. Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.
- Cruse, A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Cruse, A. (2004). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.
- Dodd, W.S. (1989). Lexic computing and the dictionary of the future. In James, G. (ed.) *Lexicographers and their words*, 89-93. Exeter: University of Exeter.
- Duckworth, T. (χ.χ.) *A dictionary of slang: English slang and colloquialisms used in the United Kingdom*. Διαθέσιμο στο: <http://www.peevish.co.uk/slang/> [10/5/2020].
- Fellbaum, C. (2014). Large-scale Lexicography in the Digital Age, *International Journal of Lexicography*, 27(4), 378-395.
- Forrest, C. (2011). *International Law and the Protection of Cultural Heritage*. London: Routledge.
- Heuberger, R. (2016). Corpora as game changers: The growing impact of corpus tools for dictionary makers and users. *English today*, 32(2), 24-30.
- Hoffman, B.T. (ed.). (2009). *Art and Cultural Heritage. Law, Policy and Practice*. Cambridge, Cambridge University Press.
- Ίδρυμα Μανόλη Τριανταφυλλίδη. (1998). *Λεξικό της Κοινής Νεοελληνικής*. Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών, Ίδρυμα Μανόλη Τριανταφυλλίδη.
- Καμηλάκη, Μ., Κατσούδα, Γ. & Βραχιονίδου, Μ. (2015). *Πιπέρι στο στόμα: όψεις των λέξεων-ταμπού στη Νέα Ελληνική*. Εκδόσεις: Καλλιγράφος.
- Κάτος, Γ. (2016). *Λεξικό της Λαϊκής και Περιθωριακής Γλώσσας*. Κέντρο Ελληνικής Γλώσσας. Διαθέσιμο στο: <http://georgakas.lit.auth.gr/dictionaries/index.php> [5/5/2010].
- Κατσίκης, Ι., Σπυρόπουλος, Δ. (1999). *Το αλφαβητάρι της γλώσσας των νέων*. Αθήνα: Δίαυλος.
- Κεχαγιά, Κ. (1997). Το “άσεμνο” λεξιλόγιο στα νέα ελληνικά: λεξιλογική και πραγματολογική προσέγγιση. Στο *Μελέτες για την Ελληνική Γλώσσα. Πρακτικά της 17ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας της Φιλοσοφικής Σχολής του Α.Π.Θ.*, 22-24 Απριλίου 1996. Θεσσαλονίκη, 591-602.
- Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3), 333-347.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography* 1(1), 7-36.
- Landau, S. I. (2001). *Dictionaries: The Art and Craft of Lexicography* (2nd Edition). Cambridge University Press.
- Μακροδήμος, Ν., Ακριτίδης, Χ. (2007). *Λεξιλόγιο της αργκό*. Αθήνα: Δρόμων.
- Μπαμπινιώτης, Γ. (2002). *Λεξικό της Νέας Ελληνικής Γλώσσας* (2η Έκδοση). Αθήνα: Κέντρο Λεξικολογίας.
- Oppentocht, L. & Schutz, R. (2003). Developments in electronic dictionary design. In Sterkenburg, P. van (ed.) *A practical guide to lexicography*. Amsterdam: John Benjamins, 215-227.
- Pautasso, C., Zimmermann, O., Leymann, F. (2008). RESTful Web Services vs. “Big” Web Services: Making the Right Architectural Decision. In *International World Wide Web Conference Com- mittee (IW3C2)*, 805-814, April 21-25, 2008, Beijing, China.
- Πετρόπουλος, Η. (1971). *Καλιαντά*. Αθήνα: Δίγαμμα.
- Rodriguez, A. (2008). RESTful Web services: The basics. *Developer Works. IBM*. Διαθέσιμο στο: <http://www.gregbulla.com/TechStuff/Docs/ws-restful-pdf.pdf> [14/5/2020].
- Σπηλιώτης, Κ., Φραγκιαδάκης, Γ. (2007). Φιδιάζει νέος; δε σε χάλασε που σ' έχωσαν αγγαρεία μαγειρεία!: Μορφολογικά Χαρακτηριστικά της κοινωνιολέκτου που χρησιμοποιούν οι Έλληνες στρατιώτες. Στο *8ο Διεθνές Συνέδριο Ελληνικής Γλωσσολογίας*, 30 Αυγούστου - 2 Σεπτεμβρίου 2007. Ιωάννινα, 1133-1148.
- Swanepoel, P. (2003). *Dictionary Typologies: A Pragmatic Approach*. Piet van Sterkenburg (Ed.). 2003: 44-69.



- The OSD. (χ.χ.) *The online slang dictionary: American, English and urban slang*. Διαθέσιμο στο: <http://onlineslangdictionary.com/> [10/5/2020].
- Xydopoulos, G.J., Iordanidou, A. & Efthymiou, A. (2011). Recent advances in the documentation of Greek slang: The case of [www.slang.gr](http://www.slang.gr). In K. Hatzopoulou, A. Ioannidou & S. Yoon *Proceedings of the 9th International Conference on Greek Linguistics* (ICGL 9, Chicago, IL, USA, 29 - 31 October 2009), 112-123.
- Ξυδόπουλος, Ι.Γ. (2008). *Λεξικολογία*. Αθήνα: Πατάκης.
- Ξυδόπουλος, Γ. & Χριστοπούλου, Κ. (υ.έ.). Μέτρηση της προσβλητικότητας σε λέξεις του περιθωριακού λεξιλογίου της νέας ελληνικής: εκτιμώντας τα πρώτα αποτελέσματα. In *Proceedings of the 14th International Conference on Greek Linguistics* (ICGL 14). Πάτρα.
- Χαραλαμπίκης, Χ. (επιμ.) (2014). *Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας*. Αθήνα: Ακαδημία Αθηνών.
- Χρηστάκης, Λ. & Επάριτος, Μ. (1995). *Λεξικό της ντάγκλας*. Αθήνα: Όπερα.
- Χρήστου, Σ., Στάμου, Α. (2013). Αναπαραστάσεις της στρατιωτικής κοινωνιολέκτου στον ελληνικό κινηματογράφο. *Multilingual Academic Journal of Education and Social Sciences* 1(2): 15-24.
- Χριστοπούλου, Κ. (2016). “Μια λεξικολογική προσέγγιση στο περιθωριακό λεξιλόγιο της Νέας Ελληνικής”. Αδημοσίευτη Διδακτορική Διατριβή. Πανεπιστήμιο Πατρών.
- Ζάχος (Παπαζαχαρίου), Ε. (1999). *Λεξικό της πιάτσας*. Β' Έκδοση. Αθήνα: Κάκτος.

**Η παρούσα έρευνα συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση Μεταδιδασκόντων ερευνητών/ερευνητριών - Β' Κύκλος» (MIS-5033021), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (ΙΚΥ).**









# «Τα σταλθέντα ή τα σταλμένα μηνύματα;» – απολιθώματα των αρχαίων μετοχών στα σύγχρονα λεξικά και στα σώματα κειμένων

Ιορδανίδου Α.

Τμήμα Επιστημών της Εκπαίδευσης και Κοινωνικής Εργασίας, Πανεπιστήμιο Πατρών, A.Iordanidou@patras.gr

## Περίληψη

Η παρούσα εργασία αναφέρεται στις αρχαιοελληνικές μετοχές παθητικού αορίστου που δεν εντάσσονται στο ρηματικό σύστημα της νέας ελληνικής, π.χ. *εκτελεσθείς, εξαχθείς, κλαπείς, λεχθείς, προβλεφθείς*. Στο ρηματικό σύστημα της αρχαίας ελληνικής οι μετοχές δήλωναν χρόνο (ενεστώτα, μέλλοντα, αόριστο, παρακειμένο), ενώ στη νέα ελληνική οι ελάχιστες που έχουν απομείνουν στον ενεστώτα και στον παρακειμένο δηλώνουν κατεξοχήν τρόπο ενέργειας (εξακολουθητικό – συντελεσμένο). Ουσιαστικά πρόκειται για την ενεργητική επιρρηματική μετοχή ενεστώτα σε -οντας (ή γερούνδιο, σύμφωνα με ορισμένους μελετητές), π.χ. *επιλέγοντας*, την παθητική μετοχή ενεστώτα σε ορισμένα μόνο ρήματα, π.χ. *επιλεγόμενος*, και την παθητική μετοχή παρακειμένου, π.χ. *επιλεγμένος*, η οποία σε αρκετές περιπτώσεις έχει λειτουργία ουσιαστικού ή επιθέτου. Η έρευνα στα κυριότερα σύγχρονα ελληνικά λεξικά (ΛΚΝ, ΛΝΕΓ και ΧΛΝΓ) αναδεικνύει διαφορετικές πρακτικές καταγραφής, ως επιθέτων ή ουσιαστικών στο ΛΚΝ και ως μετοχών παράλληλα με τις μετοχές παθητικού παρακειμένου στα ΛΝΕΓ και ΧΛΝΓ σε συγκεκριμένα ρήματα. Εξετάζεται αν και σε ποιον βαθμό οι λεξικογραφικές αυτές πρακτικές μπορούν να υποστηριχθούν από τα δεδομένα της γλωσσικής χρήσης, όπως προκύπτει από αναζήτηση σε σώματα κειμένων.

**Λέξεις-κλειδιά:** τυποποίηση· γλωσσικό κύρος· γλωσσική χρήση· σώματα κειμένων· νεοελληνικά λεξικά

## 1 Εισαγωγή

Η παρούσα εργασία εντάσσεται στο πλαίσιο ερευνών με αφετηρία τη διατριβή μου (Iordanidou 1985) που αφορούν την ανίχνευση της γλωσσικής πρακτικής μέσω καταγραφής δεδομένων από διάφορα κειμενικά είδη, σε σχέση με τη διαδικασία της τυποποίησης (standardisation) της νέας ελληνικής, που εγκαινιάστηκε το 1976 μετά από μακρά περίοδο διγλωσσίας. Το ερώτημα που τίθεται είναι αν η επιβίωση ορισμένων αρχαίων μετοχών παθητικού αορίστου (*σταλθέντα*), παρά την απουσία του είδους αυτού των μετοχών από το γραμματικό σύστημα της σύγχρονης νεοελληνικής, πρέπει να αποτυπωθεί λεξικογραφικά ως μοναδική επιλογή ή ως εναλλακτική προς τις μετοχές παθητικού παρακειμένου (*σταλμένα*) ή αν πρέπει να αποτυπωθεί ως επιθετοποιημένη χρήση σε συγκεκριμένα κειμενικά συμφραζόμενα και κειμενικά είδη.

## 2 Οι μετοχές στο γραμματικό σύστημα της νέας ελληνικής

Όπως καταγράφεται σε μελέτες ιστορικής εξέλιξης της ελληνικής γλώσσας (βλ. αναφορές στο Iordanidou 1985 και Manolossou 2005), το σύστημα των μετοχών της αρχαίας ελληνικής, που περιλάμβανε μετοχές ενεστώτα, μέλλοντα, αορίστου και παρακειμένου και στις τρεις φωνές (ενεργητική, μέση και παθητική), έδωσε τη θέση του στη νεοελληνική ενεργητική μετοχή ενεστώτα σε -οντας (γερούνδιο για ορισμένους μελετητές) και στις μετοχές παθητικού ενεστώτα σε -όμενος (για ορισμένα μόνο ρήματα) και παθητικού παρακειμένου σε -μένος.

Στη «Μεγάλη» Γραμματική ο Τριανταφυλλίδης (1978: 373-375) αναφέρει πως η παθητική μετοχή σε -μένος συχνά ισοδυναμεί με επίθετο, ενώ για τις μετοχές του παθητικού ενεστώτα οι οποίες λήγουν σε -όμενος, -ούμενος και -όμενος επισημαίνει πως συχνά πρόκειται για ρηματικά επίθετα και στις άκλιτες μετοχές του ενεστώτα που λήγουν σε -οντας/-ωντας αποδίδει δήλωση πράξης «που γίνεται εξακολουθητικά σύγχρονα με το σημεινόμενο από το ρήμα». Στην περιγραφή των μετοχών δεν εντάσσονται οι σχηματισμοί σε -ων, -ουσα, -ον (π.χ. λέγων, -ουσα, -ον) του ενεστώτα ενεργητικής φωνής της αρχαίας ούτε οι σχηματισμοί σε -είς, -είσα, -έν (π.χ. λεχθείς, -είσα, -έν) του παθητικού αορίστου. Στη σχολική Γραμματική Ε' και Στ' Δημοτικού (2009: 158-161) οι μετοχές διακρίνονται σε κλιτές (-μένος, -η, -ο), που «λειτουργούν όπως τα επίθετα, δηλ. συνοδεύουν ουσιαστικά και τα προσδιορίζουν», και άκλιτες (-οντας), που «λειτουργούν σαν επιρρήματα, δηλ. δίνουν πληροφορία για τον τρόπο και τον χρόνο που γίνεται η ενέργεια του ρήματος».

Στη Γραμματική Holton et al. (1999: 169 και 235-237) η εικόνα είναι αρκετά περίπλοκη: στο Μέρος Β, Μορφολογία, στο κεφάλαιο Ρήματα, ως συνήθως χρησιμοποιούμενες μετοχές στα νέα ελληνικά αναφέρονται η ενεργητική μετοχή ενεστώτα, με τον χαρακτηρισμό «γερούνδιο», η παθητική μετοχή ενεστώτα και η παθητική «τετελεσμένη» μετοχή, ενώ στο υποκεφάλαιο «Λόγιοι ρηματικοί τύποι» παρατίθενται ως σποραδικά εμφανιζόμενες οι αρχαίες μετοχές ενεργητικού ενεστώτα σε -ών, -ούσα, -όν (π.χ. εκλιπών) και αορίστου σε -ας, -ασα, -αν (π.χ. αποβιώσας), καθώς και η «παθητική παρελθοντική μετοχή αορίστου» σε -είς, -είσα, -έν. Στο Μέρος Γ, Σύνταξη, στο κεφάλαιο Μετοχές, αναφέρονται οι κλιτές μετοχές της ενεργητικής φωνής και η «παθητική παρελθοντική μετοχή αορίστου» ως κληρονομημένες από την καθαρεύουσα, με το σχόλιο ότι «απαντούν στον δημοσιογραφικό λόγο ή στον λόγο που χρησιμοποιεί χαρακτηριστικά της καθαρεύουσας για να επιτύχει υψηλό ύφος, επισιμότητα ή ειρωνεία». Όσον αφορά τη μετοχή παθητικού ενεστώτα σε -όμενος, στη σ. 130 δηλώνεται ότι «εκφράζει μια ενέργεια εν εξελίξει ή επαναλαμβανόμενη ή μια κατάσταση», αλλά



στη σ. 237 συνδέεται με εισαγωγή από την καθαρεύουσα, όπως και η μετοχή σε *-είς, -είσα, -έν*.

Στη Γραμματική Κλαίρη – Μπαμπινιώτη (2005: 542) στο κεφάλαιο «Μετοχή» περιλαμβάνονται οι «μετοχές του μεσοπαθητικού παρακειμένου σε *-μένος, -μένη, -μένο*», ενώ σχολιάζεται ότι σε «τυπικές μορφές γραπτής επικοινωνίας, π.χ. σε επιστημονικά κείμενα» χρησιμοποιούνται μερικές φορές και «μετοχές του μεσοπαθητικού ενεστώτα σε *-όμενος, -όμενη/-όμενη, -όμενο*» και οι μετοχές του «μεσοπαθητικού αορίστου σε *-θείς, θείσα, -θέν*», συνήθως με άρθρο, με σημασιολογική ισοδυναμία με αναφορικές προτάσεις.

Σε ειδικές γλωσσολογικές μελέτες προβάλλεται η άποψη ότι οι μετοχές παθητικού παρακειμένου σε *-μένος* εμφανίζουν ονοματικά χαρακτηριστικά που καθιστούν σε αρκετές περιπτώσεις προβληματική τη διάκριση ανάμεσα σε αυτές και στα επίθετα (ενδεικτικά Lascaratou & Φιλίππακη-Warburton 1984, Ralli 2003, Manolessou 2005, Νικολάου 2016 και 2019), όπως ότι συχνά τοποθετούνται σε θέση πριν από ουσιαστικό, επιδέχονται διαβάθμιση (σχηματισμό συγκριτικού και υπερθετικού βαθμού), συνδέονται με καθαυτό επίθετα μέσω του συμπλεκτικού συνδέσμου *και*, έχουν θέση κατηγορουμένου όταν συναφθούν με συνδετικά ρήματα όπως το *είμαι* ή το *φαίνομαι*, ο φορέας της δράσης (υποκείμενο) είναι ενσωματωμένος στον τύπο, π.χ. *ηλιοκαμένος*, μπορούν να τροποποιηθούν με το *παρα-*, π.χ. *παραμαυρισμένος*, μπορούν να σχηματίσουν επιρρήματα με την προσθήκη της κατάληξης *-α*, π.χ. *θλιμμένα*, και η παραγωγή τους είναι μη κανονική, αφού δε σχηματίζουν όλα τα ρήματα της παθητικής φωνής μετοχές, ενώ υπάρχουν ρήματα της ενεργητικής φωνής που διαθέτουν τύπους σε *-μένος*, π.χ. *μεθάω – μεθυσμένος*. Η επεξεργασία εκτεταμένου υλικού από σώματα κειμένων θα μπορούσε να δώσει πειστικές απαντήσεις σε σχέση με την τάση επιθετοποίησης των μετοχών σε *-μένος*, ανάλογα με τη σημασία του ρήματος και τη χρήση σε συγκεκριμένα συμφραζόμενα. Η φράση «υπό *αυξημένη* παρακολούθηση πυρηνικός σταθμός στη Γαλλία» φαίνεται ότι παραπέμπει στη ρηματική ενέργεια «αυξήθηκε η παρακολούθηση», ενώ η *αυξημένη* πίεση είναι η υψηλή πίεση και όχι αυτή που έχει αυξηθεί.<sup>1</sup> Βλ. και στα αγγλικά παρόμοια δυσδιάκριτα όρια μεταξύ μετοχής και επιθέτου: The *increased* investment will help stabilise the economy (past participle / adjective) <https://www.collinsdictionary.com/dictionary/english/increase>.

### 3 Οι μετοχές στα νεοελληνικά λεξικά και στα σώματα κειμένων

Για να σκιαγραφηθεί το λεξικογραφικό τοπίο σχετικά με τις μετοχές παθητικού παρακειμένου και αορίστου στα νέα ελληνικά επιλέχθηκε η αποδελτίωση στο Λεξικό Ρημάτων (Ιορδανίδου 1992) των ρημάτων από το γράμμα Α και, με παράλληλη αναζήτηση στο Διαδίκτυο, καταγράφηκαν όσα σχηματίζουν μόνο μετοχές παθητικού παρακειμένου (πρώτη στήλη στον Πίνακα 1), όσα σχηματίζουν μετοχές παθητικού παρακειμένου και αορίστου (δεύτερη στήλη) και όσα σχηματίζουν μόνο μετοχές παθητικού αορίστου (τρίτη στήλη).

ΡΗΜΑΤΑ ΜΕ ΜΤΧ. ΠΑΘ. ΠΑΡΑΚΕΙΜΕΝΟΥ ΧΩΡΙΣ ΜΤΧ. ΠΑΘ. ΑΟΡΙΣΤΟΥ [εμφανίσεις μτχ. παθ. αορ. 23/4/20 Google κάτω των 50]	ΡΗΜΑΤΑ ΜΕ ΜΤΧ. ΠΑΘ. ΠΑΡΑΚΕΙΜΕΝΟΥ ΚΑΙ ΜΤΧ. ΠΑΘ. ΑΟΡΙΣΤΟΥ [εμφανίσεις μτχ. παθ. αορ. 23/4/20 Google άνω των 50]	ΡΗΜΑΤΑ ΜΕ ΜΟΝΟ ΜΤΧ. ΠΑΘ. ΑΟΡΙΣΤΟΥ [εμφανίσεις μτχ. παθ. αορ. 23/4/20 Google άνω των 50]
141	58	6 ανατραπείς, -είσα, -έν αναφερθείς, -είσα, -έν αντιπαρατεθείς, -είσα, -έν απαχθείς, -είσα, -έν απελαθείς, -είσα, -έν απορριφθείς, -είσα, -έν

Πίνακας 1: Ρήματα από γράμμα Α – εμφανίσεις μετοχών στο Διαδίκτυο.

Όπως φαίνεται στον Πίνακα 1, περίπου 70% των ρημάτων εμφανίζουν μόνο μετοχές παθητικού παρακειμένου, 28% και μετοχές παθητικού παρακειμένου και μετοχές παθητικού αορίστου και 2% μόνο μετοχές παθητικού αορίστου. Από την επεξεργασία του διαδικτυακού υλικού προκύπτουν τα εξής:

- Στο 70% των ρημάτων με μόνο μετοχές παθητικού παρακειμένου ανήκουν κατεξοχήν ρήματα με υψηλή συχνότητα εμφάνισης σε κείμενα του καθημερινού προφορικού λόγου και της λογοτεχνίας, π.χ. *αγανακτισμένος, αγαπημένος, αγκαλιασμένος, αγριεμένος, αγχωμένος*. Η σπάνια εμφάνιση μετοχής παθητικού αορίστου σε τέτοια ρήματα δηλώνει σκωπτική χρήση, π.χ. Είναι τα τραγούδια των προδομένων και του σκληρού χωρισμού που

<sup>1</sup> Και για τις μετοχές σε *-όμενος* μπορεί να παρατηρείται διάκριση ρηματικής – επιθετικής λειτουργίας, όπως αποτυπώνεται, για παράδειγμα, στις φράσεις «ψηφιακός μηχανισμός *ελεγχόμενος* από υπολογιστή» (ρηματική ενέργεια) και «*ελεγχόμενος* και σταδιακός ο επαναπατρισμός» (ιδιότητα). Παρόμοια και για τη μετοχή παθητικού παρακειμένου του ίδιου ρήματος: «*ελεγχόμενος* από την Κεντρική Ευρωπαϊκή Τράπεζα κατασκευαστής» – «*ελεγχόμενος* βαθμός απόδοσης».



συνήθως οφείλεται σε ύπαρξη τρυφερώς *αγαπηθείσα*.<sup>2</sup> Στα ρήματα αυτά απουσιάζει ο τριτοπρόσωπος σχηματισμός του παθητικού αορίστου σε *-η, -ησαν* που απαντάται ορισμένες φορές σε ρήματα λόγιας προέλευσης, π.χ. (ανατρέπω) *ανετράπη, ανετράπησαν*.

- Η χρήση της μετοχής παθητικού αορίστου αντί της μετοχής παθητικού παρακειμένου απαντάται σχεδόν αποκλειστικά σε διοικητικά, νομικά, ακαδημαϊκά και δημοσιογραφικά ενημερωτικά κείμενα, συνήθως σε ρόλο ουσιαστικού ή επιθέτου, π.χ. *αγορασθέντα* αντί για *αγορασμένα* (προϊόντα), *αδικηθέντες* (παραγωγούς) αντί για *αδικημένους*, *αθωωθέντες* από το δικαστήριο αντί για *αθωωμένους*, *ανακηρυχθέντες* (υποψήφιους) αντί για *ανακηρυγμένους*, *ανακτηθέντα* (ακίνητα) αντί για *ανακτημένα*, *ανατεθέντα* (καθήκοντα) αντί για *ανατεθειμένα*. Σε αρκετές περιπτώσεις εμφανίζεται αποκλειστικά ή σε υψηλή συχνότητα σε συγκεκριμένες ονοματικές φράσεις, π.χ. *αναληφθέντες δάσκαλοι, αναρτηθέντες πλειστηριασμοί, αποδοθέντες φόροι*.
- Η ανοίκεια στον καθημερινό λόγο πολύπλοκη μορφολογία οδηγεί αρκετές φορές σε αποκλίνουσα χρήση, κυρίως στον σχηματισμό του θηλυκού: *αναβληθέντων αναμετρήσεων, αναλυθέντων μεθόδων, ανεγερθέντες κατοικίες, απαλλοτριωθέντων εκτάσεων, αποκτηθέντες συνήθειες, αποσταλέντων εκθέσεων*.
- Σε κάποια ρήματα η μετοχή παθητικού παρακειμένου έχει ειδική σημασία επιθέτου, π.χ. *απομακρυσμένοι* (= μακρινοί) *προορισμοί*, ενώ η μετοχή παθητικού αορίστου χρησιμοποιείται με ρηματική λειτουργία, π.χ. *οι απομακρυνθέντες από την υπηρεσία υπάλληλοι*.
- Σε ελάχιστα παραδείγματα χρησιμοποιούνται και οι δύο μετοχές με διάκριση συντελεσμένου – συνοπτικού τρόπου ενέργειας, π.χ. οι «νέοι *ασφαλισμένοι*» στρατιωτικοί (*ασφαλισθέντες* για πρώτη φορά από 1-1-1993 και μετά).<sup>3</sup>
- Τα 6 παραδείγματα αποκλειστικής εμφάνισης μετοχών παθητικού αορίστου αφορούν κατεξοχήν χρήση ως επιθέτων ή ουσιαστικών, π.χ. *οι ανατραπέντες δικτάτορες, οι απαχθέντες ναυτικοί, οι απελαθέντες μετανάστες*.

Η παρατήρηση της κειμενικής κατανομής των μετοχών παθητικού αορίστου φαίνεται να επιβεβαιώνει τα ευρήματα της διατριβής μου (Iordanidou 1985), όπου οι μετοχές παθητικού αορίστου σε *-είς, -είσα, -έν* (καταγεγραμμένες από κοινού με τις κλιτές μετοχές ενεργητικού ενεστώτα σε *-ων*) ήταν απύσυχες από κείμενα (απομαγνητοφωνημένου) προφορικού λόγου και λογοτεχνίας:

	ΑΡΘΡΑ ΚΑΙ ΕΙΔΗΣΕΙΣ ΣΕ ΕΦΗΜΕΡΙΔΕΣ	ΕΠΙΣΤΗΜΟΝΙΚΑ ΚΕΙΜΕΝΑ	ΛΟΓΟΤΕΧΝΙΑ	ΓΡΑΜΜΑΤΑ ΑΝΑΓΝΩΣΤΩΝ	ΠΡΟΦΟΡΙΚΟΣ ΛΟΓΟΣ
-ontas	20%	16%	22%	17%	6%
-menos	44%	24%	75%	40%	84%
-omenos	26%	45%	3%	30%	10%
-on / -eis	10%	15%	---	13%	---

Πίνακας 2: Κατανομή μετοχών στο σώμα κειμένων Iordanidou 1985.

Στα συμπεράσματα της διατριβής αναφερόταν ότι η παρουσία των αρχαίων μετοχών σε *-ων* και *-είς* στις άλλες κειμενικές κατηγορίες μειώνεται σημαντικά αν ληφθεί υπόψη ότι περιλαμβάνονται και επιθετοποιημένες χρήσεις. Είναι σημαντικό εδώ να υπογραμμιστεί η σπουδαιότητα των κειμενικών ειδών στην εξέλιξη της νεοελληνικής διγλωσσίας την οποία αναδεικνύει η εργασία των Fragaki & Goutsos 2018 με δεδομένα στην «Διαχρονικό σώμα ελληνικών κειμένων του 20ού αιώνα». Συγκεκριμένα, διαπιστώνεται η απουσία των τύπων της καθαρεύουσας *διά* και *εις* (ως αντίστοιχων των τύπων των προθέσεων της δημοτικής *για* και *σε*) σε λογοτεχνικά κείμενα και σε διαλόγους κινηματογραφικών ταινιών, έναντι της παρουσίας τους σε ακαδημαϊκά κείμενα, δημόσιες ομιλίες και ειδήσεις.

Στο επόμενο στάδιο της έρευνας, για να μελετηθεί η σύγχρονη ελληνική λεξικογραφική πρακτική, καταγράφηκαν ενδεικτικά από τον κατάλογο των ρημάτων του γράμματος Α καταχωρίσεις στα τρία μεγαλύτερα σύγχρονα λεξικά: *Λεξικό της κοινής νεοελληνικής* (εφεξής *ΛΚΝ*), 1998, Μπαμπινιώτη *Λεξικό της νέας ελληνικής γλώσσας* (εφεξής *ΛΝΕΓ*), πρώτη έκδ. 1998, αναθεωρημένη 2003, και *Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας* (εφεξής *ΧΛΝΓ*), Ακαδημία Αθηνών, 2014.

<sup>2</sup> [http://greek-guitar.weebly.com/uploads/1/0/9/2/10927010/rembetiko\\_to\\_rembetiko\\_gennisi\\_kai\\_sanatos\\_tou.pdf](http://greek-guitar.weebly.com/uploads/1/0/9/2/10927010/rembetiko_to_rembetiko_gennisi_kai_sanatos_tou.pdf)

<sup>3</sup> <https://www.dealnews.gr/roi/item/9428-Διευκρινίσεις-για-τις-αλλαγές-στη-συνταξιοδότηση-στρατιωτικών#.XqSjShRRXIU>



ΛΚΝ (Λεξικό Κοινής ΝΕ) Ε12γ [επιθ.]	ΛΝΕΓ (Λεξικό Μπαμπινιώτη)	ΧΛΝΓ (Λεξικό Ακαδημίας)	Google [23/4/2020] εμφανίσεις άνω των 50
	αναβλήθηκα, αναβεβλημένος	αναβλήθηκε (λόγ. ανεβλήθη, μτχ. αναβληθείς)	αναβληθείς και αναβλημένος
	αναγγέλθηκα, αναγγελμένος	αναγγέλθηκε, μτχ. αναγγελθείς	αναγγελθείς και αναγγελμένος
	ανακλήθηκα (λόγ. ανεκλήθην, μτχ. ανακληθείς), ανακλημένος	ανακλήθηκε (λόγ. ανεκλήθη, μτχ. ανακληθείς)	ανακληθείς και ανακλημένος
ανακοινωθείς, -είσα, -έν (και ανακοινωθέν, το)	ανακοινώθηκα, -μένος και ανακοινωθέν, το	ανακοινώθηκε (λόγ. ανεκοινώθη, μτχ. ανακοινωθείς), -μενος, και ανακοινωθέν, το	ανακοινωθείς και ανακοινωμένος
	αναλήφθηκα (λόγ. ανελήφθην μτχ. αναληφθείς) ανειλημμένος	αναλήφθηκε (λόγ. ανελήφθη, μτχ. αναληφθείς), ανειλημμένος	αναληφθείς και ανειλημμένος
	αναλύθηκα, -μένος (λόγ. αναλελυμένος)	αναλύθηκε, αναλυμένος	αναλυθείς και αναλυμένος

Πίνακας 3: Γράμμα Α, μετοχές παθητικού αορίστου στα λεξικά.

Όπως φαίνεται στον Πίνακα 3, στο ΛΚΝ η μοναδική μετοχή παθητικού αορίστου που καταχωρίζεται συνιστά ξεχωριστό λήμμα με τον χαρακτηρισμό «επίθετο» (και παραπομπή σε σχετικό κλιτικό πίνακα Ε12γ): *ανακοινωθείς, -είσα, -έν*, χωρίς να περιλαμβάνεται στο λήμμα του ρήματος «ανακοινώνω», ενώ τα ΛΝΕΓ και ΧΛΝΓ περιλαμβάνουν τις μετοχές στις γραμματικές πληροφορίες στην αρχή του λήμματος. Ο χρήστης των ΛΝΕΓ και ΧΛΝΓ καλείται να αποκωδικοποιήσει πολύπλοκες και αντιφατικές πληροφορίες:

- Η μετοχή σε *-είς* άλλοτε αναφέρεται και άλλοτε όχι, χωρίς διαφανές κριτήριο. Για παράδειγμα, απουσιάζει και από τα δύο λεξικά ο τύπος *αναλυθείς*, παρά το γεγονός ότι καταγράφονται αρκετές εμφανίσεις στο Διαδίκτυο, ενώ από το ΛΝΕΓ απουσιάζουν οι τύποι *αναβληθείς, αναγγελθείς, ανακοινωθείς*. Όταν αναφέρεται η μετοχή σε *-είς*, χαρακτηρίζεται «λόγ.» στο ΛΝΕΓ (και συνδέεται με παθητικό αόριστο σε *-ην*) και «λόγ.» στο ΧΛΝΓ συνδεόμενη με την τριτοπρόσωπη ποικιλία του παθητικού αορίστου *-η, -ησαν*. Να υπογραμμιστεί εδώ ότι η αναφορά της κλίσης σε *-ην* στο ΛΝΕΓ και της τριτοπρόσωπης ποικιλίας *-η, -ησαν* στο ΧΛΝΓ δε φαίνεται να προκύπτει από παρατήρηση της γλωσσικής πρακτικής: και στα έξι ρήματα του Πίνακα 3 επικρατεί η κλίση σε *-ηκα*. Γενικότερα, για ορισμένα ρήματα με αόριστο σε *-η, -ησαν* είναι συνήθης η εμφάνιση ουσιαστικοποιημένης ή/και επιθετοποιημένης μετοχής σε *-είς* (π.χ. *συνελήφθη – συλληφθείς*<sup>4</sup>), αλλά σε άλλα απουσιάζει εντελώς (π.χ. *εξερράγη, επενέβη, κατέστη, συνεπλάκη*). Η λόγια μορφολογία δεν είναι η μοναδική προϋπόθεση – η χρήση σε συγκεκριμένα είδη κειμένων όπου κυριαρχούσε η καθαρεύουσα φαίνεται ότι είναι ο καθοριστικός παράγοντας για την επιβίωση των μετοχών σε *-είς* ως ουσιαστικών ή/και επιθέτων.
- Στο ΛΝΕΓ η μετοχή *αναβεβλημένος* παρατίθεται χωρίς χαρακτηρισμό, ενώ η αντίστοιχη του ρήματος *αναλύω* (*αναλελυμένος*) χαρακτηρίζεται «λόγ.». Στο ΧΛΝΓ αναφέρεται μόνο η μετοχή *αναβληθείς*, ως συνδεδεμένη με τη «λόγ.» τριτοπρόσωπη ποικιλία του παθητικού αορίστου. Στο Διαδίκτυο απουσιάζει ο *αναβεβλημένος* αλλά εμφανίζεται σποραδικά ο *αναβλημένος*.

#### 4 Συμπεράσματα – Συζήτηση

Με δεδομένη την απουσία των μετοχών παθητικού αορίστου από το γραμματικό σύστημα της σύγχρονης ελληνικής, η επιλεκτική λεξικογραφική αναφορά τους ως μοναδικών τύπων ή παράλληλα με τις μετοχές παθητικού παρακειμένου

<sup>4</sup> Αξίζει να αναφερθεί ότι ο τύπος *συλληφθέντες* απαριθμεί 10.350 εμφανίσεις στο Sketch Engine [Greek Web 2014, πρόσβαση 23/4/20] και ακολουθούν *προαναφερθέντες* 3.146, *πληγέντες* 2.884 και *ερωτηθέντες* 2.355. Στα μικρότερης εμβέλειας ΣΕΚ (Σώμα Ελληνικών Κειμένων) για τους *συλληφθέντες* καταγράφονται 27 εμφανίσεις και στον ΕΘΕΓ (Εθνικός Θησαυρός Ελληνικής Γλώσσας) 50 [πρόσβαση 23/4/20]. Το γεγονός ότι αυτοί οι τύποι δε διαθέτουν εναλλακτικό σχηματισμό σε *-μένος* συντελεί στην απολιθωματική τους επιβίωση.



φαίνεται ότι ανάγει τη νομιμοποίησή τους στο κριτήριο του υψηλού κύρους της λόγιας μορφολογίας: εφόσον το προβαλλόμενο γλωσσικό πρότυπο χρησιμεύει και ως μηχανισμός διακρίσεων, θα πρέπει να παρουσιάζει και υψηλό βαθμό δυσκολίας και πολυπλοκότητας. Αξίζει να τονιστεί ότι δεν πρόκειται για φαινόμενο γλωσσικής ποικιλίας, όπως είναι, π.χ., η τριτοπρόσωπη παραλλαγή *-είτο, -ούντο* (*εθεωρείτο – εθεωρούντο* έναντι *θεωρούνταν*) στον παρατατικό των ρημάτων σε *-ούμαι* αλλά για αδιαφανή γραμματική δομή με ιδιαίτερα βεβαρημένη μορφολογία. Φαίνεται ότι πρόκειται για ελληνική γλωσσική ιδιαιτερότητα: στη σχολική Γραμματική Ε' και Στ' Δημοτικού και στο ΛΚΝ, που αποτελούν τους βασικούς φορείς της τυποποίησης (standardisation), απουσιάζουν μορφολογικοί τύποι και δομές υψηλής επισημότητας και πολυπλοκότητας, ενώ στο διεθνές γλωσσικό τοπίο είναι σύνηθες το αντίθετο, να προβάλλονται δηλαδή ως πρότυπα της «πρότυπης» (standard) οι τύποι και οι δομές που χαρακτηρίζουν επίσημες περιστάσεις επικοινωνίας και επίσημο ύφος λόγου (βλ. ενδεικτικά Poplack 2015, Moschonas 2019).

Η έρευνα σε σώμα κειμένων γραμμάτων αναγνωστών εφημερίδων (Ιορδανίδου 1999), με παραγωγή λόγου σε συνθήκες υψηλής επισημότητας, έδειξε ότι οι ομιλητές οριοθετούν συστηματικά και με παρατηρήσιμη κανονικότητα τις λόγιες επιλογές και δεν τις ανάγουν σε συνολικές μορφολογικές δυνατότητες. Για παράδειγμα, η μορφολογία του παρατατικού των ρημάτων σε *-ομαι* δε φαίνεται να επηρεάζεται από την εμφάνιση των λόγων καταλήξεων *-είτο, -ούντο* των ρημάτων σε *-ούμαι*, τα λόγια συμφωνικά συμπλέγματα εμφανίζονται σε συγκεκριμένα ρήματα και ο παθητικός αορίστος σε *-ηκα* δεν υποκαθίσταται από τη λόγια κλίση (τρίτο πρόσωπο *-η, -ησαν*) παρά μόνο σε πολύ περιορισμένες περιπτώσεις. Όσον αφορά τις αρχαίες μετοχές ενεργητικού ενεστώτα και παθητικού αορίστου, είχε διαπιστωθεί ότι στα γράμματα αναγνωστών εμφανίζονταν σπάνια ως μετοχές, εντελώς περιθωριακά σε σύγκριση με τις μετοχές σε *-οντας, -όμενος* και *-μένος* της κοινής νεοελληνικής.

Η πρόταση της παρούσας εργασίας, βασισμένη σε σωματοκειμενικά δεδομένα, είναι ότι θα έπρεπε να ακολουθηθεί το πρότυπο του ΛΚΝ, με καταγραφή της χρήσης των αρχαίων μετοχών παθητικού αορίστου ως ουσιαστικών και επιθέτων, αλλά και με εκτενείς πληροφορίες σχετικά με τα κειμενικά είδη και τα συμφραζόμενα όπου εντάσσονται, καθώς και σχολιασμό της δυνατότητας εναλλακτικής απόδοσης με μετοχές παθητικού παρακειμένου ή αναφορικές προτάσεις. Παρατίθεται παράδειγμα εναλλακτικών επιλογών για τον τύπο *απολυθέντες*, ο οποίος δεν αναφέρεται στο ΛΝΕΓ και στο ΧΛΝΓ αλλά εμφανίζει υψηλά διαδικτυακά ποσοστά (150 Google [23/4/20] και 334 Sketch Engine [Greek Web 2014]):

ερώτηση για τη μη καταβολή δεδουλευμένων στους <i>απολυθέντες</i>	καταβολή δεδουλευμένων στους <i>απολυμένους</i>	Έχει καταβληθεί το σύνολο των δεδουλευμένων σε όσους εργαζόμενους <i>απολύθηκαν</i> .
Σε μια συμβολική κίνηση άναψαν το φακό από το κινητό τηλέφωνο για να δηλώσουν συμπάρασταση στους <i>απολυθέντες</i> .	Συμπάρασταση στους <i>απολυμένους</i> συμβασιούχους του δήμου.	Δηλώνουμε την αμέριστη συμπάραστασή μας στους υπαλλήλους που <i>απολύθηκαν</i> .
Προσέφερε το Ευρωπαϊκό Ταμείο Προσαρμογής στην Παγκοσμιοποίηση ευρωπαϊκή προστιθέμενη αξία στην επανένταξη <i>απολυθέντων</i> εργαζομένων;	επανένταξη των <i>απολυμένων</i> εργαζομένων, κατά προτίμηση στο πλαίσιο κοινωνικού προγράμματος	την επανένταξη στην απασχόληση των εργαζομένων που <i>απολύθηκαν</i> λόγω της παγκοσμιοποίησης
Αλληλεγγύη στους <i>απολυθέντες</i> από την ΑΔΕΔΥ.	«Στερεά Υπεροχής!» <i>αλληλέγγυα</i> στους <i>απολυμένους</i> στο Μαντούδι.	Έδειξαν αλληλεγγύη σε όσους <i>απολύθηκαν</i> .

Πίνακας 4: Εναλλακτικές μετοχές / αναφορικές προτάσεις (Διαδίκτυο).

Στην εποχή της ηλεκτρονικής λεξικογραφίας, όπου δεν υπάρχει ο περιορισμός χώρου των έντυπων λεξικών, παρόμοιες πληροφορίες μπορούν να είναι διαθέσιμες στον χρήστη, όπως και πληθώρα άλλων: πλήρης κλίση, αναλυτικά ερμηνεύματα συνοδευόμενα από αυθεντικά παραδείγματα, αναλυτική παράθεση λεξικών συνάψεων, με παράλληλη απλοποίηση της μεταγλώσσας, κατάργηση συμβόλων και συντομογραφιών και συνδέσεις με ηλεκτρονικά σώματα κειμένων και πολυμεσικό υλικό (Oppenrecht & Schutz 2003). Να τονιστεί, επίσης, ότι με την ηλεκτρονική λεξικογραφία διασφαλίζεται η συνεχής ενημέρωση και επικαιροποίηση, κάτι το οποίο είναι από πολύ δύσκολο έως αδύνατο στα έντυπα λεξικά.

Σε παρόμοιο μήκος κύματος με την παρούσα εργασία, η έρευνα της Κατσούδα 2009 σχετικά με τον σχηματισμό του πληθυντικού αριθμού των μετοχών *ηγουμένη, εφαιπόμενη, προϊσταμένη, συνισταμένη, υφισταμένη* καταλήγει ότι τα λεξικά πρέπει να αφιερώνουν διαφορετικό λήμμα για τις επιθετοποιημένες μετοχές και διαφορετικό για τις ουσιαστικοποιημένες (όπως κάνει το ΛΚΝ) και, εφόσον δίνονται γραμματικές πληροφορίες, πρέπει να παραπέμπουν σε κλιτικό παράδειγμα, ώστε να γίνονται αντιληπτές οι μορφολογικές ιδιαιτερότητες, ενώ μπορούν να παρουσιάζουν τη μορφολογική ποικιλία του πληθυντικού και μέσα από τα παραδείγματα του λήμματος.



## 5 Βιβλιογραφικές αναφορές

- Fragaki, G. & Goutsos, D. (2018). The importance of genre in the Greek diglossia of the 20th century – A diachronic corpus study of recent language change. In R. J. Whitt (ed.), *Diachronic Corpora, Genre, and Language Change*. John Benjamins Publishing Company.
- Γραμματική Ε' και Στ' Δημοτικού (2009) βλ. Φιλιππάκη-Warburton et al.
- Holton, D., Mackridge, P. & Φιλιππάκη-Warburton, E. (1999). *Γραμματική της Ελληνικής Γλώσσας*. Αθήνα: Εκδόσεις Πατάκη.
- Iordanidou, A. (1985). *La diglossie en Grèce: Étude d'un cas précis, le participe*. Thèse de Doctorat Troisième Cycle, Université de Paris VII. <http://thesis.ekt.gr/thesisBookReader/id/11501#page/1/mode/2up> [23/4/2020].
- Ιορδανίδου, Α. (1992). *Τα ρήματα της νέας ελληνικής*. Αθήνα: Εκδόσεις Πατάκη.
- Ιορδανίδου, Α. (1999). Ζητήματα τυποποίησης (standardisation) της σύγχρονης νεοελληνικής, Διεθνές Συνέδριο «Ισχυρές» - «Ασθενείς» γλώσσες στην Ευρωπαϊκή Ένωση, Πρακτικά, Θεσσαλονίκη, Κέντρο Ελληνικής Γλώσσας, σ. 835-854.
- Κατσούδα, Γ. (2009). Ο σχηματισμός του πληθυντικού εφαπτομένη, ηγουμένη, προϊσταμένη, συνισταμένη, υφισταμένη: συμπεράσματα και εφαρμογές. Στο Α. Tsangalidis (ed.) *Selected Papers from the 18<sup>th</sup> International Symposium on Theoretical and Applied Linguistics* (Thessaloniki, 4-6 May 2007), 457-465. <https://ejournals.lib.auth.gr/thal/article/view/5464/0>. [23/4/2020].
- Κλαίρης, Χ. & Μπαμπινιώτης, Γ. (2005). *Γραμματική της Νέας Ελληνικής. Δομολειτουργική-Επικοινωνιακή*. Αθήνα: Ελληνικά Γράμματα.
- Laskaratu, G. & Philippaki I. (1984). Lexical versus transformational passives in Modern Greek. *Γλωσσολογία*, 2-3, pp. 99-109. <http://glossologia.phil.uoa.gr/sites/default/files/Ch.%20Lascaratu%20-%20I.%20Philippaki%20-%20Warburton%20%281983-4%29.PDF>. [23/4/2020].
- Λεξικό της κοινής νεοελληνικής (ΛΚΝ) (1998). Θεσσαλονίκη: Α.Π.Θ., Ινστιτούτο Ν.Ε. Σπουδών. [http://www.greek-language.gr/greekLang/modern\\_greek/tools/lexica/triantafyllides/index.html](http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/index.html). [23/4/2020].
- Manolessou, I. (2005). From participles to gerunds. In M. Stavrou & A. Terzi (eds.) *Advances in Greek Generative Syntax*. Amsterdam/Philadelphia: John Benjamins, pp. 241-283. <http://docslide.us/documents/from-participles-to-gerunds.html>. [23/4/2020].
- Moschonas, S. A. (2019). From language standards to a Standard Language: The case of Modern Greek. *Diachronia*, 10, pp. 1-44.
- Μπαμπινιώτης, Γ. (1998, αναθ. έκδοση 2003). *Λεξικό της νέας ελληνικής γλώσσας (ΛΝΕΓ)*. Αθήνα: Κέντρο Λεξικολογίας.
- Νικολάου, Γ. (2016). Τα νεολογικά επίθετα της κοινής νεοελληνικής. Διδακτορική διατριβή. ΑΠΘ, Τμήμα Φιλολογίας. <http://hdl.handle.net/10442/hedi/39203>. [23/4/2020].
- Νικολάου, Γ. (2019). [+ λόγιες] μετοχές της νεοελληνικής γλώσσας. Στο Α. Φλιάτουρας, Α. Αναστασιάδη-Συμεωνίδη (επιμ.) *Το λόγιο επίπεδο στη σύγχρονη νέα ελληνική: Θεωρία, ιστορία, εφαρμογή*. Αθήνα: Εκδόσεις Πατάκη, σ. 187-199.
- Oppentocht, L. & Schutz, R. (2003). Developments in electronic dictionary design. In P. Van Sterkenburg (ed.), *A Practical Guide to Lexicography*. Amsterdam: John Benjamins, pp. 215-227.
- Poplack, S. (2015). Norme prescriptive, norme communautaire et variation diaphasique. In K. Kragh & J. Lindschouw (eds.) *Les variations diasystematiques et leurs interdépendances dans les langues romanes : Actes du Colloque DIA II 2012*. Strasbourg, Société de linguistique romane/ÉliPhi, p.p. 293-319.
- Ralli, A. (2003). Morphology in Greek Linguistics: A State-of-the Art. In *Journal of Greek Linguistics* 4, pp. 77-130.
- Τριανταφυλλίδης, Μ. (1978) [1941]. *Νεοελληνική Γραμματική της Δημοτικής*. Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών (Ίδρυμα Μανόλη Τριανταφυλλίδη).
- Φιλιππάκη-Warburton, E., Γεωργιαφέντης, Μ., Κοτζόγλου, Γ. & Μ. Λουκά (2009). *Γραμματική Ε' και Στ' Δημοτικού*. Ινστιτούτο Τεχνολογίας Υπολογιστών και Εκδόσεων «ΔΙΟΦΑΝΤΟΣ». <http://digitalschool.minedu.gov.gr> [23/4/2020].
- Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας (ΧΛΝΓ). (2014). (Σύντ. & Επιμ.: Χρ. Χαραλαμπίδης). Αθήνα: Ακαδημία Αθηνών & Εθνικό Τυπογραφείο.



# Semantic Relations in the Thesaurus of English Idioms: A Corpus-based Study

Giztova G.<sup>1</sup>, Ismagilova L.<sup>2</sup>

<sup>1</sup> Kazan Federal University, Russia

<sup>2</sup> Kazan Federal University, Russia

## Abstract

This paper deals with the principles of constructing an Ideographic Dictionary of English Idioms (Thesaurus) based on corpus data. Idioms in the dictionary are arranged by their figurative meaning rather than alphabetically. The need for a new type of dictionary is motivated by the fact that at present there is no corpus-based dictionary of English idioms built on a thesaural principle. Ideographic description of idioms enables a reader to find the largest possible number of idiomatic word combinations of the language that express a given concept. The basic entry of the Thesaurus is called a taxon, consisting of a conceptual descriptor used as a label of a taxon, and a group of idioms expressing the respective taxon. English Web text corpus 2013 (enTenTen13) is used as an empirical basis of the study. The analysis of corpus data presents a range of syntactic patterns, idiom variation, synonymous and polysemous idioms which cannot be retrieved from the existing idiomatic and monolingual dictionaries of the English language, since they fail to register all meanings of an idiom. Today, as lexicography is experiencing “the corpus revolution” (Hanks 2012), this is a question of key importance. The use of corpora provides additional possibilities for compiling the idiom list and structuring entries.

**Keywords:** thesaurus, idioms, corpus, variation, synonymy

## 1 Theoretical Concept

The research is based on the main principles of cognitive linguistics and, primarily, on the system organization of structuring of semantic fields. It is assumed that ideographic classifications of different languages *basically* coincide and conceptual sphere covering phraseology of different languages in principle is the same, i.e. extralinguistic, that can be interpreted as a conceptual universal (Dobrovol'skij 1992: 280). However, every language has its own unique semantic structure. Each semantic field segments objective reality in a way that is specific only to a given language. Moreover, certain linguistic changes within the language belong to the sphere of initial concepts, which have specific linguistic differences in other languages. The thesaurus is based on an inductive method, that is – from idioms to semantic fields and not from an abstract logical outline to idioms.

Following the principles of the Conventional Figurative Language Theory (CELT) developed by Dobrovol'skij & Piirainen (2005), we use the term *idiom* in the European tradition of phraseology research and rely on their definition of an idiom as:

... phrasemes with a high degree of idiomaticity and stability. In other words, idioms must be fixed in their lexical structure (however, this does not exclude a certain limited variation), and they must be, at the same time, semantically reinterpreted units (i.e. they do not point to the target concept directly but via a source concept) and/or semantically opaque (Dobrovol'skij & Piirainen 2005:40).

Our work on the English Thesaurus was inspired by publishing of the *Thesaurus of Present-Day Russian Idioms* by Baranov and Dobrovol'skij (2007). It can be called a “lexicographic giant” and had been the first and the only profound, comprehensive and up to date ideographic dictionary of Russian idioms in international lexicography since then. Such types of dictionaries have not been developed in other languages as well. A thorough outline of existing English thematic dictionaries (not ideographic ones, since they do not exist) of idioms is given in (Gizatova 2016). Significance of ideographic dictionaries of idioms lies in their ability to reflect the “naïve” pre-scientific picture of the world, in the study of the history of human knowledge about the world surrounding us, in comparative study of “natural” world picture of different nations. Despite the importance of these issues, few studies have focused on these aspects, so practical lack of research and lexicographic accomplishments in this field motivated our interest to start our work on compiling a corpus-based dictionary of English idioms built on thesaural principle and exclusively authentic examples.

## 2 Discussion

Lexicographers encounter certain problems in connection with treatment of idioms in dictionaries. In this article the authors approach the issue of idioms in the context of organizing semantic fields of the Thesaurus. High level of variation of idiom components is a typical feature of phraseological system of languages. Very often it is difficult to make a decision whether it is a variant or a synonym of an idiom. The most important conceptual criterion in this case is the identity of inner forms of the idioms under consideration. If the inner form of an idiom is not changed in the case of its component modification, the decision is taken in favour of variants (*shake/quiver/quake in one's boots/shoes*). If the



image components underlying the idioms are different, it is natural to speak of a synonymy (*spill the beans/ let the cat out of the bag*). This leads to the reflection of semantic differences of idioms in structuring of semantic fields of the Thesaurus.

Traditionally, idiomatic variants are units in which partial difference in their constituents does not alter their meaning. In contrast to idiomatic synonyms which are interchangeable only in definite contexts, substitution between idiom variants is possible in any condition of usage. Let us have a closer look at idiom variants:

(1) *to have/bear a grudge* “to have an old resentment for someone; to be continually angry at someone”.

The idiom *to have a grudge* is traditionally considered to be more formal than *to bear a grudge*, the latter being more frequently used in poetry and fiction. But analysis of the corpus material does not prove this notion. It is sufficient to look at their contextual use in the following examples and it is evident that both idioms are used in spoken syntactic constructions, cf.:

(2) But possibly the worst job I ever had was possibly to the worst boss I ever had. Marjorie. I still remember you. Marjorie was my middle-aged upline when I was 17 years old Office Junior. And she *had a grudge* against all humanity, particularly against friendly teenage girls, it turned out (enTenTen13).

(3) There were the key words which only the adults could understand. I grew up and came to understand, and I *bore a grudge* against all men... All of them were treacherous (enTenTen13).

These two idioms have some specific characteristics: both of them are usually used in the construction with preposition *against*, at the same time, the results of corpus analysis show that *to have a grudge* is used almost twice as much with the preposition *against* as the idiom *to bear a grudge* (72% to 38% of occurrences respectively). The idioms also differ in the aspect of their tendency to be used in negative sentences. Thus, *to bear a grudge* can take negative form in 17% of its use, whereas *to have a grudge* practically does not take the negative form (only in 0.7%). These characteristics of the two idioms are of structural nature and refer to their inner organization. Both idioms have identical meanings and functions. That is why we consider them to be idiomatic variants. At the same time, we think it essential to indicate these insignificant structural differences in idioms in this paper. It is important because the specifics of inner organization of an idiom can lead to generation of semantic and other connotations and consequently play a decisive role in transformation of variants into synonyms.

As for arranging idiomatic variants in the Thesaurus taxon, it is important to define which of the variants is the base-form and which is the idiom-variant. According to our data, retrieved from the corpus, the frequency of the idiom *to have a grudge* is twice as high as that of *to bear a grudge* (1456 to 734 occurrences respectively). So, the idiom *to have a grudge* which is more frequently used in the language is the base-form and the less frequent one *to bear a grudge* is the idiom-variant, which follows the base-form after slash, cf.: *to have/bear a grudge*.

As has been mentioned above, one of the challenges in the dictionary-making process is distinguishing between variants and synonyms of idioms. In some dictionaries a range of idiomatic synonyms is traditionally presented as idiom variants. We will show (relying on corpus data) that some of these idioms have different shades of meaning or stylistic colouring and due to that they are regarded as synonymous idioms. Due to that they can fall under different taxa of the Thesaurus. Cf. two synonyms: *to put an end to* and *to put a stop to*. They have a common first component *to put*, but their second components *an end* and *a stop* are synonymous. These idioms have two meanings: 1) to hold back, to restrain; to put something on hold and 2) to destroy, to do away with. In both cases the idiom *to put an end to* has a shade of a more resolute interference, presupposing a significant force and pressure.

In their first meaning “to hold back” these two idioms do not differ in their combinatorial properties, cf.:

(4) We should stop being hypocrites – preaching in public and indulging in private... We now need movements that will *put an end to* corruption and mismanagement which is draining all our efforts and progress (enTenTen13).

(5) A group of women are sitting and talking under the shade of some large mango trees in Juba, South Sudan – a common sight. But today, instead of talking about the weather, the conversation is taking a different tack. One woman asks about the best way *to put a stop to* corruption in local government. She says she wishes their discussion could be heard by all government officials (enTenTen13).

In their second meaning “to destroy; to do away with” both idioms are used with nouns denoting processes, inanimate objects, abstract nouns, e.g.:

(6) In your insane wars you destroy millions of lives and think nothing of it. I am going *to put a stop to* your wholesale destruction of human beings. I want laughter, not slaughter (enTenTen13).

(7) “Mankind must put an end to war before war *puts an end to* mankind” – John F. Kennedy (enTenTen13).

As noted above, the idiom *to put an end to* in comparison with the idiom *to put a stop to* denotes terminating something



involving more resolute pressure and interference. For this reason, it expresses not only an uncompromising determined will of a person to do away with something, but his desire to subdue herewith the will of another person. It is easy to see from corpus examples above that both idioms are used with abstract nouns. But due to corpus data we were able to reveal a significant difference in combinatorial properties of these two idioms. The key issue is that the idiom *to put an end to* in the meaning “to do away with” can be used with the names referring to people due to specific characteristics of its semantic functions described above, cf.:

(8) The right-wing opposition once again erred - Confident in the idea that the world crisis and the fall in oil prices would finally bleed dry the revolution and *put an end to* Chavez (enTenTen13).

Sometimes the ability of an idiom to refer to people is transformed into a new meaning: “to kill”, cf.:

(9) The jealous Queen, vowing to rid herself of “Miss goody-goody Snow White” once and for all, sends the girl off into the forest with the Huntsman, bidding him to “*put an end to her*”. Despite his fear of the Queen’s powers, the Huntsman cannot bring himself to obey her command, and takes the Princess to the home of the seven Dwarfs, who vow to protect her (enTenTen13).

In arranging semantic fields of the English Thesaurus, we follow the basic principle underlying the organization of Thesaurus of Present-day Russian Idioms (Baranov & Dobrovol’skij 2007). According to Dobrovol’skij: “As far as idioms are concerned, it seems appropriate to put together not only synonyms proper, but all idioms belonging to the same conceptual domain without taking into account even their word class identity” (1994: 266). Following this concept, for instance, within the domain *Insanity, Silliness* such idioms as *go nuts* (verb), *a mare’s nest* (noun), *a few cards shy of a full deck* (adjective), *take leave of one’s senses* (verb), *not all there* (adverb) can be organized in one taxon. “The reason for this decision is the vagueness of many idioms, their ability to change the formal parameters depending on the context, and the resulting ineffectiveness of postulating artificially precise distinctions” (Dobrovol’skij 1994: 266).

See below an example-taxon *Punishment* with its conceptual variations:

*Punishment*: Reproach, Blame, Reprimand, Accusation, Criticism

*knock s.o.’s block off; give s.o. hell; chew/bawl s.o. out; haul s.o. over the coals; tear s.o.’s head off; finish s.o. off; put and end to s.o.; skin s.o. alive; raise one’s voice; lift one’s voice; drag s.o. /s.o.’s name through the mud/ the dirt/ the mire/ the muck; blacken/ sully s.o.’s name; give a telling-off to s.o.; give s.o. a sound thrashing; beat the daylight out of s.o.; plow s.o. under; teach s.o. a good lesson etc.*

284 contexts having direct connection to the taxon *Punishment* have been found in the corpus and we have displayed here only some of them for reasons of space. Let us consider two phraseological units: *lift one’s voice* and *raise one’s voice*, often being confused as variants of the same idiom. Their direct meaning is ‘to speak louder’ (physically); the usage of idioms in authentic texts is presented in the following examples.

(10) *lift one’s voice*

Turning to face the others, the Weyrleader *lifts his voice* to be heard by all, not that he has to speak loudly. ‘No sense in us wasting energy either’ (enTenTen13).

(11) *raise one’s voice*

Turning, he *raises his voice* so the others can hear too (enTenTen13).

Their secondary figurative meaning according to the New English-Russian Dictionary is ‘to raise a voice of protest against somebody/something’ (Apresjan 1993:352; 1994:13). But the search in the corpus allowed us to reveal many important semantic differences between these two idioms. First of all, they differ in their semantic functions: if the idiom (10) has a sense of reason, the other idiom (11) is more emotional and often lacks the sense of reason. Consider examples:

(12) I watched some of the commentary after the speech... They swooned over the greatness of Obama’s speech. He *lifted his voice* in authority a couple times, but he stumbled on basic words. He came on excited to Accept the Democratic Nomination, then turned into Just Another Politician (enTenTen13).

(13) “In my entire career in the industry I have never had someone react to an evaluation the way Robert did...” the regional manager would later admit. “It’s not unusual for people to be disappointed or defensive when given a poor job performance rating, but Robert immediately began *to raise his voice* ... He was shouting at me and telling me I didn’t know what I was doing. It took me over an hour to calm him down” (enTenTen13).

The second difference is connected to the secondary figurative meaning of these two idioms: ‘to raise a voice of protest against somebody/something’. Out of 273 contexts found in the corpus the idiom *lift one’s voice* is used only once in the meaning ‘to raise a voice of protest against somebody/something’. In other cases, the idiom is used in its direct meaning ‘to speak louder’. Let us consider this single example in the corpus, cf.:

(14) This failure was a fresh and yet mortifying disappointment, and his end was a gloomy and somewhat obscure one, but he will always be remembered with gratitude as one of the first who in the Irish Parliament *lifted his voice* against



those restrictions under which the prosperity of the country lay shackled and all but dead (enTenTen13).

The third distinguishing characteristics of the idiom (10) is that in 70 % of its usage in authentic discourse it is practiced in religious texts and mainly in the meaning: ‘to chant in the praise of God’, cf.:

(15) Wherever stress there is in preparation for the day melts away during the Shabbat morning service as your child leads the congregation in prayer; chants beautifully from Torah and Haftarah and teaches the congregation with his (or her) d’var torah. Samuel is also a ‘sweet singer in Israel’ so it was a special treat to hear him *lift his voice* to God (enTenTen13).

Analysis of semantic differences of idioms *lift one’s voice* (10) and *raise one’s voice* (11) helps us in structuring semantic fields of the Thesaurus. Thus, *lift one’s voice* falls under three taxa: 1. *Punishment* with its sub-taxa: ‘reproach’, ‘blame’, ‘reprimand’ or ‘accusation’. 2. Religion (chanting). 3. Communication (chanting). The idiom *raise one’s voice* falls under several taxa as well. We can raise voice under different circumstances: attracting someone’s attention, in the process of argument, in the state of irritation, frustration, anger, annoyance and other emotional conditions. The main outstanding semantic difference between idioms (10) and (11) is that the idiom *to raise one’s voice* has the meaning ‘to raise a voice of protest against somebody/something’, whereas the idiom *to lift one’s voice* is used only once in this meaning. So, the idiom (11) falls in our Thesaurus under the following taxa: 1. Protest. 2. Punishment in connection with its sub-taxa ‘reproach’, ‘blame’, ‘reprimand’, ‘accusation’ or ‘blame’. 3. Attention and its sub-taxa ‘attracting attention’. 4. Behavior with its sub-taxa ‘improper argument behavior’. 5. Emotions with its sub-taxa: ‘irritation, annoyance’, ‘frustration’, ‘anger’ ‘grief’ etc. Definitely, this is not the full picture of presenting idioms in the semantic network, because besides hierarchical links there are as well paradigmatic (horizontal) links in the taxon structure. The zone of paradigmatic references is presented by the sign →. Thus, idioms of the semantic field *Prison* are connected semantically with the idea of *Punishment*, that is why the reference from the taxon *Punishment* → is applied to the taxon → *Prison*. The taxon organization with all its paradigmatic references is presented in the Table 1 in next section of the paper.

### 3 Idiom Classification: Results

#### 4.1. Stage 1. Collection of data

At the first stage of our work (in 2007) the list of English and American idioms consisted of 2000 units collected manually from fiction, academic journals, popular newspapers and magazines. Now it comprises about 6300 idioms and 11500 contexts of their usage in authentic contexts from the corpora. The process of collection of empirical material continued at the first stage of research where idioms were drawn from monolingual, bilingual and phraseological dictionaries. The goal of the first stage was to assign a certain descriptor to each idiom under consideration. This was followed by classification of idioms on the basis of their semantic description. Idioms of the same conceptual field were organized under a taxon that is the basic unit of the thesaural representation of idioms and therefore is the main entry-form of the dictionary labeled by a relevant descriptor. For example, the following idioms are organized in the dictionary in one group under the taxon *True-Untrue*:

- (1) *to talk turkey* “to discuss something directly and honestly”, which is regarded as being True;
- (2) *to cook the books* “falsify a company’s financial records”, which is regarded as being Untrue;
- (3) *snow job* “a systematic deception; a deceptive story that tries to hide the truth” which is regarded as being Untrue.

Table 1 below displays taxon *True-Untrue*, which is one of the 82 idiom-thesaurus taxa. The sign (→) indicates paradigmatic references to other taxa of the dictionary. We display the contents of the taxon concisely, presenting only two idiom-examples illustrating each taxon and its sub taxa. In general, our empirical material consists of 234 idioms belonging to this taxon.

<i>True-Untrue</i> 1. Truth – Lie, Deception 1.1. Truth <i>talk turkey; call a spade a spade</i> 1.2 Lie, Deception →false status →betrayal 1.2.1. Participants, Instruments of situation of lie and deception <i>decoy duck; stool pigeon</i> 1.2.2.2. Non-verbal lie →theft, stealing <i>draw a red herring; take s.o. for a ride; cook the books</i> 1.2.2.1. Pretence →dishonesty, insincerity, hypocrisy	<i>play the dumb; crocodile tears</i> 1.2.2.2. Pseudo-art <i>soap opera; cock-and-bull-story</i> 1.2.3. Deceit, Verbal lie, Misrepresentation <i>a snow job; full of hot air</i> 1.2.4. Self-delusion, Illusions <i>a mare’s nest; build castles in the air</i> 1.3. Honesty – Dishonesty 1.3.1. Honesty, Openness, Sincerity →oath <i>a square deal; an honest Joe</i> 1.3.2. dishonesty, insincerity, hypocrisy →cunning...→prentence→immorality <i>brown nose; play possum</i>
---	--



Table 1

Big size taxa can fall into sub-taxa, e.g. the conceptual field TRUE-UNTRUE has one sub-taxon TRUTH-LIE, DECEPTION. The concept LIE, DECEPTION in its turn has its own four sub-taxa. Each sub-taxon can break down further, thus the concept NON-VERBAL LIE falls into two sub-taxa: PRETENCE and PSEUDO-ART. In some cases, the depth of such division can be up to five levels. Due to space limitations we cannot provide authentic corpora examples in the Table 1.

## 4.2. Stage 2. Verification of Conceptual Marking

The second stage of classification is connected with checking the correctness of the conceptual marking of the idioms under study, which is of primary importance for regrouping the existing stock of idioms.

## 4.3. Stage 3. Idioms in Corpora

At present time the third stage of work is being carried out and it is connected with the search of idioms in corpora with the purpose of verification of idiom usage in contemporary discourse. Since the language changes, in many cases information about idioms in corpora differs from that in the dictionaries and for that reason we had to make some changes in classification which had been done on the previous stages of our work. For example, due to corpus analysis new polysemous idioms have been retrieved from authentic contexts of their usage. As a result, some idioms fell under other taxa of the thesaurus than they had been grouped earlier. We will present advantages of corpora in retrieving new polysemous idioms. To illustrate, majority of dictionaries give two meanings of an idiom *on the cuff*, cf.:

1) 'on credit'

His cleaning lady has volunteered to go *on the cuff* when he explained to her about his cash-flow problems (BNC).

2) 'free of charge'

The press agent gets no pay but only a certain amount of drinks *on the cuff* (BNC).

However, the comprehensive study of corpora and retrieving authentic examples allows registering two additional meanings of the idiom:

3) 'confidentially'

But strictly *on the cuff* I'm willing to bet he never did see it and that he never heard of Mildred ... (BNC).

4) 'spontaneous, without previous preparation'

We have a little segment here 'They play it and I say it'. They are going to just pick out some things from the speech and I am going to respond to them *on the cuff* (COCA).

As a result, the idiom *on the cuff* now falls under five taxa of the dictionary:

1) In its first meaning 'on credit' it falls under a taxon Money with its sub-taxon Debts;

2) In its second meaning 'free of charge' it also falls under the same taxon Money, but under different sub-taxon, which is Free of Charge;

3) In its third meaning 'confidentially' the idiom falls under a taxon Mystery, Secrecy;

4) In its fourth meaning 'spontaneous' without previous preparation' the idiom falls under three taxa:

a. Behaviour with its sub-taxon Spontaneity;

b. Time and its sub-taxon Spontaneity;

c. Freedom and its sub-taxon Natural Action in the meaning: 'not forced action', 'an action of free will'.

## 5. Conclusions

The research based on theoretical concepts developed by Baranov, Dobrovol'skij and Piirainen enabled us to apply their strategy to construction of Thesaurus of English Idioms. Results of the study introduce a new approach to phraseography, that is a thesaural principle of structuring of semantic fields of English idioms.

Our analysis clearly demonstrates that advantages of application corpus approach to lexicographic research are evident. Due to corpus data, the dictionary presents a range of syntactic patterns, idiom variation, synonymous and polysemous idioms which cannot be retrieved from the existing idiomatic bilingual and monolingual dictionaries of the English language. Apart from its theoretic relevance as an instrument of description of the mental lexicon, a new ideographic dictionary of English idioms based on corpus data can be used for purposes of language acquisition and translation. Most set phrases can be translated correctly only if we take the context into account, something that many dictionaries fail to do in a systematic way. The compilation of a corpus-based Thesaurus of English idioms based on authentic data is a question of vital importance for modern theoretical phraseology and practical lexicography.

## References



- Apresjan J. (1993). New English-Russian Dictionary. Vol.1-2.– Moscow: Russkij yazyk.
- Apresjan J. (1994). New English-Russian Dictionary.Vol.3. – Moscow: Russkij yazyk.
- Baranov, A., Dobrovol'skij, D. (2007). *Thesaurus of Present-day Russian Idioms*. – Moscow: Avanta.
- Dobrovol'skij, D. (1992). Phraseological universals: theoretical and applied aspects. – In: *Meaning and grammar. Cross-linguistic perspectives*. – Berlin – New York: Mouton de Gruyter, pp.279-301.
- Dobrovol'skij, D. (1994). Idioms in a Semantic Network: Towards a New Dictionary-Type. In: *Proceedings of the 6th EURALEX International Congress*, Amsterdam, pp.263-270.
- Dobrovol'skij, D., Piirainen,E. (2005). Figurative language. Cross-cultural and cross-linguistic perspectives. – Amsterdam: Elsevier.
- Gizatova, G. (2016). A corpus-based Approach to Lexicography: Towards a Thesaurus of English Idioms. In: *Proceedings of the 17th EURALEX International Congress*, Tbilisi, pp. 348-354.



# Intensifiers/moderators of verbal multiword expressions in Modern Greek

Mexa M.<sup>1</sup>, Markantonatou S.<sup>2</sup>

<sup>1</sup> University of the Peloponnese, Greece

<sup>2</sup> Institute for Language and Speech Processing/Athena R.C., Greece

## Abstract

We present a comprehensive view of the expression of degree modification of Modern Greek (MG) verbal multiword expressions (VMWEs) with the use of lexical elements that are not part of the VMWE. Our research draws on about 550 natural examples, retrieved from the web, for 63 VMWEs denoting ANGER, SURPRISE, AGONY, FRIGHT, ANXIETY and LOVE. Three general categories of modifiers of this type were recognized: (i) lexical elements that display intensifying or attenuating/mitigating functions as a result of grammaticalization or emphatic stress, (ii) the definite and indefinite article, intensifying *και* ‘and’ and, (iii) lexical elements expressing levels of gradable properties. The lexical elements in the first two categories seem to apply with a much wider VMWE population than (most of) the items of the group (iii) which is the only group of degree modifiers that seems to need to be recorded in a VMWE lexicon.

**Keywords:** Verbal Multiword Expressions; degree modification; lexicography

## 1 Introduction

To the best of our knowledge, this the first attempt to provide a comprehensive view of the expression of degree modification (amplification/intensification or attenuation/mitigation) of Modern Greek (MG) verbal multiword expressions (VMWEs) with the use of lexical elements (either single word ones or multiword expressions) that are not part of the VMWE (We will also use the term “lexicalized” for the non-free parts of a VMWE, for instance, for the English VMWE “She will kick the bucket” lexicalized are the words “to kick”, “the”, “bucket”. The term has been introduced by Savary et al. (2018)). The aim is the updating of IDION, a web-based lexicographic environment for the multidimensional documentation of MG MWEs (Markantonatou et al. 2019) with information about idiosyncratic behavior as regards degree modification phenomena. 63 VMWEs denoting ANGER, SURPRISE, AGONY, FRIGHT, ANXIETY and LOVE<sup>1</sup> were retrieved from IDION. We should point out that we do not make a distinction among the various categories of VMWE according to semantic criteria, for instance manner of contribution of the lexicalized parts of the VMWE to the overall semantics of the construct, or syntactic criteria, for instance degree of fixedness. The reason for this choice is that the classification of the modifiers according to their distribution is beyond the scope of this work that includes the identification of these words, the description of their semantic contribution as intensifiers or mitigators and the brief discussion of other discourse functions these words may fulfill. Our research draws on 544 actual usage examples featuring these VMWEs that were retrieved with the Google browser. We focused on the discovery of all possible structures involving intensification/mitigation lexical elements rather than the quantitative study of the phenomenon which we leave for future research. Three general categories of lexical elements used for degree modification of Modern Greek VMWEs were recognized: (i) lexical elements that display intensifying or mitigating functions as a result of grammaticalization or emphatic stress, (ii) the definite and indefinite article, intensifying *και* ‘and’ and, (iii) lexical elements expressing levels of gradable properties. We observe that these lexical elements fall in two groups in terms of distributional behavior. Lexical elements of the first two categories seem to apply with a much wider VMWE population than (most of) the items of the group (iii); distribution particularities are a fact that lexicographic practice might take into account.

## 2 Grammaticalized lexical items

Grammaticalized expressions include the adverbs *λίγο* ‘a little’, *κυριολεκτικά/στην κυριολεξία* ‘literally’, *πραγματικά, όντως* ‘really’, *πράγματι* ‘in fact’, *ειλικρινά* and the prepositional phrase *στ’ αλήθεια* ‘frankly’.

To the best of our knowledge, there is little to nothing as regards the literature on the impressively frequent use of *κυριολεκτικά/πραγματικά* with MG VMWEs. Our data indicate that *κυριολεκτικά* has similar functions with its English translational equivalent “literally”: it marks the speaker’s commitment to what his/her utterance denotes, it indicates the dual –literal and metaphorical– meaning of an idiom and assigns the maximal intensity to a metaphorical expression that expresses exaggeration (Israel 2002; Nerlich & Dominguez 2003). *Κυριολεκτικά* may occupy a pre- (1) or post- (2) VMWE position but it strongly prefers the one immediately after the verb (3):

<sup>1</sup> There is extensive literature with reference to VMWEs denoting emotions, but this subject is beyond the scope of our work. For a recent review on the topic of how emotions are expressed by VMWEs in Modern Greek (in comparison to French), see Fotopoulou & Giouli (2018).



(1) Κυριολεκτικά τα πήρα στο κρανίο!  
 Kiriolektika ta pira sto kranio!  
 literally them took.01.SG on.the cranium  
 ‘I literally got mad!’

(2) Μένω άναυδος κυριολεκτικά.  
 Meno anavdos kiriolektika.  
 stay.01.SG speechless literally  
 ‘I am literally stunned.’

(3) Μου κόπηκαν κυριολεκτικά τα γόνατα όταν τον ξαναείδα.  
 Mou kopikan kiriolektika ta yonata otan ton xanaída.  
 my.GEN cut.03.PL literally the knees.NOM when him saw.01.SG again  
 ‘I literally went weak at the knees when I saw him again.’

*Πραγματικά* (4-6) has the same syntactic preferences with *κυριολεκτικά* and has functions similar to its English translational equivalent “really”: it ensures the truth of the meaning of the utterance (Israel 2002; Paradis 2003) and indicates emphasis (Paradis 2003) or exaggeration as regards the expression of the speaker’s emotions (Bordet 2017). *Όντως, πράγματι, ειλικρινά* and *στ’ αλήθεια* have the same functions with *πραγματικά*.

(4) Πραγματικά έμεινα εμβρόντητος από την απάντηση του υπουργού Υγείας.  
 Praymatika emina emvrontitos apo tin apadisi tu ipurygu Iyias.  
 really stayed.01.SG stunned from the answer of.the minister of Health  
 ‘The minister’s of Health answer came as a great surprise to me.’

(5) Βέβαια υπάρχουν φορές που χάνει την υπομονή της και τότε γίνεται πραγματικά πυρ και μανία.  
 Vevea iparxun fores pu xani tin ipomoni tis ke tote ginete praymatika pir ke mania.  
 of course are.03.PL times that loses the patience hers and then becomes really fire and mania  
 ‘Of course there are times when she loses patience and she really gets furious.’

(6) Κάθε απάντηση είναι χρήσιμη γιατί είμαι σε αναμμένα κάρβουνα πραγματικά...  
 Kaθe apadisi ine xrisimi yati ime se anamena karvuna praymatika...  
 every answer is useful because am on lit coals really  
 ‘Every answer would be useful because I am very worried really...’

*Λίγο* functions as a politeness marker (Κλαίρης & Μπαμπινιώτης 2004) and as a verbal diminutivizer denoting emotion as a result of grammaticalization (Canakis 2012). However, *λίγο* may have an ambiguous function, namely either as an adverbial quantifier – when the emphatic stress is on *λίγο* or as a verbal diminutivizer – when the emphatic stress is on the verb; this is a similar though distinct function from that of the prefix *ψιλο-* ‘a little’ or ‘(s)lightly’ as in *ψιλοδοιλέω* ‘work a little’ (Canakis 2012: 178). Σαββίδου (2012), also, argues that the *ψιλο-* may have a mitigating descriptive meaning or mitigating pragmatic function.

Since our data are textual, it is not always easy to distinguish between the two possible functions. In our data *λίγο* most probably functions as a verbal diminutivizer – like the mitigating pragmatic *ψιλο-* – though it could be a quantifier as well (7):

(7) Και μένει λίγο σέκος η κοπελιά  
 Ke meni liyo sekos i kopelia  
 and stays a bit numb the girl.NOM  
 ‘The girl was left speechless in a sense’

In the light of the discussion in Canakis (2012), we searched the web for VMWEs that can both accept *λίγο* as a modifier and appear with their verb head modified by the prefix *ψιλο-*. The retrieved data confirm this possibility, at least in certain contexts, and this fact indicates that for some speakers *λίγο* and *ψιλο-* function in a similar way (8-9):

(8) Όταν διάβασα ότι δεν είναι πλέον εν ζωή, έμεινα λίγο κάγκελο...  
 Otan diavasa oti den ine pleon en zoi, emina liyo kagelo...  
 when read.01.SG that not is any more in life, stayed.01.SG a bit rail  
 ‘When I read that he has passed away, I was a bit shocked...’

(9) Ψιλοέμεινα κάγκελο και έφυγα με το Johnnie μου λίγα λεπτά αργότερα!  
 Psiloemina kagelo ke efiya me to Johnnie mu liya leptá arýotera!  
 slightly-stayed.01.SG rail and left with the Johnnie my a few minutes later  
 ‘I was a bit shocked and left with my Johnnie a few minutes later!’

### 3 The articles



In our data we observe that both the definite and indefinite article may have an intensifying/emphatic function which depends on whether there is an emphatic stress on them or not.

The definite article *ο, η, το* appears with VMWEs with a fixed noun phrase functioning as a subject or an object complement. It functions as an intensifier when it is emphatically stressed especially in the informal, colloquial speech (Κλαίρης & Μπαμπινιώτης 2004; Apostolou-Panara 1994; Σαλτίδου 2018; Τσιακμάκης 2017). In the following example (10) from our data the emphatic stress on the article is already represented with capital letters both for the article *ΤΟ* ‘the’ and the word *ΣΟΚ* ‘shock’.

- (10) Και έφτασαν τα προϊόντα και έπαθα ΤΟ ΣΟΚ!  
 Ke eftasan ta proioda ke epaθα to sok!  
 and arrived.03.PL the products.NOM and suffered.01.SG the shock.ACC  
 ‘I was so shocked when the products arrived!’

The indefinite article *ένας, μια, ένα* appears with VMWEs with a fixed noun phrase functioning as a subject or object or copula complement. With the appropriate intonation it functions as an intensifier (Τζάρτζανος 1946; Κλαίρης & Μπαμπινιώτης 2004), as an emphaticizer (Holton et al. 2000; Mackridge 1999; Χιώτη 2010) and as a moderator (Σαββίδου 2012). In our data the intensifying or mitigating function of the indefinite article depends on whether an emphatic stress can be assumed or not (since our data are only textual) (11):

- (11) Εκείνη την ώρα με έπιασε ένας κρύος ιδρώτας.  
 Ekini tin ora me epiaσε enas krios idrotas.  
 that the hour me took.03.SG a cold sweat.NOM  
 ‘At that very moment I broke out in a cold sweat.’

The conjunction *και* ‘and’, among others, also has the function of an intensifier/emphaticizer (Τζάρτζανος 1953; Canakis 1996; Κλαίρης & Μπαμπινιώτης 2004; Mackridge 1999; Χιώτη 2010). Canakis (1996), who is the first to systematically analyze the functions of *και* ‘and’, points out that it functions as a filler in a negative context and it means ‘after all’ [‘πια’ or ‘δα’]. In our data this is the most common context where *και* ‘and’ appears and although its absence has no effect in the utterance meaning, its use denotes emphasis (12):

- (12) Είμαι ερωτευμένη. Δεν κόβω και φλέβα.  
 Ime erotevmeni. Den kovo ke fleva.  
 am in love. not cut.01.SG and vein.  
 ‘I am in love with him but I wouldn’t die for him.’

As regards the distribution of the grammaticalized lexical items, *κυριολεκτικά* seems to apply freely. However, we have not been able to find an example of the use of the VMWE *μου κόβονται τα ήπατα*, Lit. to.me cut.PASS the liver.NOM, ‘I am terrified’ with *κυριολεκτικά* and this might be an indication that the particular VMWE, and probably other VMWEs that we have not studied yet, do not combine with this adverb. A much wider search of the web, probably supported by Natural Language Processing tools, as well as well-designed studies of native speaker intuitions would be necessary in order to arrive to a precise picture about the distribution of *κυριολεκτικά*. The situation seems to be the same with *πραγματικά* and *λίγο*. The distribution of the emphatic stress on the articles seems to be constrained only by the syntax of the VMWE, that is whether there is an article to be stressed or not. The remaining adverbs and PPs are in less use since their functions overlap with the functions of the widely used *κυριολεκτικά*, *πραγματικά*, *λίγο* and there are no indications that they are subjected to particular selection constraints by the VMWEs. This overall picture suggests that there is no need to record information about the distribution of these lexical items in the lexical entries of the VMWEs.<sup>2</sup>

## 4 Degree modification

In this section we focus on lexical elements expressing levels of gradable properties, such as inherently degree adverbs, and various manner adverbs as well as adjectives, which function as intensifiers.

### 4.1 Πολύ/εντελώς

Degree adverbs occurring with VMWEs may denote upper or lower points in a climax (maximizers *εντελώς/τελείως* ‘perfectly, completely’ and approximators *σχεδόν* ‘nearly’ (13)). They may also be degree modifiers of a property/situation (boosters *πολύ* ‘a lot, much’, moderators *σχετικά, κάπως* ‘somewhat’ (14)) (Paradis 1997; Γαβρηλίδου 2013; Gavriilidou 2015).

- (13) Μπορώ να πω πως έμεινα σχεδόν μαλάκας!  
 Boro na po pos emina σχεδόν malakas!  
 can to say.01.SG that stayed.01.SG nearly jerk  
 ‘I could admit that I was nearly blown away!’

<sup>2</sup> We would like to thank an anonymous reviewer for drawing our attention to this point.



- (14) Εγώ ίσως και να ήμουν κάπως τσιμπημένη μαζί του, μα αυτός δεν ήθελε να το δει.  
 Ego isos ke na imoun kapos tsibimeni mazi tu, ma aftos den ithele na to di.  
 I maybe and was somewhat pinched with him, but he didn't want to it see  
 'I was probably a bit soft on him, but he did not want to see it.'

Given that Γαβριηλίδου (2013) argues that the modification of inherently intensified verbs with *πολύ* is redundant and considers VMWEs as inherently intensified expressions, it can be inferred that the modification of VMWEs with *πολύ* would be redundant. Additionally, Gavriilidou & Giannakidou (2016) point out that *πολύ* modifies gradable verbs/participles that do not combine with maximizers, such as *εντελώς*. Coming to the object of this research, although the studied VMWEs express the maximum of a property/situation, a considerable percentage of them (11 out of 63) were found to be gradable and some of them occurred with both *πολύ* (15) and *εντελώς* (16):

- (15) Κάτι φώναζε εκεί πέρα, κι εγώ τα πήρα πολύ στο κρανίο!  
 Kati fonaxe eki pera, ki ego ta pira poli sto kranio!  
 something shouted.03.SG over there, and I them took a lot on.the cranium  
 'Someone shouted something over there and I got really furious!'

- (16) Τα πήρα εντελώς στο κρανίο ...  
 Ta pira edelos sto kranio...  
 them took.01.SG completely on.the cranium  
 'I got absolutely furious'

## 4.2 Adverbs of manner

In our data we found the manner adverbs *χοντρά*, Lit. 'fatly', *άγρια* 'wildly', *γερά*, *δυνατά* 'strongly', *τρελά* 'madly', *κανονικά* 'normally', *σοβαρά* 'seriously', *επικίνδυνα* 'dangerously' which, in the spirit of Γαβριηλίδου (2013), we could consider as boosters (17). We observe that these adverbs select the categories ANGER, LOVE but not SURPRISE, AGONY, FRIGHT. This could be an indication that selection restrictions hold as regards the distribution of these adverbs. Of course, the further study on these categories would lead us to safer conclusions. We suggest that such restrictions are recorded in the lexicographic description of a MG VMWE because it does not seem possible to obtain this information from some general rule of the language (Hanks, 2013: 54).

- (17) Ποδοσφαιριστής δάγκωσε χοντρά τη... λαμαρίνα με σαγηνευτική αοιδό!  
 Podosferistis dagose ti... lamarina me sayineftiki aido!  
 football player bit fatly the sheet iron with seductive singer  
 'A football player fell hard for a seductive singer!'

## 5 External modification with VMWEs

The following cases can be considered "external modifiers" (Ernst 1981). External modification by an adjective that modifies a lexicalized noun of the VMWE occurs when the adjective can be paraphrased with an adverb (in (18) the relevant adverb is "sociologically") that takes the meaning of the whole VMWE in its scope rather than the meaning of the modified noun only (19); "internal modification" occurs if the adjective takes only the meaning of the modified noun in its scope.

- (18) Don't rock the sociological boat with your ideas. (Gehrke & McNally 2019: 782)

- (19) The federal agency decided to take the project under its well-muscled wing. (Gehrke & McNally 2019: 781)

Examples (20-21) below are instances of external modification because the adjective can be paraphrased with the adverbs *πολύ/πάρα πολύ* 'very much' applying to the meaning of the whole VMWE and not to the modified part of the VMWE only:

1. The head noun of the Noun Phrase that functions as the lexicalized subject, object or copula complement of a VMWE can be modified by an intensifying adjective. In our data, the most frequent intensifier of this type is (*πολύ*) *μεγάλος* '(very) big' while the maximum degree is expressed with the adjective *τεράστιος* 'huge'.

- (20) Έπαθα μεγάλο σοκ  
 Epaθα megalο sok  
 suffered.01.SG big shock  
 'I was shocked very much.'

The adjectives *μεγάλος* and *πολύς* seem to have a wide distribution constrained by the same rules as in the general language, therefore their application will not be mentioned in the lexicographic entry of the VMWE.



2. Particular VMWEs select particular intensifying adjectives: *μου έρχεται βαριά/χοντρή κεραμίδα*, Lit. to me comes heavy/thick roof tile, *παθαίνω διπλό/τριπλό/τετραπλό/πενταπλό/... εγκεφαλικό*, Lit. I suffer double/triple/four-/five-tuple/... stroke, *παθαίνω απίστευτο/τρελό σοκ*, Lit. I suffer unbelievable/mad shock, *παθαίνω τρελή κολούμπρα*, Lit. I suffer mad CRANBERRY WORD, *έγινα σωστό/πραγματικό/άγριο θηρίο/θηρίο ανήμερο*, Lit. I became proper/real/wild beast/beast untamed.

- (21) Και βέβαια μας ήρθε χοντρή κεραμίδα...  
 Ke vevea mas irthe xodri keramida...  
 and of course us came.03.SG thick roof tile.NOM  
 ‘And of course, we were floored’

In this case, each VMWE selects specific intensifying adjectives; the distribution of the adjectives is clearly idiosyncratic and this is a fact that has to be recorded in the lexicographical description of a VMWE. A similar idea has been implemented in DUELME (Grégoire, 2010) as a special device for encoding the modifiers of the lexicalized nouns in a VMWE (such as the heads of fixed object NPs) whether these modifiers are fixed as in the cases of *ανήμερο* θηρίο and *χοντρή/βαριά κεραμίδα*, or less fixed such as the adjectives “sociological” and “well-muscled” in the (18) and (19) respectively.

The extent of the fixed subject or object can be used as an intensifier (Gehrke & McNally (2019) mention *blow off steam* < *blow off a lot of steam*). In MG we find the adjective *όλος* ‘all, whole’ (22) and numerals (23). This property should be mentioned in the entries of the VMWEs that are found in the corresponding structures.

- (22) Του Μήτσου του είχε ανέβει όλο το αίμα στο κεφάλι.  
 Tu Mitsu tu ixe anevi olo to ema sto kefali.  
 the Mitsos he.GEN had ascended all the blood to.the head  
 ‘Mitsos was furious.’

- (23) Το επεισόδιο ξεκινάει με το άγρυχο κορμί του Jon και ο θεατής περιμένει –  
 To episodio xekinai me to apsiχο kormi tu Jon ke o theatis perimeni –  
 the episode begins with the dead body the.GEN Jon and the viewer waits –  
 έχοντας φάει και τα 20 νύχια του – το πότε θα αναστηθεί.  
 exodas fai ke ta 20 nixia tu – to pote tha anastithi.  
 having eaten and the 20 nails his – the when will resurrect.03.SG.MIDDLE  
 ‘The episode begins with John’s dead body and the viewer waits – in great agony – when it will be resurrected.’

Lastly, the (very) formal/learned lexical elements *ολίγον τι*, *ολίγον* ‘a little’, *ελαφρώς* ‘slightly’, *όντως* ‘really’, *αγρίως* ‘wildly’ were found functioning as degree modifiers. For instance, in the example below, *κράνα* ‘cranium’ is strongly colloquial; its co-occurrence with the formal *ολίγον* may result to intensification/emphasis (Καμηλάκη 2009) or add a humorous tint to the expression (Αναστασιάδη-Συμεωνίδη & Φλιάτουρας 2004) or, as in this example, function as a politeness marker moderating the probably negative impression created by the colloquial *κράνα* (24):

- (24) Συγγνώμη και εγώ που τα πήρα ολίγον στην κράνα, αλλά δεν είσαι ο πρώτος,  
 Siynomi ke ego pu ta pira oliyon stin krana, ala den ise o protos  
 sorry and I that them took a little on.the cranium, but not be.02.SG the first  
 ούτε καν ο δέκατος...  
 ute kan o dekaτος...  
 neither even the tenth  
 ‘I am also sorry that I got somewhat furious, but you are not the first, not even the tenth...’

## 6 Conclusion

To sum up, our data show that most of the studied VMWE co-occur with the adverbs *κυριολεκτικά* ‘literally’, *πραγματικά* ‘really’, which have multiple functions including intensification/emphasis, as well as *λίγο* ‘a little’, whose function is ambiguous – either as a quantifier or a mitigating pragmatic indicator. Also, it seems that the distribution of MG definite and indefinite article depends on the syntactic structure of the VMWEs. Therefore, it is not necessary to record the above lexical items in the lexicographical documentation of the VMWEs. On the other hand, the fact that some VMWEs choose specific intensifying adverbs (e.g. *χοντρά*, ‘fatly’, *άγρια* ‘wildly’) or adjectives (e.g. *ανήμερο*, ‘untamed’, *βαριά*, ‘heavy’) makes their lexicographic record necessary as it is indicative of idiosyncratic behavior. In addition, it would be useful to record the simultaneous modification of some VMWEs by both *πολύ* ‘very’ and *εντελώς/τελείως* ‘perfectly/completely’, because it indicates that the intensity of the emotion can not only be enhanced but also maximized despite the fact that VMWEs are inherently intensifying lexical elements. Undoubtedly, the present study could be expanded to cover more VMWEs of the same or different semantic fields in order to shed more light to the idiosyncratic way(s) in which degree



modification is expressed and enriches their lexicographic documentation. Finally, further study of the phenomenon could include classification of the modifiers according to their distribution.

## 7 References

- Αναστασιάδη-Συμεωνίδη, Α., Φλιάτουρας, Α. (2004). Η διάκριση [λόγιο] και [λαϊκό] στην ελληνική γλώσσα: Ορισμός και ταξινόμηση. In *Πρακτικά του 6ου Διεθνούς Συνεδρίου Ελληνικής Γλωσσολογίας, ICGL6, 18-21 Σεπτεμβρίου 2003*. Πανεπιστήμιο Κρήτης, Ρέθυμνο.
- Γαβριηλίδου, Ζ. (2013). *Όψεις επίτασης στη Νέα Ελληνική*. Θεσσαλονίκη: Αφοί Κυριακίδη.
- Καμηλάκη, Μ. (2009). Τα λόγια στοιχεία στη νεανική επικοινωνία: κοινωνιοπραγματολογική διερεύνηση της ποικιλίας [± ΛΟΓΙΟ] στον γραπτό λόγο νεαρών ομιλητών της Νέας Ελληνικής. Διδακτορική Διατριβή. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Αθήνα.
- Κλαίρης, Χ., Μπαμπινιώτης, Γ. (2004). *Γραμματική της Νέας Ελληνικής: Δομολειτουργική – Επικοινωνιακή*. Αθήνα: Ελληνικά Γράμματα.
- Holton, D., Mackridge, P. & Φιλίππακη-Warburton E. (2000). *Γραμματική της Ελληνικής Γλώσσας*. Αθήνα: Πατάκης.
- Mackridge, P. (1999). *Η Νεοελληνική Γλώσσα*. Αθήνα: Πατάκης.
- Σαββίδου, Π. (2012). Μετριασμός και επίταση με τη χρήση των μορφημάτων *ψιλο-* και *θεο-*: ανάλυση σε σώματα κειμένων. In Z. Gavrilidou, A. Efthymiou, E. Thomadaki, P. Kambakis-Vougiouklis (eds.), *Selected Papers of the 10<sup>th</sup> International Conference of Greek Linguistics*, Komotini, 1-4 September 2011. Komotini: Democritus University of Thrace, pp. 1090-1099.
- Σαλτίδου, Θ. (2018). Η «γλώσσα των νέων» ως υφολογικός πόρος όλων των ηλικιών στον λόγο τηλεοπτικών σειρών. Διδακτορική Διατριβή. Πανεπιστήμιο Δυτικής Μακεδονίας. Φλώρινα.
- Τζάρτζανος, Α. (1946). *Νεοελληνική Σύνταξις (της Κοινής Δημοτικής), Τόμος Α'*. Αθήνα: Οργανισμός Εκδόσεως Σχολικών Βιβλίων.
- Τσιακμάκης, Ε. (2017). Το οριστικό άρθρο της Νέας Ελληνικής. Σκέψεις πάνω στην οριστικότητα και την α-πρόσωπη αναφορά. Μεταπτυχιακή Διπλωματική Εργασία. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης. Θεσσαλονίκη.
- Χιώτη, Α. (2010). Οι παγιωμένες εκφράσεις της Νέας Ελληνικής: Ιστορική διάσταση, ταξινόμηση και στερεοτυπικότητα. Διδακτορική Διατριβή. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης. Θεσσαλονίκη.

## NON-GREEK REFERENCES

- Apostolou-Panara, A. (1994). Language change under way? The case of the definite article in Modern Greek. In I. Philippaki-Warburton, K. Nicolaidis, M. Sifianou (eds.), *Themes in Greek Linguistics. Papers from the First International Conference on Greek Linguistics*. Amsterdam: John Benjamins.
- Bordet, L. (2017). From vogue words to lexicalized intensifying words: the renewal and recycling of intensifiers in English: A case-study of *very*, *really*, *so*, and *totally*. *Lexis* [Online], 10. Accessed at: <https://journals.openedition.org/lexis/1125> [25/04/2020].
- Canakis, C. (1996). KAI. The Story of a Conjunction. PhD Thesis. University of Chicago. Illinois, Chicago.
- Canakis, C. (2012). *liyo*: towards grammaticalized verbal diminutivization? In Z. Gavrilidou, A. Efthymiou, E. Thomadaki, P. Kambakis-Vougiouklis (eds.), *Selected Papers of the 10<sup>th</sup> International Conference of Greek Linguistics*, Komotini, 1-4 September 2011. Komotini: Democritus University of Thrace, pp. 177-185.
- Ernst, T. (1981). Grist for the linguistic mill: Idioms and 'extra' adjectives. In *Journal of Linguistic Research*, 1(3), pp. 51-68.
- Fotopoulou, A., Giouli, V. (2018). MWEs and the Emotion Lexicon: Typological and cross-lingual considerations. In M. Sailer, S. Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*. Berlin: Language Science Press, pp. 63-91.
- Gavrilidou, Z. (2015). Phraseology and degree modification. In *Μελέτες για την Ελληνική γλώσσα*, 35. *Πρακτικά της 35ης Ετήσιας Συνάντησης του Τομέα Γλωσσολογίας του Τμήματος Φιλολογίας του Α.Π.Θ.*, 8-10 Μαΐου 2014. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Θεσσαλονίκη, pp. 241-247.
- Gavrilidou, Z., Giannakidou, A. (2016). Degree modification and manner adverbs: Greek: *poli* 'very' vs. *kala* 'well'. In *Selected Papers of the 21<sup>st</sup> International Symposium of Theoretical and Applied Linguistics, ISTAL 21, 5-7 April 2013*. Aristotle University of Thessaloniki, Thessaloniki.
- Gehrke, B., McNally, L. (2019). Idioms and the syntax/semantics interface of descriptive content vs. reference. In *Linguistics*, 57(4), pp. 769-814.
- Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44 (1-2), pp. 23-39.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. The MIT Press.
- Israel, M. (2002). Literally speaking. In *Journal of Pragmatics*, 34, pp. 423-432.
- Markantonatou, S., Minos, P., Zakis, G., Moutzouri, V. & Chantou, M. (2019). IDION: A database for Modern Greek multiword expressions. In *Proceedings of Joint Workshop on Multiword Expressions and WordNet, MWE-WN 2019, Workshop at ACL 2019, 2 August 2019*. Association of Computational Linguistics.
- Nerlich, B., Dominguez, P. J. C. (2003). The use of *literally*: Vice or virtue? In F. J. R. de Mendoza Ibáñez (ed.) *Annual Review of Cognitive Linguistics*, 1, Amsterdam: John Benjamins, pp. 193-206.
- Paradis, C. (1997). Degree modifiers of adjectives in spoken British English. *Lund Studies in English*, 92. Lund: Lund University Press.
- Paradis, C. (2003). Between epistemic modality and degree: the case of *really*. In R. Facchinetti, M. Krug, F. Palmer



(eds.), *Modality in Contemporary English*. Mouton de Gruyter.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryigit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., & Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary, V. Vincze (eds.) *Multiword expressions at length and in depth, Extended papers from the MWE 2017 workshop, Valencia, 4 April 2017*. Berlin: Language Science Press, pp. 87-147.







# Building a Controlled Lexicon for Authoring Automotive Technical Documents

Miyata R., Sugino H.

*Nagoya University, Japan*

## Abstract

We describe the framework and the process of building a controlled lexicon, specifically intended for authoring Japanese automobile repair manuals. Focusing on verbs, we seek to control two types of linguistic variations: (1) synonymous words and (2) case (argument) order variations. For synonymous words, we comprehensively extracted verb tokens from a large text data set and classified each verb type as approved or unapproved. For case order variations, we descriptively analysed case structures of Japanese sentences in the data set and defined the canonical order. We also examined the status of the constructed lexicon in terms of coverage, which enables us to establish a tangible goal of future lexicon building. The resultant controlled lexicon with 910 verbs and 954 case patterns can help authors choose appropriate words and construct consistent sentence structures. In order to accomplish effective and efficient authoring, we further proposed and designed two types of authoring support tools: a sentence diagnostic tool that identifies unapproved variations of verbs and sentence structures, and a template-driven writing tool that helps writers compose controlled sentences by completing canonical case patterns.

**Keywords:** controlled lexicon building; technical authoring; descriptive analysis; variation management; grammatical case; automotive domain

## 1 Introduction

Controlled lexicon, or controlled vocabulary, is a list of approved words in a certain domain, which may further determine unapproved words and provide the definition and usage of the registered words (Nyberg et al. 2003; Warburton 2014). The deployment of controlled lexicon helps enhance the consistent use of words—in particular verbs, nouns, adjectives and adverbs—in writing technical documents and prevents ambiguous and difficult expressions, which will lead to not only improved readability but also translatability of the documents. Furthermore, in combination with well-managed bilingual dictionaries, we can envisage the improved quality of machine translation outputs.

In this study, we describe the framework and the process of building a controlled lexicon specifically intended for authoring the Japanese automobile repair manuals of Toyota Motor Corporation. For every new model of automobile, huge volumes of technical documents, such as repair manuals, are created and translated, and an assemblage of writers and translators are involved in the document production workflow. This makes it difficult to ensure linguistic consistency across documents, eventually inducing a lack of clarity in the readers' understanding. In this context, we are now developing a controlled language for Japanese automotive technical documents. Controlled languages for authoring and translation basically consist of a syntactic and a lexical component (Nyberg et al. 2003; Kuhn 2014). In this paper, we report on the compilation of a controlled lexicon with specific focus on verbs, since verbs are crucial building blocks for composing operational instructions for repair manuals and governing sentence structures such as predicate-argument structures.

Although many controlled languages have been developed for particular domains, including the automotive domain (Means & Godden 1996; Godden 2000), few are publicly available. One of the few exceptions is ASD Simplified Technical English, or ASD-STE (ASD 2017), which was originally developed for aerospace/maintenance documentation, and is now widely used in other industries. It defines writing rules that restrict certain syntactic/textual features, including sentence length and compound nouns, and provides a lexicon of approved and unapproved words. While ASD-STE is useful in its own right, it is not easy to directly port it to other purposes, domains and languages. In the case of controlled authoring of Japanese automotive technical documents, we also need extensive information of word usage. Thus, referring to the ASD-STE as a model example, we decided to build our controlled lexicon from scratch.

However, the practical problem is that few studies have established the general process of controlled lexicon building; in many cases, a controlled lexicon has been developed chiefly based on the tacit knowledge of domain experts and researchers. In this study, our lexicon building proceeded as follows: we first collected verb occurrences from existing texts, and then defined approved verbs and their canonical usage by analysing their occurrences. One of the important contributions of this paper is to document the controlled lexicon building process in detail, which will be helpful for related endeavours in the future.

The remainder of the paper is structured as follows. In Section 2, we design our controlled lexicon that enhances consistent writing of technical manuals. Section 3 describes the process of collecting verbs and controlling the variations to prepare a list of approved and unapproved verbs, presenting the growth of coverage in accordance with the building process. In Section 4, we further extend our controlled lexicon by defining the canonical case (argument) structure patterns for approved verbs. We then propose and design authoring support tools in Section 5 and conclude this paper with future outlook in Section 6.



## 2 Design of Controlled Lexicon

We address the two problems of inconsistent use of verbs: (1) synonyms and (2) case (argument) order. In automobile repair manuals, different verbs are sometimes used for the same operations, such as *koukansuru* and *torikaeru*, both of which mean ‘replace’. These variations violate the basic principle of controlled lexicon, that is, ‘one word – one meaning’ (Nyberg et al. 2003; Møller & Christoffersen 2006), and may hinder readers’ comprehension of the documents. Further, the notion of *case* (Fillmore 1968) is significant for controlled writing as Japanese case order is fairly free (Masuoka & Takubo 1992; Sasano & Okumura 2016), which sometimes causes structural variations. The following two sentences present an example of the different case orders of the Japanese verb *setsuzokusuru* (connect):

(1) GTS を DLC3 に接続する。 / GTS *o* DLC3 *ni* *setsuzokusuru*.

(2) DLC3 に GTS を接続する。 / DLC3 *ni* GTS *o* *setsuzokusuru*.

The order of the accusative case (-*o*) and the dative case (-*ni*) is different from each other. Both sentences are grammatically correct in Japanese and can be translated as ‘Connect GTS to DLC3’. They do not even necessarily hinder readers’ comprehension of the text. From the viewpoint of consistent authoring, however, these variations should be avoided. In addition, if we can fully reduce these variations in the source, we can expect the improved results in parsing, text retrieval and machine translation.

Here, we propose a controlled lexicon that can enhance the consistency of writing in Japanese by extending the existing framework of controlled languages. To address the problem of synonyms, based on the ASD-STE, we create a list of approved and unapproved words with word definitions and examples. To address the problem of case order, we further define the canonical case order for each verb.

Figure 1 shows examples of approved and unapproved words in our controlled lexicon. Each entry word has the part of speech, semantic category and example sentence(s) of the word. Unapproved words have the links to the approved words, while approved words have definitions. These descriptions of the words help writers consistently select an appropriate word in writing text. Furthermore, the entries of approved words accompany the canonical case order(s) to support writers to appropriately construct sentences in a controlled manner.

<b>Approved word</b>	交換する / <i>koukansuru</i>
<b>Part of speech</b>	verb
<b>Semantic category</b>	Action > Part
<b>Definition</b>	‘To remove an item and to install a new or serviceable item of the same type’. (ASD 2017)
<b>Canonical case order</b>	[PART/ITEM <i>o</i> ] [PART/ITEM <i>ni</i> ] <i>koukansuru</i>
<b>Example</b>	センサーを新品に交換する。 / <i>Sensa o shinpīn ni koukansuru</i> . (Replace the sensor with a new one.)
<b>Unapproved word</b>	取り替える / <i>torikaeru</i>
<b>Part of speech</b>	verb
<b>Semantic category</b>	Action > Part
<b>Approved alternative</b>	交換する / <i>koukansuru</i>
<b>Unapproved example</b>	必ず新品に取り替える。 / <i>Kanarazu shinpīn ni torikaeru</i> . (Always replace it with a new one.)

Figure 1: Examples of entries in the controlled lexicon: *koukansuru* and *torikaeru* (replace). For explanation, the definition of ‘replace’ is extracted from the specification of ASD-STE as a definition for *koukansuru*.

## 3 Construction of Controlled Lexicon

In this section, we elucidate how we collected approved and unapproved verbs from the text data of automotive technical documents. We also present the semantic categories of collected verbs and detailed analyses of the frequency of verb occurrence in the text data, which enables us to understand the status of lexicon in terms of coverage.

### 3.1 Verb Collection and Variation Control

From 17 sets of repair manuals that cover 10 types of automobiles from Toyota Motor Corporation, we comprehensively extracted verb tokens used in the main clauses of sentences, using Japanese sentence analysis tools JUMAN++V2 (Morita et al. 2015; Tolmachev et al. 2018) and KNP (Kawahara & Kurohashi 2006). We assume that they cover a sufficient range of verbs in this domain as we extensively investigated huge volumes of text data containing more than one million sentences. Subsequently, we eliminated verbs which were wrongly identified as verbs by the tools and rare compound verbs that can be replaced by simpler verbs. For example, *sokutei-kaishisuru* is a compound verb which combines the two simple expressions *sokutei* (measurement) and *kaishisuru* (start), and can be rephrased into *sokutei o*



*kaishisuru* (start taking a measurement) or, simply, *sokuteisuru* (measure). We finally collected 1,058,424 verb tokens and 910 verb types.

The next task was to classify whether each verb (type) is approved or unapproved. We conducted the following steps:

1. Gather semantically similar verbs
2. If a verb is mostly interchangeable with another verb in actual sentences, regard them as verb variations
3. Define one of the verb variations as approved and the rest unapproved
4. Link the unapproved word(s) to the approved verb

In Step 3, we used the frequency of the words in the data set as an important evidence for decision-making; the more frequently the word occurs in the data, the more likely that it is approved.

Through this process, we finally identified 822 approved and 88 unapproved verbs. Table 1 presents the basic statistics of the constructed controlled lexicon, showing that approximately 10% of verb types and 3% of verb tokens (i.e. verb occurrences in our data set) were labelled as variations. It suggests that even documents that were authored by technical writers contained a certain amount of inconsistent use of verbs and we can expect an improvement in document consistency by employing our constructed lexicon.

Table 2 shows the verb variation categories with their frequencies in data set and examples. The major category is the synonym, i.e. a word that conveys almost the same meaning. We also regarded the use of prefix to verbs as variations. The prefixes are productive and potentially create many verb types, which is not desirable from the viewpoint of controlled writing. The important point is that prefixed verbs can be decomposed into a simple combination of a verb and an adverb as follows:

(3) ダイアグノーシスコ드를再確認する。/*Daiagunoshisu-kodo o sai-kakuninsuru*.  
(**Re-check** the diagnosis code.)

(4) ダイアグノーシスコ드를再び確認する。/*Daiagunoshisu-kodo o futatabi kakuninsuru*.  
(**Check** the diagnosis code **again**.)

In this case, *sai-kakuninsuru* (re-check) can be decomposed into two simple words, *futatabi* (again) and *kakuninsuru* (check). We prohibited the following six types of prefixes and identified unapproved verbs: *sai-* (re-, again), *kari-* (temporary), *go-* (false), *zen-* (all), *ryo-* (both) and *shi-* (trial). Compound verbs combine two similar verbs. In many cases, it is possible to convert them into simpler verbs. The last category is the notational variation, which is peculiar to Japanese language. For example, the Japanese verb *atatameru* (warm up) can be written in different forms of kanji such as 温める and 暖める. The meaning of the two notations are almost similar and can be controlled.

	All #	Approved #	%	Unapproved #	%
Type	910	822	90.33	88	9.67
Token	1,058,424	1,027,659	97.09	30,765	2.91

Table 1: Statistics of the constructed controlled lexicon of verbs.

Category	#	Example (unapproved verb: corresponding approved verb)
Synonym	49	<i>jisshisuru</i> (conduct): <i>okonau</i> (do), <i>jokyosuru</i> (eliminate): <i>torinozoku</i> (remove)
Use of prefix	33	<i>sai-kakuninsuru</i> (re-check): <i>kakuninsuru</i> (check), <i>kari-koteisuru</i> (temporarily-fix): <i>torinozoku</i> (fix)
Compound verb	3	<i>ooi-kakusu</i> (cover and conceal): <i>oou</i> (cover), <i>arainagasu</i> (wash away): <i>arau</i> (wash)
Notation	3	温める: 暖める / <i>atatameru</i> (warm up), 判る: 分かる / <i>wakaru</i> (know)

Table 2: Verb variation types with examples.

### 3.2 Semantic Category of Verbs

In conjunction with controlling verb variations, we analysed each verb and labelled a semantic category. Based on the lexicographic categories defined in WordNet,<sup>1</sup> which is often called *supersenses* (Ciaramita & Altun 2006; Paaß & Reichartz 2009), we created a typology of semantic categories of verbs. Table 3 presents the typology of a semantic category of verbs with examples and frequency information. While WordNet defines 15 categories for verbs, namely, body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative and weather, we adopted the necessary classification for the automotive technical documentation and modified them as necessary to create top level categories. Second level categories were defined in a bottom up manner through the analysis of verbs.

All the 910 verb types were classified into one of the semantic categories. The dominant category is the Action > Part. The verbs in this category are used to express operational actions regarding automobile parts, which are the core building blocks of procedural instructions for automobile repair tasks. These semantic categories can help writers choose appropriate verbs when authoring documents.

<sup>1</sup> <https://wordnet.princeton.edu>



Level 1	Level 2	Example	Type		Token	
			#	%	#	%
Action	Part	<i>toritsukeru</i> (install), <i>kirihanasu</i> (disconnect)	343	37.7	395,890	37.40
	Software	<i>hyojisuru</i> (display), <i>kiokusuru</i> (store)	54	5.9	97,944	9.25
	Diagnosis	<i>tenkensuru</i> (check), <i>kanshisuru</i> (monitor)	23	2.5	96,457	9.11
	General	<i>shiyousuru</i> (use), <i>sousasuru</i> (operate)	91	10.0	50,436	4.77
	Auxiliary	<i>suru</i> (do, make), <i>okonau</i> (perform)	10	1.1	189,140	17.87
Stative	Existence	<i>aru/iru</i> (be, exist), <i>ichisuru</i> (be located)	11	1.2	41,544	3.93
	Composition	<i>kouseisuru</i> (compose), <i>yuusuru</i> (have)	21	2.3	10,855	1.03
	Denotation	<i>shimesu</i> (indicate), <i>arawasu</i> (show, denote)	18	2.0	8,606	0.81
	Relation	<i>kankeisuru</i> (be related), <i>kotonaru</i> (differ)	28	3.1	2,439	0.23
	Function	<i>kinousuru</i> (function), <i>eikyosuru</i> (affect)	31	3.4	5,084	0.48
	State	<i>taikisuru</i> (wait), <i>nokoru</i> (remain)	25	2.7	18,931	1.79
Change	Auxiliary	<i>hajimeru</i> (start), <i>keizokusuru</i> (continue)	29	3.2	8,195	0.77
	State	<i>modosu</i> (return), <i>hasseisuru</i> (occur, generate)	65	7.1	7,305	0.69
	Quantity	<i>atatameru</i> (warm up), <i>joshosuru</i> (increase, rise)	34	3.7	3,967	0.37
Communication	General	<i>henkasuru</i> (change), <i>hendousuru</i> (vary)	5	0.5	4,326	0.41
	Exchange	<i>soushinsuru</i> (send), <i>tsuuchisuru</i> (inform)	36	4.0	6,869	0.65
	Record	<i>kirokusuru</i> (record), <i>hozonsuru</i> (save, store)	14	1.5	5,764	0.54
Cognition	Performance	<i>kinshisuru</i> (prohibit), <i>shitagau</i> (follow)	22	2.4	2,717	0.26
		<i>chuiisuru</i> (pay attention to), <i>handansuru</i> (judge)	44	4.8	101,541	9.59
Perception		<i>miru</i> (see), <i>kanjiru</i> (feel), <i>kiku</i> (hear)	6	0.7	414	0.04

Table 3: Typology of the semantic categories of verbs with examples and frequency information.

Rank	Verb	Frequency		Cumulative Frequency (Coverage)	
		#	%	#	%
1	<i>suru</i> (do, make)	94,613	8.94	94,613	8.94
2	<i>okonau</i> (perform)	93,750	8.86	188,363	17.80
3	<i>kakuninsuru</i> (confirm)	87,099	8.23	275,462	26.03
4	<i>torihazusu</i> (remove)	70,011	6.61	345,473	32.64
5	<i>kirihanasu</i> (disconnect)	65,451	6.18	410,924	38.82
6	<i>toritsukeru</i> (install)	64,508	6.09	475,432	44.92
7	<i>setsuzokusuru</i> (connect)	55,546	5.25	530,978	50.17
8	<i>sokuteisuru</i> (measure)	50,703	4.79	581,681	54.96
9	<i>tenkensuru</i> (check)	44,254	4.18	625,935	59.14
10	<i>aru</i> (be)	40,728	3.85	666,663	62.99
11	<i>koukansuru</i> (replace)	23,614	2.23	690,277	65.22
12	<i>sentakusuru</i> (select)	19,632	1.85	709,909	67.07
13	<i>taikisuru</i> (wait)	18,448	1.74	728,357	68.82
14	<i>shoukyosuru</i> (clear)	16,736	1.58	745,093	70.40
15	<i>shutsuryokusuru</i> (output)	14,171	1.34	759,264	71.74
16	<i>hyojisuru</i> (be displayed)	10,499	0.99	769,763	72.73
17	<i>yomu</i> (read)	10,004	0.95	779,767	73.67
18	<i>shiyousuru</i> (use)	8,989	0.85	788,756	74.52
19	<i>sadousuru</i> (operate)	7,301	0.69	796,057	75.21
20	<i>naru</i> (become)	7,003	0.66	803,060	75.87

Table 4: The 20 most frequent controlled verbs that occurred in our data set.

### 3.3 Coverage of Verbs

Currently, the number of approved verb types in our lexicon is 822. However, from the practical point of view, it is still too many and writers—even professional technical writers—may find it difficult to appropriately make use of the controlled lexicon. Here, it is more valuable to define the core set of verbs, or further reduce the number of verb entries. To set the goal of the lexicon refinement process, in this section, we investigate the coverage of verbs and estimate how many verbs are necessary for authoring automotive technical documents.

Table 4 presents the 20 most frequent controlled verbs that occurred in our data set with their individual frequencies and cumulative frequencies, which can be regarded as coverage. It is worth noting that only seven verbs cover half of the verb use in the automotive repair manuals and 20 verbs three quarters. Figure 2 illustrates the growth curve of coverage as the



number of verbs increases in the descending order of frequency. The curve grows rapidly by around 100, then tapers off and, around 400, is flattened out. From this observation, we can recognise that approximately 100 verb types are particularly important and can be considered a core set of verbs, and 300–400 verb types will suffice for our controlled lexicon. Further refinement of the lexicon will be part of future research.

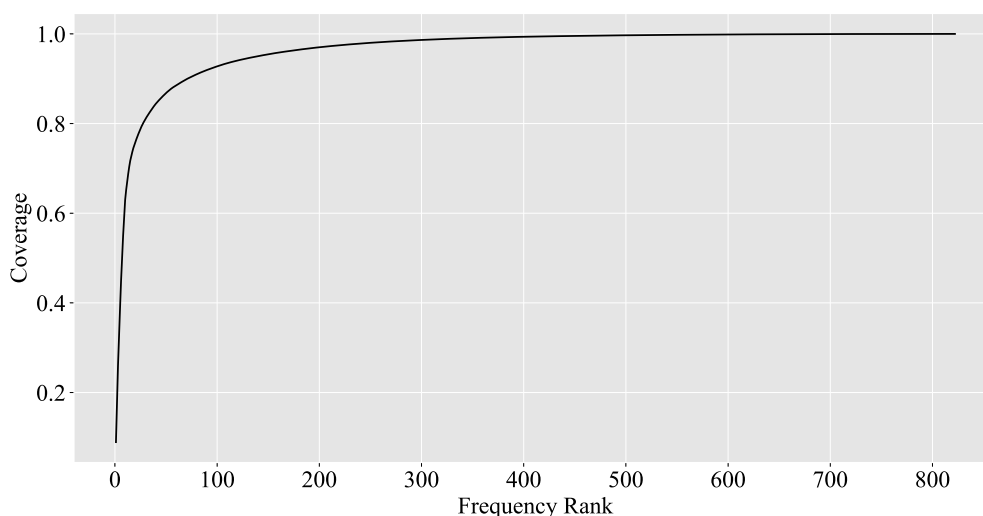


Figure 2. Growth curve of the coverage of the controlled verb lexicon.

## 4 Formulation of Canonical Case Order

### 4.1 Procedure

To further increase the utility of our controlled lexicon, we formulated the canonical case orders for all the approved verbs in the following two processes.

1. For each verb, we abstracted the case structure of sentences (main clauses) using Japanese sentence parser KNP with JUMAN++V2 (see also Figure 3). For each sentence, a verb at the end of the sentence, which is a predicate of the main clause in Japanese, was identified. Subsequently, noun phrases with postpositional particles, such as *ga*, *o*, *ni* and *de*, that directly attach to the verb were extracted as cases, or arguments, for the verb. Only the particles were reserved to form an abstract case pattern. At this stage, supplementary adverbial phrases were omitted because they were usually irrelevant to the core structure of sentences.
2. We selected the preferred case orders based on the frequency of usage and combined them to create canonical orders that could cover frequent types of the case structures for the verb (see also Figure 4). It should be noted that there were many less frequent types that are not covered by the defined case orders, which we will discuss in the next section.

After conducting these processes for all the 822 approved verbs, we finally formulated 954 canonical case orders. Some of the verbs have multiple canonical case structures; for example, the verb *hyojisuru* (display) has two structures: [*~o*] [*~ni*] *hyoji-saseru* and [*~ni*] [*~ga*] *hyoji-sareru*. Here are some examples of such sentences:

(5) ダイアグノーシスコード確認画面を表示させる。/*Daiagunoshisu-kodo kakunin gamen o hyoji-saseru*.

(Display the diagnosis code check screen.)

(6) 画面にダイアグノーシスコードが表示される。/*Gamen ni daiagunoshisu-kodo ga hyoji-sareru*.

(The diagnosis code will be displayed on screen.)

The verb *hyojisuru* can be used both in causative form *hyoji-saseru* for human actions and in passive form *hyoji-sareru* for machinery functions, and these two forms have different case structures. The detailed description of the usage of the different verb forms and case structures will be an important future task.

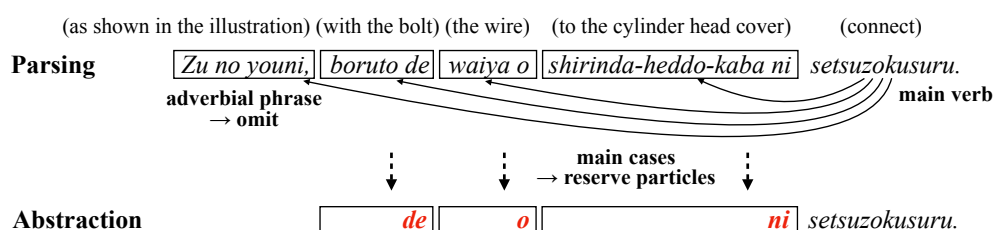


Figure 3. Automatic process to abstract the case structure of a Japanese sentence.



Frequency	Abstract case order	Combine	Canonical case order
3466	[~ o] <i>setsuzokusuru</i>	variation	[~ de] [~ o] [~ ni] <i>setsuzokusuru</i>
1934	[~ o] [~ ni] <i>setsuzokusuru</i>		
215	[~ ni] [~ o] <i>setsuzokusuru</i>		
172	[~ ni] <i>setsuzokusuru</i>		
61	[~ de] [~ o] <i>setsuzokusuru</i>		
⋮	⋮	not covered	⋮
22	[~ de] <i>setsuzokusuru</i>		
18	[~ niwa] [~ ga] <i>setsuzokusuru</i>		
⋮	⋮	not covered	⋮
1	[~ ga] [~ ni] <i>setsuzokusuru</i>		

Figure 4. Examples of the formulation of canonical case order: *setsuzokusuru* (connect).

## 4.2 Coverage of Defined Case Structure

As mentioned above, it is difficult to comprehensively cover all the case order patterns, although we assume that the formulated canonical case structures covered a substantial portion of case patterns. Here, we calculate the coverage of the 954 canonical patterns using the same data set of the automobile repair manuals. For each sentence, we abstracted the case pattern in the same manner described in Figure 3. If the abstracted case pattern contained the same set or a subset of case elements defined in the canonical structure, we regarded it as the covered pattern. In addition, if the order of the case elements violates the canonical order, we considered it as a variation.

Table 5 shows the results; 85.61% of the pattern tokens were covered by our formulated patterns, which demonstrated the fairly high coverage. However, the coverage of case pattern types is low. It indicates that a large number of rare types of case patterns have not been captured by the current set of patterns. The coverage needs to be increased by defining other types of canonical orders.

To understand the relationship between the number of case structure types and coverage of tokens, we observe the growth curve in Figure 5. This figure illustrates how coverage increases as case pattern types are included in the descending order of frequency. We can see that the 2,000 most frequent types can cover almost all tokens in our data set. The formulated canonical case patterns already covered 1,807 types, while they do not necessarily include the frequent ones. We assume we can soon attain higher coverage, namely more than 90%, by adding frequent patterns.

	All	Covered		Variation	
	#	#	%	#	%
Type	5,363	1,807	33.69	170	9.41
Token	1,058,424	906,134	85.61	24,366	2.69

Table 5: Coverage of the set of canonical case structures and ratio of variation in our data set.

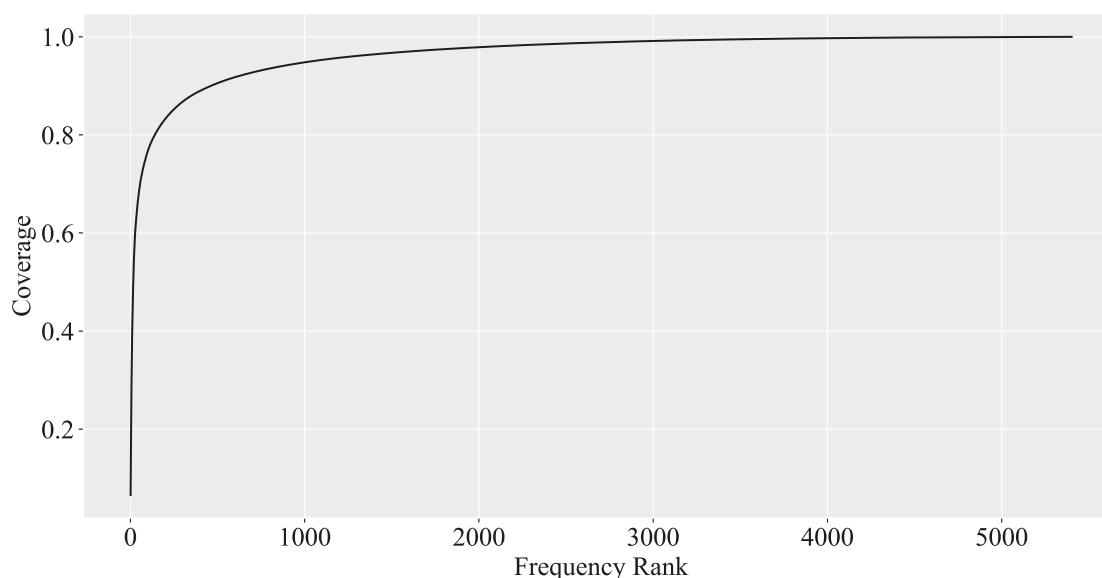


Figure 5. Growth curve of coverage of case structure patterns in the original forms.



### 4.3 Case Structure Variation

Finally, we investigate the issue of structural variations in our data set. As Table 5 shows, 2.7% of the covered pattern tokens were found to be variations. Although the ratio is not high, given that the automobile manuals were reasonably controlled by professional writers, there is still room for improvement in terms of consistent use of sentence structures. An example of case structure variations for the verb *tofusuru* (apply, coat) is shown below:

- (7) 新品の O リングに コンプレッサオイルを塗布する。 / *Shinpin no O-ring ni conpuressa oiru o tofusuru*.  
(Apply compressor oil to a new O-ring.)
- (8) グリースを SST のボルトに塗布する。 / *Gurisu o SST no boruto ni tofusuru*.  
(Apply grease to the SST bolts.)

Example (7) conforms to the defined canonical pattern, [*~ de*] [*~ ni*] [*~ o*] *tofusuru*, and Example (8) is regarded as a variation since the order of the [*~ o*] case and [*~ ni*] case is the reverse of the canonical order. In this example, we can modify (8) into (9) as shown below without changing its meaning.

- (9) SST のボルトに グリースを塗布する。 / *SST no boruto ni gurisu o tofusuru*.  
(Apply grease to the SST bolts.)

Although (8) is grammatically correct and the core meaning of the sentence remains unchanged, it is important to control these structural variations for consistency, which will eventually lead to high usability of the text. Importantly, if we define canonical patterns in advance, these variations can be automatically detected in combination with sentence analysis tools, which we will discuss in the next section.

## 5 Towards Authoring Support

The constructed controlled lexicon of verbs with the definitions of canonical case orders is a basis for controlled authoring; it helps writers recognise which verb should be used in what sentence structure. To be more effective, we will further discuss the mechanisms for supporting the controlled authoring process of writers. Authoring support scenarios can be broadly divided into two types: *post hoc revision* and *writing from scratch*. Correspondingly, we propose a sentence diagnostic tool for revision and a template-driven writing tool based on the canonical case patterns. In the following two sections, we outline them respectively.

### 5.1 Diagnostic Tool

The constructed lexicon can be used to control the two types of variations, that is, the use of unapproved words and non-canonical sentence structures. We propose a tool to support the diagnosis and revision of both types of variations in three processes: *detect*, *suggest* and *rewrite*.

With regard to unapproved words, the three processes can be simply implemented if the synsets of the unapproved and approved words are defined (Warburton 2014). The tool first searches the input sentence for unapproved words referring to the lexicon and, if any unapproved word is discovered, it retrieves the corresponding approved word. If the suggestion is adopted, the unapproved word in the input sentence is automatically corrected.

Figure 6 depicts the process of detection and suggestion of unapproved words using our controlled lexicon checker. The following example is used as an input:

- (10) 新品の乾いた布で異物を除去する。 / *Shinpin no kawaita nuno nado de ibutsu o jyokyosuru*.  
(Eliminate foreign matter with a new dry cloth.)

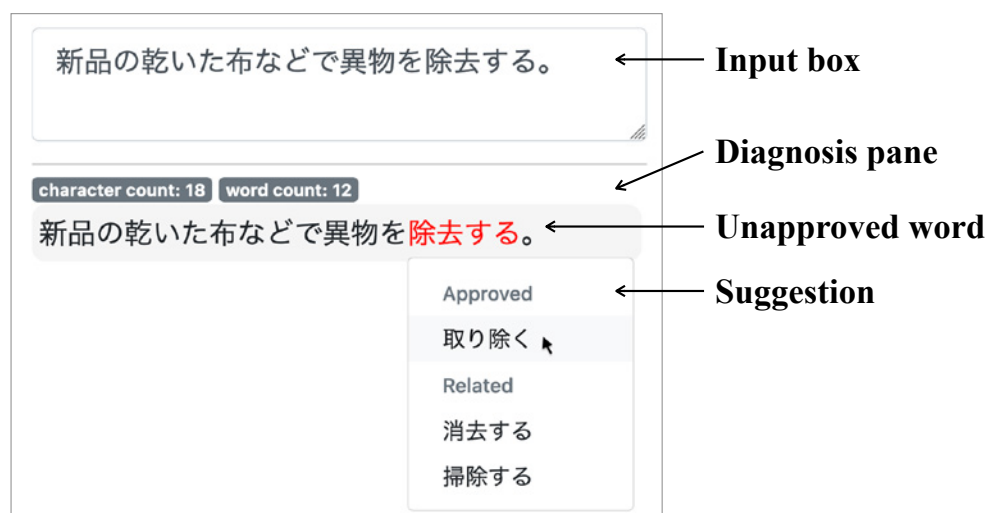


Figure 6. Prototype interface of the controlled lexicon checker.



The unapproved word *jokyosuru* (eliminate) is highlighted in red and the approved word *torinozoku* (remove) is recommended. To further support human decision-making, this tool also provides semantically-related words *shoukyosuru* (delete) and *soujisuru* (clean), which might be more suitable in a certain context. Writers can select any of the candidates to replace the unapproved word.

Conversely, case structure variations are more difficult to handle automatically. The first bottleneck is the parsing of the input sentence to abstract its case structure. Although the high-performance Japanese parsers, such as KNP (Kawahara & Kurohashi 2006) and CaboCha (Kudo & Matsumoto 2002), are available, the accuracy of parsing of long, complex sentences is still not sufficient for this task. Another difficulty is the ambiguity of suggestion. The canonical case patterns we defined do not specify the order of supplementary elements such as adverbial phrases. Even if a sentence is correctly parsed and the violated case order is detected, there may be multiple suggestions for rewriting. While the final decision making will be left to human writers, we need to further elaborate rules to place elements of sentence in proper positions.

## 5.2 Template-driven Writing Tool

As a more preemptive solution for possible violations to the controlled lexicon, we also plan to develop a template-driven writing tool. Here, we explain the design principle of the tool and challenges for its development.

The basic flow of template-driven writing is shown in Figure 7. To write instructional sentences, verbs are the most important elements that govern the core meanings and sentence structures. Thus, writers first select a main verb from the controlled lexicon of verbs. At this stage, the tool helps them to select an appropriate verb by displaying the hierarchic structure of semantic categories of verbs defined in Section 3.2. Once the verb is fixed, registered sentence templates, i.e. canonical case patterns for the verb, will be presented and writers can choose one of them. However, the default sentence template may not be sufficient for accommodating necessary information. Hence, writers can further add supplementary sentence components with slots, such as an adverbial component [*~ youni*] (as ~), from a list of possible components provided by the tool. In conjunction with populating the template with additional components, writers fill in the slots with content words such as *boruto* (the bolt), *waiya* (the wire) and *shirinda-heddo-kaba* (the cylinder head cover) to complete the sentence. At this stage, the tool provides two functions: (1) suggestion of the probable candidate words and technical terms and (2) validation of the conformity of content words to the controlled lexicon of nouns and adjectives.

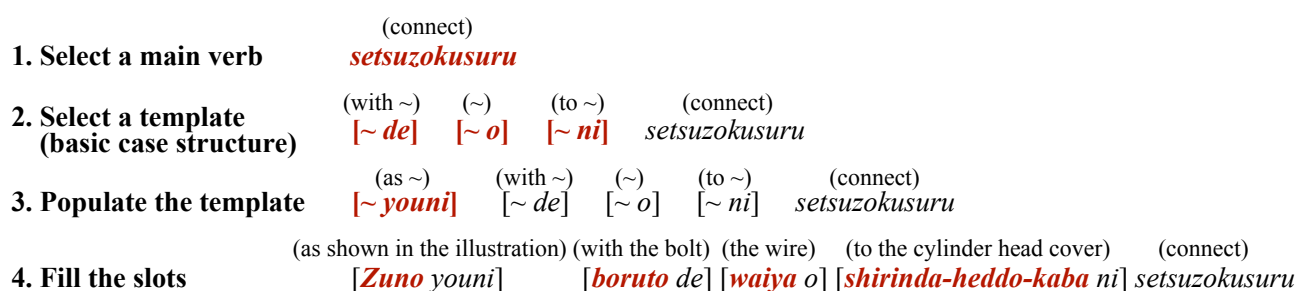


Figure 7. Basic flow of template-driven writing with a simple sentence as an example.

To implement the functions described above, the following items are specifically necessary:

1. Definition of supplementary sentence components (specifically, adverbial phrases), besides basic case patterns.
2. Construction of controlled lexicon of content words (specifically, nouns and adjectives), besides verbs.
3. Construction of controlled terminology of automotive domain (specifically, part and tool names).

Furthermore, to fully implement the template-driven writing tool, we need to tackle the challenges of constructing compound/complex sentences by combining multiple simple sentences. To understand how sentences are constructed in our data set, focusing on coordinate and adverbial clauses, we first automatically extracted compound/complex sentences and discovered that 525,966 of the 1,058,424 sentences are compound/complex. Sentence (11) is an example of a compound sentence, with two coordinate independent clauses, and Sentence (12) is an example of a complex sentence, with one main clause and one subordinate clause.

(11) エンジンを始動し、アイドリング状態にする。 /

*Enjin o sidou-shi, aidoringu joutai ni suru.*

(Start the engine **and** keep the engine idling.)

(12) トラブルシュートを実施する前に、この回路のヒューズの点検をすること。 /

*Toraburu-shuto o jisshisuru maeni, kono kairo no hyuzu no tenken o suru-koto.*

(Inspect the fuses for circuits **before** performing the troubleshooting.)

As shown in bold in the examples above, certain connective expressions are used to construct compound/complex sentences. We categorised surface connectives automatically extracted from 525,966 compound/complex sentences. Table 6 shows the results of the categorisation with frequency for each category. The total number of connectives is 739,651, which means that many of the compound/complex sentences have more than two clauses. The dominant category is the resultative coordination such as (11), occupying nearly 60%. We also notice that adverbial clauses to



express timing and conditions of events ('when', 'in case' and 'if') appear frequently in the data set, which indicates that conditional branching of tasks are crucial building blocks for authoring instructional documents. Based on the results, we plan to implement functions to support the construction of compound/complex sentences.

Finally, from the viewpoint of controlled authoring, we should emphasise that various surface connectives are used to signify almost the same meaning. For example, to mean 'when' in adverbial clauses, various connectives are used, such as *toki*, *tokini*, *tokiwa*, *tokiniwa*, *sai*, *saini*, *saiwa* and *sainiwa*. In many cases, these connectives are interchangeable. Therefore, it is effective to define the approved usage of connectives to further control the sentence structural variations.

Level 1	Level 2	Level 3	Surface connectives	#	%
coordinate	resultative	and	V (continuative form), V- <i>te</i>	421,324	56.96
		such as	<i>tari</i>	6,389	0.86
	contradictory	but	<i>ga</i>	4,563	0.62
adverbial	time	when	<i>toki (ni/wa/niwa)</i> , <i>sai (ni/wa/niwa)</i>	39,051	5.28
		each time	<i>tabi (ni)</i>	101	0.01
		before	<i>mae (ni)</i>	17,946	2.43
		after	<i>nochi (ni)</i> , <i>ato (ni/de)</i>	12,305	1.66
		then	<i>ue (de)</i>	1,526	0.21
		until	<i>made</i>	10,805	1.46
	condition	in case	<i>baai (ni/ha/niwa)</i>	90,773	12.27
		if	<i>to</i> , <i>nara</i> , <i>ba</i>	37,373	5.05
		only if	<i>dakedemo</i>	210	0.03
		though	<i>mo</i>	6,491	0.88
		as long as	<i>kagiri</i>	102	0.01
	method	by	<i>kotode (mo/niyori)</i>	9,198	1.24
	attendant circumstances	while, with	<i>nagara</i> , <i>mama</i> , <i>tsutsu</i>	9,426	1.27
		without	<i>zu ni</i> , <i>nai de</i>	3,394	0.46
	state	in the state	<i>jotai de</i>	12,273	1.66
		in the way	<i>youni</i>	11,013	1.49
	purpose	in order that	<i>tame</i>	27,325	3.69
	reason	because	<i>tame (ni)</i> , <i>node</i> , <i>kekka</i> , <i>kara</i>	17,779	2.4
	contradictory	but	<i>noni</i>	270	0.04
	extent	to the extent	<i>hodo</i>	14	0.00

Table 6: Compound/complex sentence patterns and surface connectives observed in our data set.

## 6 Conclusion

In this study, we have built the controlled lexicon of verbs that is useful for consistent authoring of automotive technical documents. The lexicon building proceeded in both descriptive and prescriptive manners: we first descriptively observed a huge volume of existing text data, and then prescriptively defined approved words and their canonical usage. Although we dealt with Japanese verbs as a starting point, this lexicon building process is applicable to other lexical units and languages.

Currently, the constructed lexicon consists of 822 approved and 88 unapproved verbs, which comprehensively cover the analysed data set containing more than one million verb tokens. The detailed analysis of coverage revealed that we can reduce the size of the lexicon to 300–400 words with little loss of coverage. The significant feature of our lexicon is the definition of canonical case orders for each approved verb, which helps writers compose sentences in consistent structures. We have defined 954 canonical case patterns that are estimated to cover 85% of the existing sentences.

We have also proposed authoring support tools that employ controlled lexicon. For two different scenarios, that is, post hoc revision and writing from scratch, we designed a sentence diagnostic tool and a template-driven writing tool, respectively. These tools are designed to assist writers in using appropriate words in accordance with their controlled usage. We also discussed necessary components and technologies to implement these tools.

In future research, we will refine the constructed lexicon and extend it to cover nouns, adjectives and adverbs. We also plan to build an English lexicon and link the approved words between Japanese and English, which enables consistent translation by both human translators and machine translation systems. In particular, we will examine the effectiveness of controlled bilingual lexicon for machine translation. We assume that the use of controlled lexicon can facilitate the reduction of vocabulary size of the text, which may have a positive impact on machine translation, including recent neural models. Finally, the implementation and evaluation of the authoring support tools we designed is an important practical goal of this research project, which we will address in future studies.

## 7 References

ASD. (2017). ASD Simplified Technical English. Specification ASD-STE100, Issue 7. <http://www.asd-ste100.org> [08/07/2020].



- Ciaramita, M. & Altun, Y. (2006). Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 594–602.
- Fillmore, C. J. (1968). The Case for Case. In Bach, E. & Harms, R. T. (eds.) *Universals of Linguistic Theory*, 1–88. New York: Holt, Rinehart and Winston.
- Godden, K. (2000). The Evolution of CASL Controlled Authoring at General Motors. *Proceedings of the 3rd International Workshop on Controlled Language Applications*, Seattle, WA, USA, 14–19.
- Kawahara, D. & Kurohashi, S. (2006). A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, USA, 176–183.
- Kudo, T. & Matsumoto, Y. (2002). Japanese Dependency Analysis using Cascaded Chunking. *Proceedings of the 6th Conference on Natural Language Learning*, Stroudsburg, Pennsylvania, USA, 63–69.
- Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1): 121–170.
- Masuoka, T. & Takubo, Y. (1992). *Kiso Nihongo bunpo [Basic Japanese Grammar]*. Tokyo: Kuroshio Shuppan.
- Means, M. & Godden, K. (1996). The Controlled Automotive Service Language (CASL) Project. *Proceedings of the 1st International Workshop on Controlled Language Applications*, Belgium, Leuven, 106–114.
- Morita, H., Kawahara, D. & Kurohashi, S. (2015). Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2292–2297.
- Møller, M. H. & Christoffersen, E. (2006). Building a Controlled Language Lexicon for Danish. *LSP & Professional Communication*, 6(1): 26–37.
- Nyberg, E., Mitamura, T. & Huijsen, W. O. (2003). Controlled Language for Authoring and Translation. In Somers, H. (ed.) *Computers and Translation: A Translator's Guide*, 245–281, Amsterdam: John Benjamins.
- Paaß, G. & Reichartz, F. (2009). Exploiting Semantic Constraints for Estimating Supersenses with CRFs. *Proceedings of the SIAM International Conference on Data Mining*, Sparks, Nevada, USA, 485–496.
- Sasano, R. & Okumura, M. (2016). A Corpus-Based Analysis of Canonical Word Order of Japanese Double Object Constructions. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2236–2244.
- Tolmachev, A., Kawahara, D. & Kurohashi, S. (2018). Juman++: A Morphological Analysis Toolkit for Scriptio Continua. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, System Demonstrations*, Brussels, Belgium, 54–59.
- Warburton, K. (2014). Developing Lexical Resources for Controlled Authoring Purposes. *Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use*, Reykjavik, Iceland, 90–103.

### Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19K20628 and 19H05660, and by the Naito Research Grant, Japan. The automobile manuals used in this study were provided by Toyota Motor Corporation.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Bi- and Multilingual Lexicography**







# Reconceptualizing Lexicography: The Broad Understanding

Leroyer P.<sup>1</sup>, Køhler Simonsen H.<sup>2</sup>

<sup>1</sup> Aarhus University, Denmark

<sup>2</sup> Copenhagen Business School, Denmark

## Abstract

Lexicography has changed radically over the past 20 years and numerous scholars have discussed, in a vast number of theoretical contributions, whether lexicography is a science of its own and how it should be defined. But has the whole idea of lexicography, the way we see it in the first place, also changed? Is lexicography all about dictionaries one way or another? Can it be understood differently? In this light, the purpose of our paper is to propose a broad understanding spurred by the closing remark in Adamska-Salaciak's article on lexicography and theory as follows 'theoretical lexicography in its present form is unlikely to offer any such theoretical perspective', cf. (Adamska-Salaciak 2018:14). We do not wish to continue the somewhat tautological discussion of whether lexicography is a science or not, and bring in yet another definition. Instead, we intend to take Adamska-Salaciak up on her call for further theory development and introduce a reconceptualization of lexicography founded on a social-constructivist position paving the way to a broad understanding. In our discussion, we have drawn on established, seminal lexicographic theory, but reconceptualization requires a break with current views. Consequently, we have also drawn on theories discussed in Simonsen (2012), Christensen (2017), Fadel et al. (2015), Leroyer and Simonsen (2018a, 2018b), Liew (2013), Osterwalder and Pigneur (2010), Osterwalder et al. (2014), Weill and Woerner (2018), etc. Elaborating on the model of Verlinde et al. (2010) and Simonsen (2012), we explain how what we call 'lexicographic meaning-construction processes' are at the heart of lexicography. In this light, we present a seven-faced model showing how current and novel elements of lexicographic theory interplay and can be reinterpreted.

**Keywords:** social-constructivist position; reconceptualization; lexicographic meaning-construction processes; seven-faced model

## 1 Grounding the field: towards a broad understanding

The questions of whether or not lexicography is a science and what constitutes lexicography have been extensively discussed in a vast number of theoretical contributions over the years, such as for example Scerba (1940), Zgusta (1971), Wiegand (1984), Hausmann (1985), Tarp (2008), Bogaards (2010), Verlinde et al. (2010), Tono (2010), Bergenholtz and Gouws (2012), Rundell (2012), Piotrowski (2013), Bergenholtz and Agerbo (2017), Tarp (2018, 2019), Adamska-Salaciak (2018) and Margalitadze (2018), to name but a few.

Being a polysemous term, i.e. a complex, conceptual unit of meaning and understanding (Temmerman 2000) in an ever-changing environment, the term lexicography can be understood, not only as a product of our mind, i.e. a conceptual blend, but also as a series of continuing conceptualization processes framed by competing views on science (c.f. the above mentioned contributions). Yet, terms tend to be established at the terminal stage of processes, and lexicography is then defined into competing terms. This entails competing definitions drawing up distinctive conceptual limits and borders (*definire*), and attempting to keep things apart and make sense. As a result, definitions bring processes to an end, so meaning-construction comes into terms with concepts so to speak. Meaning freezes. In a field of expert knowledge as lexicography, in which definitions play a predominant role, defining not surprisingly seems to be one of the main issues so far in the realm of theory making addressing the seminal questions: What is lexicography? How should it be understood?

Therefore, attempting to redefine lexicography yet another time is not our purpose here. Rather, we intend to move borders and contribute to further theory development through a reconceptualization of the understanding of lexicography. In our attempt to do that, we base our work on a realization that we need a respectful, but necessary, detachment from past and present considerations. At the same time, we base it on a social-constructivist position to the ontology and epistemology of lexicography. A short voyage through the past and present stages of lexicographic (r)evolution will illustrate our position. The fact that lexicography has been subject to a number of revolutions over the centuries since the use of ancient clay tablets and roll and codex seems to be the rule rather than the exception. Static dictionaries of the print era have extensively given way to Internet dictionaries and lexicographic cyberspace. Trap-Jensen (2018) speaks of three revolutions: the descriptive, the corpus, and the digital revolution. The first two relate to dictionary making and are seen from the position of the lexicographer. The third goes further, as it includes not only new opportunities of dictionary making for human users, but also NLP and AI applications in which lexicography is turned into a language data provider of so-called machine users. We do not consider the first two revolutions as full-fledged revolutions proper. Both of them are the result of changes of an approach in language description, with the emergence of the big empirically based national dictionaries, moving from a prescriptive to a descriptive perspective, on the one hand, and with the use of corpus methodologies to improve language description, on the other. One should remember, however, that the prescriptive approach is still valid, as it is the cornerstone of terminographic description and standardization bodies. As far as the corpus revolution is concerned, we agree with Trap-Jensen. It has improved the quality of language description, but has not led to radical changes and has remained unnoticed by non-lexicographers. The digital revolution (which incidentally is closely connected to the corpus revolution) is a true,



ongoing revolution, leading to metamorphoses not only in dictionary making processes and dictionary forms, but also in dictionary use and in the general status of lexicography. Fuertes Olivera (2016) speaks of it as a “Cambrian Explosion”.

This gets us to the point where lexicography in existing theory is conceptualized as a discipline on a continuum between linguistics and IT – the mirroring of language through systematic language descriptions in dictionaries – including descriptions designed to become input in NLP and AI applications. Admittedly, the functional theory of lexicography (Tarp 2008, 2019) pushes lexicography further as an idealist, user-oriented science of its own, although with a lot in common with information science. Still, its tenets depart from objective views on dictionaries as objects of use responding to needs (objectivism), and of critical, empirical views on their use (phenomenalism), i.e. a post-positivist position.

This is what we choose to call the narrow, focalized understanding of lexicography. It is naturally prompted by lexicographers and their organisations doing dictionary work in theory and practice. They pursue comprehensive, objective language descriptions in order to mirror language in dictionaries, automated or not, that are designed to be used, whereby dictionaries are turned into language tools. However, the tool-driven conceptualization does not discern that lexicography is far from being limited to the making and use of lexicographic tools. Lexicography also includes lexicographic and metalexicographic processes whenever we take part in discussions and activities on and about language in meaning-construction departing from linguistic signs (lexical items), whether isolated or in combinations, or other semiotic signs.

This constructivist principle is *de facto* integrated today in online dictionary platforms, but not recognized as such. It is quite common for platforms today to include word-based services such as word games, critical debates and discussions about words, their definitions and usage, their history, chats, numerous questions and answers concerning words, word-centered events (word of the day, word of the year), suggestions for inclusion or creation of new words and new meanings. The list is far from being exhaustive. Common to all these is the critical study, construction, interpretation and sharing of meaning mainly triggered by lexical items, a social behaviour involving a vast number of actors, not simply users. The reason is that meaning really matters to us, and that interrogating and negotiating meaning makes sense. It is not only, as in the communicative model of (Jakobson 1960), a mere reflexive function of language in which the code becomes the object of the message. It concerns the making of the code itself, and our reflexive relationship to the code.

From now on we will speak of ‘Lexicographic Meaning-Construction Processes’, or in short LMCPs. LMCPs are reflexive processes of meaning construction, interpretation, structuration and representation. They are extensively triggered by – and applied to – units of meaning and understanding from our lexicon of semiotic signs, and fostered by the idea of getting to mutual understanding.

One might object that such processes are exactly revolving around the making, use and management of dictionary platforms and that lexicography therefore must be determined by dictionaries. Yet, meaning-construction processes are constantly taking place in a huge variety of social contexts other than dictionary making or use. They occur in politics, companies and organisations (henceforth C&Os), workplaces, language-planning bodies, institutions, educational contexts, at home, etc., whenever we take actively part in lexicographic processes as we discuss, create and record the meaning of lexical items, names, designations and concepts which are at the core of our private and professional life. This is not a matter of language mirroring but of language making, i.e. meaning-construction as part of the general processes of semiotic sense-making in our social life, much in line with the processes described by Weick (2009). Consequently, lexicographic processes are deeply embedded in the semiotic production and management of signs in a social and organisational environment. A great deal of lexicographic processes is certainly determined by the making and use of dictionaries for humans or data sets for machines, whereby processes come to an end and freeze, but most of these are not, and keep moving. The purpose of lexicography is not only to make lexicographic tools. It is also to support, through meaning-construction processes, cognitive sense-making mechanisms, individually as well as in organisations and social groups. At a higher level, the purpose is to frame semiotic meta-perception and construction of what we sense and call reality through language.

Many perspectives have changed, just as new ones appear. One perspective is the epistemology of lexicography. As explained above, lexicography goes far beyond tools of language description. It also includes the ongoing processes of meaning-construction by lexicographers AND by all of us as both users and actors, as we all engage actively in meaning-construction processes. Thus, we suggest a new word: ‘user-actors’. As notable as it is, dictionary making and use is only one of the user-actor roles we engage in. The overall lexicographic landscape is much broader.

Another perspective concerns the multimodal nature of the objects of lexicography. These have changed immensely. Dictionaries used to be stand-alone products, but are now autonomous and integrated in users’ tools and devices such as mobile phones and office software. Dictionaries or lexicographic products are ubiquitous and take on new hidden forms for us. They have changed from being full, self-contained objects to becoming partial, autonomous objects. More and more publishing houses sell lexicographic components to software producers and game producers etc. Lexicography has been fragmented and diversified, and new meaning-construction processes are at work because of these metamorphoses.

A third important perspective is the activity of lexicography. Dictionaries used to be a one-user experience. Now, lexicographic meaning-construction processes have turned into shared, collaborative activities, also on social media. They play a crucial role in language planning in many countries and regions of the world, and are guided by the co-construction and preservation of regional or national identity and culture. They are also at the heart of language policies and organizational communication in C&Os, and support corporate language management and language engineering systems. Finally, the users of lexicography deserve another role. Since Wiegand (1988: 778), lexicography was phenomenologically conceptualized as an action involving the meeting of user-determined use-objects having a “Genuiner Zweck” with “bekannte Unbekannte” users. Following the toolification phenomenology, lexicography has been developed into a system of individually, user-needs adapted lexicographic functions (Tarp 2008) and user profiles. Yet, dictionaries and lexicographic processes are also collectively determined. They support for instance the missions of C&Os and involve a great number of us as user-actors, not simply users. We will get back to this in the following sections.

The main thesis upon which we base our work is not simply that the subject matter, object, activity, actors, and orientation



of lexicography have changed considerably and that we should re-think them, but that changes have disclosed a theoretical view on lexicography that hitherto has remained under-represented. Therefore, time is ripe for a broader understanding.

## 2 Epistemological objectives and methodology

The epistemological objectives of this paper are to discuss and challenge current conceptual constructions of the term lexicography and to reconceptualize it from a social-constructivist position. In this light, the theory is applied to explain how present and new elements of lexicographic theory making can be understood in a broad sense. The methodology we used is founded on a critical literature analysis as described for example by Saunders et al. (2019). Prior to the analysis, we formulated a number of views on science and epistemological themes in the field that we looked for in the literature using the thematic analysis grid approach (Saunders et al. 2019:122-124). Firstly, we will briefly introduce relevant theoretical contributions and current conceptualizations of the term lexicography. Secondly, we will propose an expanded conceptual interpretation of the meaning of the term lexicography based on the key concept of LMCPs. We will then apply it in a seven-faced model showing the interplay between LMCPs and each of the faces.

## 3 Dictionary-centered structural and functional conceptualizations

Prevailing theoretical models of lexicography are based on structural (Wiegand 1988) or functional conceptualizations (Tarp 2008). They make the dictionary as the paragon research object of lexicography and place the relationship between dictionary, individual users and usage at the heart of lexicographic studies. Structures and functions frame the understanding of the key concepts of access, data, data selection, adaptation and presentation, and for the presentation of data to fully match the lexicographic information needs of intended users depending on their personal profile and situation in the world. It is worth noting, however, that lexicographers, editors and C&Os are largely absent from theory. So is every one actively taking part in lexicographic debates, discussions, and activities. In addition to that, lexicography normally refers to all of us as users, not as actors. Not much has been said about who we truly are or which role we actively play. With the exception of, to some extent, user studies and studies of the impact of dictionaries on society throughout history, the numerous social aspects of lexicography seem rather underestimated. Still, socialisation is highly visible and can be observed in the proliferation of lexicographically driven and inspired actions in crowd and collaborative lexicography such as for example *Wordnik* or *Thefreedictionary*, as well as debates among us all on the genesis, use and acceptability of new lexical items and meanings.

The recording of new words and meanings may frequently turn into ethical issues and challenges towards norms and policies in society. This shows that lexicography is actively shaping society through meaning construction and interpretation of linguistic signs and non-linguistic signs as units of meaning – images, sounds, icons, etc. C&Os make extensive use of lexicographic processes such as designating, naming, defining, explaining, exemplifying, recording their own vocabulary. These processes are aimed at the creation and management of all specialized words, expressions and other non-verbal signs that make up their “company-speak” (de Vecchi 2013). Company-speak includes brands and product names, and reflects their own way of building up and communicating their own knowledge and express their own culture and identity in the workplace and on the market. The same holds true for organizational communication.

The social context of lexicography, particularly its business dimension, whatever being profit-driven or not, tends also to be underestimated. Simonsen (2002a, 2002b), Leroyer (2007) and Leroyer (2018) are among the few to acknowledge this dimension and speak of “corporate lexicography”. In business, lexicography is more than just gratification, it also serves value-creating or value-adding agendas.

## 4 Understanding lexicography from a social-constructivist position: seven-faced model

So how should lexicography be understood? Lexicography might be referred to as a convergence discipline in line with Verlinde et al. (2010: 3), who describe it “as an information science at the crossroads of three basic perspectives, user, data and access” or as “a science based on six dimensions” (Simonsen 2012). We will refrain from conceptualizing lexicography in an objective manner, as the discipline aiming at the making of lexicographic objects as dictionaries, i.e. descriptions of language, and as a corollary, at the study of their usage and impact in and on society. We use dictionaries to find the correct, appropriate words (and other signs) to express our thoughts and emotions, to represent or acquire knowledge, to communicate (read, write, translate) and to take appropriate actions. However, we do much more than that. We want to find the right words, not simply the correct or appropriate words, but the words that precisely express the meaning we try to convey, the knowledge we keep working on. Language is socially constructed and so is lexicography. Lexicographic work is constantly being performed, words are transformed into right wordings that sound right so real mutual understanding, although utopic, can be achieved. Meaning-construction is the cognitive face of what Wittgenstein describes in his *Tractatus Logico Philosophicus* when he claimed “Dass die Welt meine Welt ist, das zeigt sich darin, dass die Grenzen der Sprache (der Sprache, die allein ich verstehe) die Grenzen meiner Welt bedeuten” (Wittgenstein 1922:145).

Therefore, we will reconceptualize lexicography not in objective and phenomenological terms, i.e. as object-determined activities or actions, but in social-constructivist terms. In line with Fauconnier (1994), we see lexicography as socially determined mental spaces of LMCPs triggered by the interpretation by user-actors of units of meaning and understanding in language (in the Saussurean sense of the word = language as a semiotic faculty). We even go a little further and do not limit meaning construction to natural language. Language consists of linguistic signs and non-linguistic signs as well, such as images, sounds, gestures, icons, etc. Signs are actively created, used and reflected upon every day in our private or



professional life, at home or in the workplace.

LMCPs are cognitive processes. They may – or may not – lead to the construction and use of lexicographic objects in part or totality. Most of the time, they do not. LMCPs are at the center of our reconceptualization and govern the interplays of the seven-faced model below. The model offers answers to the following seven ‘how’ questions:

1. User-actors and LMCPs: How do LMCPs turn us into user-actors?
2. Information and LMCPs: How do LMCPs determine satisfaction of information needs?
3. Mission and LMCPs: How do LMCPs help C&Os realize their communicative mission?
4. System and LMCPs: How do LMCPs form meaningful interfacing in lexicographic systems?
5. Data and LMCPs: How do LMCPs stand for all lexicographic data generation, including polysemiotic data?
6. Access and LMCPs: How do LMCPs rely on the multimodality of human as well as automated access?
7. Business and LMCPs: How do LMCPs bring in true value adding and creation in business?

The model is illustrated in Figure 1 below:

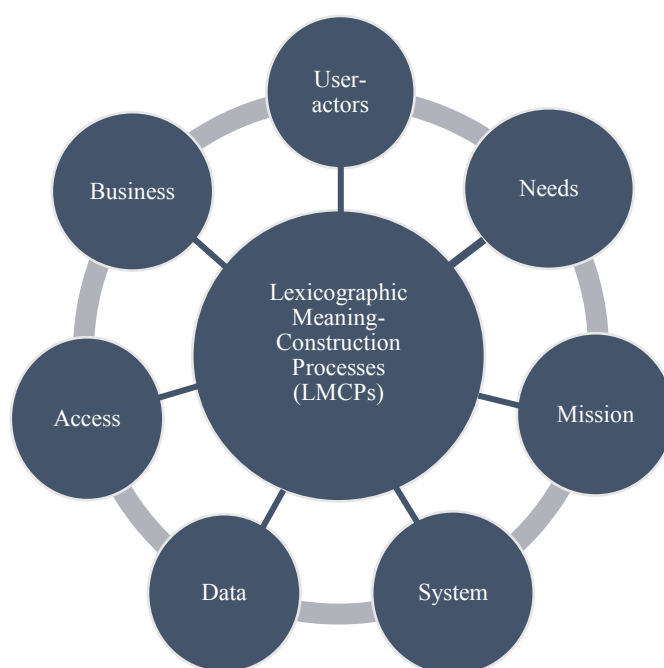


Figure 1: The seven-faced model of the broad understanding of lexicography

#### 4.1 User-actors

A vast number of theoretical contributions on the concept of the user have dealt with the question of who the users are, for example Wiegand (1988) and Tarp (2008). From our position, the term ‘user’ is anchored in a narrow use and toolification perspective, an observation perspective objectively focussing “from the outside” on individual identity, profile and behaviour, whereas social roles and actions are underestimated. Rather than users, we should truly identify ourselves as user-actors. We may have multiple identities and profiles, but are all equally active in LMCPs related to individual or collective situations. User-actors are also involved in C&Os in line with Simonsen (2002:46), as we participate in business-related LMCPs, for instance in connection with sales and branding activities and in efforts to support the mission and vision of the C&O (Simonsen 2002, 2005, 2007a, 2007b; Leroyer 2007, 2013, 2018). The C&O dimension will be discussed further in sections 4.3 and 4.7 below.

User-actors are humans. The idea of non-human machine users replacing us in lexicography in automated systems is questionable. We are not being replaced. Simply, specific LMCPs are involved in the construction of the meta-rules that govern the construction of the algorithms of AI-systems that make them ‘intelligent’, and of NLP applications. As a corollary, critical LMCPs govern our meaning interpretation of machine output.

Furthermore, user-actors are co-producers of lexicographic data and often interact not only with the owner of lexicographic products but also between peers in action and co-action, cf. Leroyer (2016) and Simonsen (2002b). In this sense, user-actors, either individually or collectively, are at the heart of LMCPs. As already explained, LMCPs participate in our study, construction, interpretation and sharing of meaning triggered by linguistic signs and other signs, inside AND outside dictionaries. In fact, our collaborative making and use of lexicographic objects (dictionary platforms, glossaries, termbases, encyclopaedias, etc.) is only one aspect of LMCPs.



## 4.2 Needs for information and satisfaction of needs

Function theory (Tarp 2008) is similar to “Uses and Gratifications Theory” (Katz et al. 1973), which is individualistic and focussing on gratification that can be derived from the use of information and media. So, what characterizes our information needs as lexicographic user-actors? Moreover, what are our ‘lexicographically relevant’ needs? Are they simply needs that (somehow tautologically) can be satisfied by means of a dictionary? In our hypermediated society, we should remember that understanding the constructive power of language is even more important than ever. We subscribe to the realization of language as one of the most important building blocks of our perceived and interpreted reality, as new needs for meaning outside the dictionary arise. Our needs are much more comprehensive and are embedded in our every-day life.

One example of how needs constantly arise, and how they are constructed, can be illustrated by the corona virus in the spring of 2020. Corona has boosted lexical creation. New competing words and expressions are created, forms and meaning become unstable, and to various degrees, we all take part in meaning construction processes. Discussions flourish on social media. What is the meaning of new corona words? Are they the right words to use? How should they be spelled? Defined? Preferred? Rejected? Translated into other languages? What kind of cognition is at work behind them, since words are never innocent? When the French speak of “gestes barrière” (barrier gestures) to stop the virus through “distanciation sociale” (social distancing), they live in a war metaphor. The term “gestes de protection” (protection gestures) might just as well have been used to signify a far less combative view on the crisis. And even more dramatic: what is a virus in the first place? And what do we know exactly about this one? What does it mean for us? In situations like that, LMCPs are at the heart of the search for meaning construction and satisfaction. The purpose is to establish mutual understanding in a moving world created and manipulated by changing words and meanings, and by changing knowledge constructions.

Because dictionaries refrain from including neologisms as they appear, waiting for meaning to stabilize and fossilize, we correspondingly take on LMCPs to satisfy our needs to find the right words (at least right in our minds). The idea of mirroring language is utopian, because the reflection is always novel and changing. This explains why we here choose to reconceptualize ‘lexicographically relevant information needs’ and ‘satisfaction of needs’ as the search for meaningfulness – finding the right words – through activation of situated, relevant LMCPs. Function theory alone cannot explain how information needs can be satisfied if not extended to processes outside dictionaries. Relevance certainly matters, as the theory claims (Bothma & Tarp 2012), but relevance and cognitive consonance alone cannot explain it all. Cognitive dissonance, not falling into terms with meaning, is an equally important factor in the dynamics of lexicographic meaning-construction. Under consultation, satisfaction is determined by the interplay of LMCPs and cognitive structuration.

## 4.3 Missions and organizational goal-achievement in C&Os

C&Os have communicative needs and goals, but how do they achieve these? Not only do we as individuals have communicative needs or cognitive needs (Tarp 2008), but also C&Os and work communities have communicative needs (Simonsen 2002a). For this reason, they develop communication strategies. LMCPs are deeply involved in organizational communication, in efforts to make sense and construct the right discourse. This is necessary to support their mission. In their resolution to continuously legitimate their business and differentiate their value proposition in an even more competitive business environment, C&Os include in their business and organizational communication an array of strategies based on formats and registers of knowledge communication and language mediation, which hitherto were exclusively used by lexicography (Leroyer 2018). By doing so, C&Os simultaneously pursue their sales objectives and support the branding of their image and reputation (Leroyer 2007, 2011). These strategies may include explicative, didactic and pedagogic formats and have a strong focus on dialogue and interaction with primary stakeholders, the consumers or end users as part of their mission. This shows how LMCPs are actively at work. New company words are created, old ones become obsolete. Words are discussed, interpreted, negotiated, and are included into C&Os’ company speak. Others become parts of mission statements and of the in-house vocabulary and lexical assets of the organization.

At another level, non-profit, national lexicographic institutions and organizations use similar strategies in their public affairs efforts to promote the branding and ongoing development of scholastic, national dictionaries and gain legitimacy and political support. This includes fund raising strategies. In this communicative context, lexicography undergoes radical changes and develops new formats in which LMCPs are even more salient. From being normative or descriptive, lexicography becomes reflexive, interactive and co-active. C&Os do not simply compile lexicographic resources such as specialized in-house glossaries, termbases, query systems etc., but actively use these to discuss the making itself of glossaries, their design, their functions, transforming them into collaborative branding instruments, c.f. Simonsen (2007) and Leroyer (2018). It is no longer sufficient to craft definitions of company keywords and terms; focus is on discussion of principles for the crafting and understanding of terms. Similarly, multimedia components are not simply added and made accessible online, they are explicitly used and marketed to achieve pedagogic, didactic or gamification goals (Leroyer & Simonsen 2018a, b). These are all expressions of situated, organizational LMCPs at work.

Questioning knowledge construction and communication when discussing the construction of the meaning of terms is turned into a dynamic, collaborative activity. Key user-actors – C&Os and their customers or clients – negotiate the meaning of terminologies that are central to their business and brand experience in line with the ongoing transformation of the environment. Above, we used the corona virus example, but C&Os often launch new technologies, brands and services motivated by for example climate change or sustainability agendas, which is yet another example of dynamically constructed communicative needs. LMCPs are at the heart of all term-related activities of designing, naming, defining, and discussing, when new professional knowledge is constructed and needs to be communicated, shared and understood. The same reflexive activity can be experienced in the mission of non-profit, national or European lexicographic institutions in the development of new lexicographic resources and networks. They include the preparation and delivery of data



resources for NLP businesses, involving all key user-actors, the general public, C&Os, as for example through the European Lexicographic Infrastructure (ELEX.IS 2020). National and European language and research policies are determinant for collective, institutional LMCPs, and for the framing and achievement of communicative goals and missions.

#### 4.4 Systems and meaningful interfacing

What are the relations between us and the structuration of lexicographic systems, including dictionaries? How do LMCPs relate to systems? Interfacing determines how data can be meaningfully accessed or generated through individuals or organizational systems, and how specific LMCPs come into play. This is for example the case of the Oenolex dictionary platform (Leroyer 2018) which allows professional organizations of the wine business in French Burgundy to edit and access data according to their business agenda (production, marketing, sales, distribution, organizational communication, public relations). LMCPs steer the navigation of wine actors in the heterogeneous lexical landscape of wine (production, business, law, consumption). Wine actors use them to construct and express their identity and culture through wine speak. The platform is also used to differentiate their position towards competitors or legislators, whom they often do not agree with, for instance when new categories of terms are created and imposed to designate new properties of the wine (mineral, minerality). LMCPs are then used to resist neology.

Interfacing between us, dictionaries and man-machine systems has changed. Specifically situated LMCPs are involved in which concepts like ergonomics and intuition are subject to meaning constructions. What is to be understood as ergonomic or intuitive or not? Some systems are clearly designed for us, while others are designed to be accessed by other systems through APIs or web services. As interfacing is increasingly becoming multimodal, new specific LMCPs come into play. Systems include mobile or ubiquitous systems featuring constantly updated interfaces on our smartphone. Systems also become increasingly multimedial, as they are accessible on multiple media platforms (Leroyer & Simonsen 2020). The concept of proximity is of high interest here, because data are accessible everywhere, at any time. Besides proximity, the key word is integration, and specific LMCPs will format and determine whether integration is meaningful or not.

Interfacing also becomes highly personal. In patient-centered health information systems for instance, in connection with electronic patient health records, medical data meaning interpretation – opaque terms, complex figures, graphs, unknown symbols, and interpretation of findings – the so-called epicrisis – is at the heart of integrative interpretation systems. Lexicographic solutions are involved and highly personal LMCPs are at work: They have a dramatic impact on a patient's ability to interpret personal, medical information and turn it into meaningful patient-doctor-hospital communication.

#### 4.5 The generative and polysemiotic nature of data

The next dimension of our reconceptualization addresses the nature of what is normally understood by lexicographic data. We think of them as constructs generated by LMCPs. We also think of them in terms of polysemioticity.

What characterizes lexicographic data? From our constructivist position, we refuse to see lexicographic work as the exclusive selecting, recording, explaining and structuring of lexical data, and the securing of access to these so we can turn them into useful information. We do not believe that data are already present in the world, given to us, and waiting for us to be selected and recorded. Data are generated by lexicographers and actors and acquire their status of data as the outcome of specific LMCPs governing lexicographic work, including the compiling of corpora, which are themselves the outcome of LMCPs. Data are constructed by us as user-actors through shared LMCPs to generate meaning and secure mutual understanding. Therefore, we prefer to speak of data generativeness. Outside dictionary making, lexicographic data are constantly generated as new meaning data and metadata related to word creation and critical reflexion on usage of new words appear. This is why neonymic processing, as in the case of the Corona virus example above, is so highly relevant for us. The word 'virus' is a key to conceptual knowledge. What is a virus? Is it alive? How are we at risk? What can we do to protect ourselves? In times of crises and dramatic changes in society, language is challenged, and new LMCPs rapidly spread to bring meaning back on track. This entails massive data generation and massive LMCP activity.

In addition to that, there is also an increasing need for high-quality lexicographic data in both conventional lexicographic systems and in new ones such as for example writing assistants or augmented writing tools (Simonsen 2020a, 2020b). Without lexicographically curated data, these systems are not good enough because they lack the world knowledge and relational knowledge of humans. Here also, we engage in LMCPs to keep a critical distance to the output of machines, because we are intuitively aware of how world knowledge is constructed through word knowledge.

Lexicographic data go far beyond linguistic signs. There are other meaning modalities. Lexicographic systems today provide multimedial data (Leroyer & Simonsen 2020) in which lexical data are supplemented or replaced by multimedia data types. In LMCPs, other signs than linguistic signs are processed. These include simple and complex semiotic units of meaning and understanding, including icons, gestures, pictures, sounds, animated pictures, graphs etc. As already explained, C&Os provide many interactive, polysemiotic data types already through their communication (*Vins de bourgogne*) or (*Altomhus*) that are supported by LMCPs. In this light, it is surprising that lexicography, which is extensively language-determined, seems to have failed to remember that language is a faculty, not simply a system. As lexicographic user-actors we demand and process a huge variety of meaningful, non-linguistic signs. We do so by way of easy-to-use technology (Weill & Woerner 2018) because we want to understand and navigate in today's highly polysemiotic world. LMCPs rely on data generativeness and polysemioticity to generate a vast array of meaning constructions in many languages (language being used here in the Saussurean sense of the word, and not simply in the sense of natural language).



#### 4.6 The multimodal nature of access

The sixth dimension of our reconceptualization is the multimodal nature of access. LMCPs are unfolded through two very different types of access modalities, search and explore. Access can similarly be divided into human-driven vs. machine-driven access (Simonsen 2020b; Colson 2019). Search is conventionally understood as a dictionary determined behavior, but need not be. We also perform a great number of dictionary independent searches all the time.

Search should not be understood as being the initial stage of LMCPs. Search never starts out of the blue, but is preceded by LMCPs leading to it and shaping it. Searches will then be accompanied and followed by other related LMCPs.

Explore is a serendipitous access modality. It is dynamically framed by specific LMCPs that establish sudden meaning on the move, as a surprise element, in the course of what is randomly discovered and suddenly makes sense.

With writing assistants and augmented writing technology, new modalities emerge. Lexicographic data are not only used by us, but are increasingly shared in automated writing technologies. As a result, there is a shift from human-based decision-making to human-based, machine-assisted decision-making (Simonsen 2020b; Colson 2019). Making decision on meaning, and deciding whether the right words are used or not, is subject to specific, highly conscious LMCPs that aim to qualify our understanding of the output of the machine and help us keep a critical distance, whenever the machine is suspected of failing to make itself understandable. New modalities of interpreting and understanding come into play.

Access modalities are changing dramatically, and the concept of access itself calls for reconceptualization. In our IT world where (seemingly) meaningful information can be provided instantly by search engines and AI algorithms, we may need to turn access upside down (Simonsen 2020a, 2020b). We may have to introduce a new division of labour and let machines do the labour-intensive and big data-intensive work, while we shall provide world knowledge that is missing. The point here is that it is no longer a matter of access in the traditional lexicographic sense. New conceptual LMCPs are increasingly being called upon to bring in world knowledge to the extent it is needed to validate sense.

#### 4.7 Value-creation and -adding in lexicographic business

What is the business of lexicography? What type of value does it create or add in business? When we speak of 'lexicographic business', we choose to make a distinction between lexicography as a business proper, whether private or not (editors, dictionary platforms, language academies and departments, dictionary societies etc.) and lexicography in C&Os as part of a business. Because two different types of business-related LMCPs are involved.

In the former, business proper, total LMCPs are at work and determinant for all business processes. Lexicography then coincides (or should ideally coincide) with value creation inside and outside the business, and LMCPs at work are determinant for all processes in the organization. That goes for lexicographic business objectives (legitimacy, profitability) as well as internal and external organizational communication, including decision-making, power relations and ethics.

In the latter, lexicography as part of a business, partial LMCPs govern C&Os' ability (or not) to include lexicographic processes as value adding to the business and to internal and external organizational communication.

Business is the domain in which lexicography has changed the most, and probably the most visible part. Dictionaries are no longer the golden eggs they used to be, and business is in the process of constantly revising its models to stay in business (Simonsen 2017). The value chain has also been challenged (Simonsen 2017). In the Internet age, we want our information needs to be satisfied easily and swiftly, and we want it for free (Weill & Woerner 2018). Editorial production systems of the past are disrupted by IT providers, who have developed language platforms on which conventional lexicographic data are but a minor part of the entire value proposition (Ordbogen.com 2020). The concept of market is also subject to reconceptualization, as lexicography creates or adds value on radically different markets. New business models emerge and are being constructed, as lexicography, either directly or indirectly, is now implied in the delivery of lexicographic data sets for artificial intelligence applications, or experience-based services in C&Os (Simonsen 2007: 411-412).

In any case, the lexicographic business is challenged (Simonsen 2016). There seems to be a gap between what we as user-actors demand (Customer Profile) and what the business can offer us (Service Profile), as discussed by Osterwalder & Pigneur (2010) and Osterwalder et al (2014). Business models are dynamic constructs and need to be constantly revisited in order to create or add the value we demand. The more closely and harmoniously integrated LMCPs are in the efforts to identify and achieve business objectives, whatever these are – to market dictionary platforms, support national language policies, inspire CALL applications, enrich branding experience, sales- and experience-based services, fuel AI applications – the better it is for the business and for the sense making processes of the business.

We live in the information and knowledge society, but drown in information (and misinformation) flooding and knowledge communication asymmetries. New knowledge is constructed and radical changes in society challenge language, making meaning fluctuating. These are problems that lexicography throughout centuries has addressed, providing strong solutions. Whether C&Os today are business actors creating full lexicographic value, or partial actors, using lexicography for value-adding, reconceptualizing the business of lexicography as organizational processes determined by business-determined LMCPs provides a better understanding of what is going on. Whether we do lexicography for a living, or simply take advantage of it for a better understanding of how meaning is constructed in our life, be it private or professional.

### 5 Concluding remarks: dictionaries are the tip of the iceberg

We have argued that the understanding of the term lexicography can benefit from a reconceptualization of its ontology and epistemology, and consequently an expansion of the object centered user-data-access model. While lexicography, guided by centripetal forces of reasoning, is extensively understood as the result of a toolification process – the making, use and life of lexicographic use-objects – we have argued, guided by centrifugal forces of reasoning, that lexicography should also



be seen as lexicographic meaning-construction processes, whether these are focalised on lexicographic use-objects or not. Although extremely visible, objectivation is but one mode of the ontology and epistemology of lexicography. Dictionaries are the tip of the iceberg. We have introduced a seven-faced model, which should be sufficiently broad to allow both scientific positions to be placed in a continuum, lexicography as product-determined at the one end and lexicography as process-determined at the other. The model reconceptualizes the following elements:

- An extension from lexicographic ‘users’ to lexicographic ‘user-actors’ dynamically authoring and shaping LMCPs.
- A distinction between individually driven functions and collectively driven missions, i.e. between information needs aiming at individual satisfaction and communicative needs aiming at mutual understanding and organizational goal achievement. In both cases, LMCPs can explain how satisfaction and goal achievement can be achieved.
- A recognition of the meaningfulness of interfacing as a condition for LMCPs to unfold.
- A recognition of the generativeness and polysemiotic nature of data through LMCPs.
- An understanding of LMCPs as keys to lexicographic business throughout a large number of organization processes, in which business-related LMCPs are building blocks of organizational communication.

It is our hope that the suggested reconceptualization of lexicography will inspire work framed by the broad understanding in the future, and will expand the fields of investigations. Particularly, research will be needed to describe and explain how LMCPs appear and evolve, and also determine how they impact the interactions of the different elements of the model. It should be possible to expand LMCP theory beyond the scope of a single, unifying theory, to a network of interdependent social-constructivist theories in all fields where lexicography is involved. This is only a modest beginning and the broad, all embracing understanding of lexicography presented here should be a key to new social research agendas in the field. Several disciplines will have to work together with lexicography e.g. social cognitive linguistics and social terminology, communication and media sciences, information sciences, business sciences, ethnographic science and social IT. The social-constructivist position has revealed a whole range of processes that hitherto were not identified as lexicographic as such. Perhaps a case of not seeing the forest for the trees? Where lexicographers seem to worry about the future of lexicography because dictionaries may well be becoming obsolete, we do not. Lexicography has always been very important to us right from the very beginning, but there has probably never been so much lexicography around us and between us as there is now. A reconceptualization of lexicography is the key to renewed optimism. Lexicography helps us make sense of our language faculty and of ourselves, and helps us better understand and make sense of the world we build, transform, share and ultimately care for.

## 6 References

- Adamska-Sałaciak, A. (2018). Lexicography and theory: clearing the ground. In *International Journal of Lexicography*, 2019, Vol. 32, No. 1, pp. 1–19.
- Altomhus. Accessed at <http://www.altomhus.dk/>. [19/04/2020].
- Bergenholtz, H. & Agerbo H. (2017) Types of Lexicographic Information Needs and their Relevance for Information Science. In *Journal of Information Science Theory and Practice* 5(3), 15-30.
- Bergenholtz, H. & Gouws R.H. (2012). What is Lexicography? In *Lexikos* 22: 31-42.
- Bogaards, P. (2010). Lexicography: Science Without Theory? In G.-M. de Schryver (Ed.) *A Way with Words (Festschrift for Patrick Hanks)*. Kampala, Uganda: Menha Publishers, pp. 313-322.
- Bothma, T, Tarp, S. (2012). Lexicography and the relevance criterion. In *Lexikos* 22(1), pp. 86-108.
- Christensen, C. (1997). The innovator's dilemma. When new technologies cause great firms to fail. Boston: Harvard Business School Press.
- ELEX.IS (2020). *European Lexicography Infrastructure*. Accessed at: <https://elex.is/>. [19/04/2020].
- Fadel, C., Trilling, B., Bialik, M. (2015). *Four-Dimensional Education: The Competencies Learners Need to Succeed*. Center for Curriculum Redesign. ISBN-13: 978-1518642562.
- Fauconnier, G. (1994). *Aspects of Meaning Construction in Natural Language*. Cambridge: Cambridge University Press.
- Fuertes Olivera, P. A. (2016). A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. In *International Journal of Lexicography* 29(2):226-247.
- Hausmann, F. J. (1985). Lexikographie. In C.Schwarze & D. Wunderlich (Eds.) *Handbuch der Lexikologie*. Königstein/Ts.: Athenäum, pp. 367-411.
- Jakobson, R. (1960). Linguistics and poetics. In T. Seboek (Ed.) *Style in language*. Cambridge: MIT Press, pp. 350-377.
- Katz, E., Blumler, J. G. & Gurevitch, M. (1973). Uses and gratifications research. In *The Public Opinion Quarterly*, 37(4), pp. 509-523.
- Leroyer, P. (2007). Bringing corporate dictionary design into accord with corporate image: From words to messages and back again. In H. Gottlieb & J.E. Mogensen (Eds.) *Dictionary Visions, Research and Practice*. Selected papers from the 12th International Symposium on Lexicography. Amsterdam: John Benjamins Publishing Company, pp. 109-117.
- Leroyer, P. (2011). Change of Paradigm in Lexicography. From Linguistics to Information Science and from Dictionaries to Lexicographic Information Tools. In P. A. Fuertes Olivera & H. Bergenholtz (Eds.) *E-lexicography: Internet, Digital Initiatives and Lexicography*. London, New York: Continuum, pp. 121-140.
- Leroyer, P. (2013). Oenolex Burgundy: New Directions in Specialised Lexicography. In D. A. Kwary, N. Wulan & L. Musyahda (Eds) *Lexicography and Dictionaries in the Information Age*. Selected papers from the 8<sup>th</sup> ASIALEX International Conference. Airlangga: Airlangga University Press, pp. 228-235.



- Leroyer, P. (2016). Bruger- og ekspertinddragelse ved udarbejdelse af online (fag)ordbøger: det kooperative princip i leksikografi. In *Nordiske Studier i Leksikografi* 13. Rapport fra Konference om Leksikografi i Norden. Københavns Universitet, pp. 177-190.
- Leroyer, P. (2018). The Oenolex Wine Dictionary. In P. A. Fuertes-Olivera (Ed.): *The Routledge Handbook of Lexicography*. London and New York: Routledge, pp. 438-454.
- Leroyer, P. & Simonsen, H.K. (2018a). Indlæring af fagtekstproduktion på fremmedsprog: om fagordbøger som prædeterminerede læringsværktøjer. In *LexicoNordica* 25, pp. 115-133.
- Leroyer, P. & Simonsen, H.K. (2018b). When Learners Produce Specialized L2 Texts: Specialized Lexicography between Communication and Knowledge. In *Proceedings of XVIII EURALEX International Congress*. Accessed at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2969-1>. [19/04/2020].
- Leroyer, P. & Simonsen, H.K. (2020). Multimediale ordbøger: hvordan og hvorfor. In *Nordiske Studier i Leksikografi* 15, 2019, Rapport fra Konference om Leksikografi i Norden (In press).
- Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. In *Business Management Dynamics*. Vol. 2, Issue 10, April 2013, pp. 49-62.
- Margalitadze, T. (2018). Once Again Why Lexicography Is Science. In *Lexikos* 28 (AFRILEX-reeks/series 28: 2018), pp. 245-261.
- Ordbogen.com (2020). <https://www.ordbogen.com/en/> [accessed 22 may 2020].
- Osterwalder, A. & Pigneur, (2010). *Business Model Generation: A Handbook For Visionaries, Game Changers, and Challengers*. Hoboken: John Wiley & Sons, Inc.
- Osterwalder, A., Pigneur, Y., Bernada, G. & Smith, A. (2014). *Value Proposition Canvas: How to create products and services customers want*. Hoboken: John Wiley & Sons, Inc.
- Piotrowski, T. (2013). A Theory of Lexicography – Is there one? In H. Jackson (Ed.) *The Bloomsbury Companion to Lexicography*. London and New York: Bloomsbury Academic, pp. 303-320.
- Rundell, M. (2012). It Works in Practice but Will it Work in Theory? The Uneasy Relationship Between Lexicography and Matters Theoretica. In R.V. Fjeld and J. M. Torjusen (Eds) *Proceedings of the 15th EURALEX Congress*. Oslo: University of Oslo, pp. 47-92.
- Saunders, M.N.K, Lewis, P., Thornhill, A. (2019). *Research Methods for Business Students*. Eight Edition. New York: Pearson.
- Scerba, L.V. (1995[1940]). Towards a General Theory of Lexicography. Transl. by D. M. T. Cr. Farina. In *International Journal of Lexicography* 8.4, pp. 305-349.
- Simonsen, H. K. (2002a). TeleLex - Theoretical Considerations on Corporate LSP Intranet Lexicography: Design and Development of TeleLex - an Intranet-based Lexicographic Knowledge and Communications Management System. Ph.d.-afhandling, 436 sider. Århus: Handelshøjskolen i Århus.
- Simonsen, H. K. (2002b). User Involvement In Corporate LSP Intranet Lexicography. In: H. Gottlieb, J. E. Mogensen and A. Zettersten (Eds): *Proceedings of the Eleventh International Symposium on Lexicography*. Lexicographica - Series Maior. Copenhagen: the University of Copenhagen, pp. 489-509.
- Simonsen, H. K. (2005). ZooLex: The Wildest Corporate Reference Work in Town? In: *Proceedings of XII EURALEX International Congress*, Volume II, pp. 787-793.
- Simonsen, H. K. (2007a). Virksomhedsleksikografien viser tænder: leksikografiske løsninger i København Zoo og Fagerberg A/S. In: *Nordiske studier i leksikografi* 9. Rapport fra Konference om leksikografi i Norden. Island, Akureyri: Nordisk forening for leksikografi, pp. 383-398.
- Simonsen, H. K. (2007). ZooLex: Det er for vildt! In: *LEDA-Nyt nr. 43* - marts 2007, pp. 4-12.
- Simonsen, H. K. (2012). Et informationsvidenskabeligt serviceeftersyn af Medicin.dk. In *Nordiska Studier i Lexikografi* 11. Rapport från Konferens om lexikografi i Norden. Lund: Nordisk forening for leksikografi, pp. 563-574.
- Simonsen, H. K. (2016). Hvor er forretningsmodellen? En analyse af de forretningsmæssige udfordringer i forlags- og informationsindustrien med særlig fokus på opslagsværker. MBA-afhandling. Institut for Økonomi og Ledelse. Aalborg Universitet.
- Simonsen, H. K. (2017). Lexicography: What is the Business Model? In I. Kosem, C. Tiberius, M. Jakubiček, J. Kallas, S. Krek & V. Baisa (Eds.): *Electronic Lexicography in the 21st Century*. 19–21 September 2017, Leiden, the Netherlands. Brno: Lexical Computing CZ, 395415, pp. 395-415. Accessed at: [elex.link/elex2017/proceedings-download/](http://elex.link/elex2017/proceedings-download/). [19/04/2020].
- Simonsen, H. K. (2020a). Augmented Writing: nye muligheder og nye teorier. In *Nordiske Studier i Leksikografi* 15. Rapport fra Konference om Leksikografi i Norden. Helsingfors: Nordisk forening for leksikografi (in press).
- Simonsen, H. K. (2020b). Når Augmented Writing og leksikografi går hånd i hånd. In *LEDA-nyt nr. 69*, april 2020, pp. 3-13.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographic Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Tarp, S. (2018). Lexicography as an Independent Science? In P. A. Fuertes-Olivera (ed.) *The Routledge Handbook of Lexicography*. London and New York: Routledge, pp. 19-33.
- Tarp, S. (2019). Connecting the Dots: Tradition and Disruption in Lexicography. In *Lexikos* 29, pp. 224-249. *Thefreedictionary*. Accessed at [www.thefreedictionary.com](http://www.thefreedictionary.com). [19/04/2020].
- Temmermann, R. (2000). *Towards New Ways of Terminology Description. The Sociocognitive Approach*. Amsterdam and Philadelphia: Benjamins
- Trap-Jensen, L. (2018). Lexicography between NLP and Linguistics: Aspects of Theory and Practice. In *XVIII EURALEX International Congress*, pp. 25-37. Accessed at: <https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/2969-1> [19/04/2020].



- Tono, Y. (2010). A Critical Review of the Theory of Lexicographic Functions. In *Lexicon* (Journal of the Iwasaki Linguistic Circle) 40, pp.1-26.
- Verlinde, S., Leroyer, P. & Binon, J. (2010). Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats. In *International Journal of Lexicography* 23.1: pp. 1–17.
- Vins de bourgogne*. Accessed at: <https://www.vins-bourgogne.fr/>. [19/04/2020].
- Vecchi, D. de (2013). Company-Speak: A Managerial Perspective On Corporate Languages Seen From The Inside. *Global Business & Organizational Excellence* 33-2, pp. 64-74.
- Weick, K. E. (2009). *Making sense of the Organisation Volume 2. The Impermanent Organization*. Chichester: John Wiley and Sons.
- Weill, P. & Woerner, S. L. (2018). *What's Your Digital Business Model? Six Questions to Help You Build the Next-Generation Enterprise*. Boston, Massachusetts: Harvard Business Review Press.
- Wiegand, H. E. (1984). On the Structure and Contents of a General Theory of Lexicography. In Hartmann, R. R. K. (Ed.) *LEXeter '83 Proceedings*. Tübingen: Max Niemeyer, pp. 13-30.
- Wiegand, H. E. (1988). Was eigentlich ist Fachlexikographie? Mit Hinweisen zum Verhältnis von sprachlichem und enzyklopädischem Wissen. In H. H. der Munske, P. von Polenz, O. Reichmann, R. Hildebrandt (Hrsg). *Deutscher Wortschatz. Lexikologische Studien. Ludwig Erich Schmitt zum 80. Geburtstag von seinen Margurger Schülern*. Berlin, New York: De Gruyter, pp. 729-790.
- Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Accessed at <https://www.gutenberg.org/files/5740/5740-pdf.pdf> [19/04/2020].
- Wordnik* . Accessed at [www.wordnik.com](http://www.wordnik.com). [19/04/2020].
- Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia.



# A Morpho-Semantic Digital Didactic Dictionary for Learners of Latin at Early Stages

Márquez Cruz M., Fernández-Pampillón A.M<sup>a</sup>.

Complutense University of Madrid

e-mail: manmarqu@ucm.es, apampipi@filol.ucm.es

## Abstract

*Diccionario Didáctico Digital de Latín* (Digital Didactic Dictionary for Latin) is an open-access lexicographic work, created and hosted by the Universidad Complutense de Madrid. It is a bilingual dictionary (Latin-Spanish) that faces the challenge of providing Spanish-speaking students of Latin with an innovative lexicographic tool that facilitates the learning of basic Latin. Based on the theoretical principles of valences described in Tesnière's Dependency Grammar, Lyons' ontologies, and Fillmore's theory of semantic frameworks, the dictionary has been conceived as a linguistic tool to understand how Latin works at the semantic, morphological, and syntactic levels. It is a qualitative dictionary created ad hoc, used as an auxiliary tool to answer linguistic questions raised by an inductive didactic methodology that makes the Latin learning-teaching process available even to those students who lack basic syntactic knowledge. A structure of Hierarchical Faceted Categories that constitutes the lexicographic model provides various ways to access the lemmatised lexicon, facilitating intuitive navigation through the dictionary.

**Keywords:** Computational Lexicography; Learner's Dictionaries; Latin Lexicography; Bilingual Lexicography, Digital Dictionaries, Dictionary for special needs

## 1. Introduction

Learner's lexicography as currently conceived has its roots in the 1930s, when M. West, H.E. Palmer, and A.S. Hornby, teachers of English as a second language, took part in various lexicographic projects aimed at improving English learning (Jackson 2002). As a result of this work, in 1938 West published *The New Method English Dictionary*, which is regarded as the first monolingual English learner's dictionary (Heuberger 2016)<sup>1</sup>, Plamer published *A Grammar of English Words*, aimed at the description of verbal models, and in 1942 Hornby – with Gatenby and Wakefield – published the *Idiomatic and Syntactic English Dictionary*, which years later was reprinted as *A Learner's Dictionary of Current English*, and which in 1952 was reprinted again as the *Oxford Advanced Learner's Dictionary of Current English* (Miller 2018). Since then, language learning dictionaries have been produced on an uninterrupted basis. However, this process has not been equal for all languages: Rothenhöfer (2013: 414), cited by Bugueño (2019: 68), and Onieva (2019: 149), states that the production of learner's dictionaries in languages like German and Spanish has not reached the levels of English lexicography. This is a striking situation, given that these are languages with a similar social prestige to English, with millions of speakers, and so their lexicographic treatment should be similar. The situation is even more striking in the case of learner's lexicography applied to a corpus (not dead) language, such as Latin. In fact, in the case of Spain, Latin learner's dictionaries, that is, those that in general are used in schools, are, as we shall see, nothing but lexicographic compendia that, both in their macrostructure and in their microstructure, follow the guidelines set by general Latin dictionaries (*A Latin Dictionary* by Ch. T. Lewis and Ch. Short, *Dictionnaire latin-français* by F Gaffiot, *Diccionario Latino-Español, Español-Latino* by A. Blánquez, and *Totius Latinitatis Lexicon* by E. Forcellini, the core of all these dictionaries), following the same method to list lemmas but restricting meanings, the description of complements, and examples. They are dictionaries intended to help Latin students, particularly in translation.

After observing that Latin learner's dictionaries in Spain over time followed a model inherited from the main general dictionaries, four years ago, we wondered if we could face the challenge of creating a Latin learner's dictionary for Spanish students - or students whose native language was Spanish - that improved learning effectiveness in an innovative way. The goal was to provide students with a lexicographic learning tool that, thanks to its form and contents, would facilitate the process of learning Latin and would motivate its study. The answer is *Diccionario Didáctico Digital de Latín*, which we present in this paper.

## 2. Background. The lexicography of Latin learning in Spain: from 1950 to 2020

The landscape of the lexicography of Latin learning in Spain currently comprises a small set of dictionary options for students – mainly the duo *Diccionario Ilustrado Latín: Latino-Español / Español-Latino* (DIL-VOX-VOX) published by Bibliograf S.A.<sup>2</sup> and *Diccionario Latín (DL-SM)* published by SM<sup>3</sup>, which are the top sellers in Spanish bookshops.

<sup>1</sup> According to Heuberger, citing Cowie (1999).

<sup>2</sup> <https://www.vox.es/>



*DIL-VOX*, at 715 pages, has perpetuated a lexicographic model that, since its third revised and extended 1950 version—which is the basis for all later editions, bringing together the dictionary and a brief grammatical summary of Latin – reached its 24<sup>th</sup> edition in 2011, and is one of the most popular dictionaries in schools. In 2001 the *DL-SM*, with 928 pages and a grammatical appendix, was published. Its renewed aesthetics (García Ferrer 2013) has attracted the attention of new generations of Latin students, coexisting in courses with *DIL-VOX*, a milestone not achieved by the other school dictionaries published from 1950 to 2020, which include *Diccionario de bolsillo latino-español, español-latino* by A. Vives, from 1954, *Diccionario manual latino-español y español-latino* by P. Múgica (7th edition), from 1958, *Diccionario español-latino* by Llauro y Márqués, from 1965, *Diccionario manual latino-español y español-latino* by A. Blánquez, from 1984, *Latín iter 2000* published by Ramón Sopena in 1989, and *Diccionario Cumbre latino-español, español-latino* published by Everest in 1999<sup>4</sup>. All of these are small lexicographic compendia of the larger works cited. Most likely, the reason for their lack of relevance and presence in courses is that the information they give is so extremely scarce that they are not sufficiently rich for text translation, the main skill that Spanish students need to acquire in their learning process, in accordance with the learning standards and goals established by Spanish educational institutions<sup>5</sup>. We would also like to highlight *Diccionario Latín-Español, Español-Latín* by J. Pimentel, which, even though it was published outside Spain, in Mexico (its 12<sup>th</sup> edition is from 2017), is also a leading seller in Spanish bookshops as an alternative to *DIL-VOX* and *DL-SM*, possibly because it provides a set of lexicographic items in line with the dictionaries mentioned. We will discuss this later, in section 4.1.

This is the production in analogical format. In digital format, the Latin learner's dictionaries that have been published are *Didacterion*<sup>6</sup>, with 8,966 lemmas; *Latine Disce*<sup>7</sup>, with 1,950 lemmas; *Glosbe*<sup>8</sup>, a website that brings together bilingual dictionaries in various languages, one of which is Latin-Spanish, the number of whose lemmas is not given but which is continually updated, due to its nature as an ongoing collaborative project; and *Diccionario Didáctico Digital de Latín*<sup>9</sup>, a work which we will present and analyse in this paper.

As for metalexicographic studies of the situation of the lexicography of Latin learning and the study of specific school dictionaries, the only reference that we have found is a paper by García Ferrer (2013), which is a comparative study of *DIL-VOX* and *DL-SM*. This section also includes the papers by Happ (1976) and Favarin (1979), which, despite being published outside Spain, proposed a Latin dictionary that would follow the descriptive valence model. Even though it is conceived more as a general work, we believe it should be included in this section as it is the first theoretical proposal along the lines of one of the theoretical frameworks for the dictionary to which this paper refers.

### 3. The Development of *Diccionario Didáctico Digital de Latín*

#### 3.1. Theoretical foundations

A dictionary can be regarded, in general terms, as a “lexicographic artifact” whose goal is to meet the linguistic, cultural, social, intellectual, and professional needs (among others) of its potential users. In terms of its structure, following Wiegand (2010), it can be understood as a textual conglomerate that can be segmented into a number of components. In the case of language learner's dictionaries, according to Heuberger (2018), each entry would have to include the following elements: definitions, examples and images, grammatical and use information, co-locations, pronunciation, data accessibility for intuitive and productive use of the dictionary, prologue and appendices, information on frequent words, and etymological information. However, depending of the language of the dictionary and the type of user targeted (beginner, intermediate, or advanced), some of these components may be optional: in the case of Latin, for example, the information on co-locations, pronunciation, etymology, and word frequency could be optional, depending on the dictionary's approach. Thus, for a beginner, the absence of etymological information, co-location, and information on frequent words does not make the dictionary lose functionality.

On the basis of these concepts, as well as the analysis of Latin learner's dictionaries in the Spanish-speaking world and the principles of the Functional Theory of Lexicography, a new Latin learner's dictionary model was created. The Functional Theory of Lexicography, posited by Bergenholtz and Tarp (2002; 2003), argues that, when creating a dictionary, its users' social and cultural environment should be first analysed to effectively meet their extra-lexicographic needs. Along these lines, and on the basis of our teaching experience, we designed a model that would facilitate understanding of how Latin works as much as possible. As an innovation, the cognitive function was established as the leading function of the dictionary (Grows 2011). Thus, it was assumed that a Latin learner's dictionary should not limit itself to providing information on the specification of the grammatical category, the description of the specifics of the lemma, complements, or contextual examples - all of which are very useful in translation, but not so much to teach Latin in a way that helps to understand how the language works – but rather should be an instrument whose use in different types of practice would intuitively and inductively provide knowledge of Latin on the basis of the students' native language (in this case Spanish), considering users' cognitive needs and academic profile at all times.

<sup>3</sup> <https://www.grupo-sm.com/>

<sup>4</sup> For a more specific study of the lexicography of Latin learning in Spain, see Márquez (under review).

<sup>5</sup> On the contents and competences to be developed by students of Latin, see the Spanish Official State Gazette legislative provision: <https://www.boe.es/boe/dias/2015/01/03/pdfs/BOE-A-2015-37.pdf> [retrieved on 09/04/2020]

<sup>6</sup> <https://www.didacterion.com/esddl.php> [retrieved on 09/04/2020]

<sup>7</sup> <http://www.latinedisce.net/Dictionary.latin> [retrieved on 09/04/2020]

<sup>8</sup> <https://es.glosbe.com/la/es> [retrieved on 09/04/2020]

<sup>9</sup> <http://repositorios.fdi.ucm.es/DiccionarioDidacticoLatín> [retrieved on 09/04/2020]



### 3.2. Development methodology

To create and implement the new *Diccionario Didáctico Digital de Latín (DDDL)*, a method based on spiral prototype development was used [Boehm, 1986], with four sequential phases: (1) requirement analysis, (2) design and implementation, (3) test, and (4) revision and improvement. Three iterations of phases 2, 3, and 4 have been conducted so far. The results of each phase are given in the following subsections.

### 3.3. Specification of the *DDDL* objectives and requirements

To specify the requirements, an analysis of the type of user for whom the dictionary would be intended was first conducted, to later establish the basic functions that would guide the design of *DDDL*: students learning for the first time a language that is different from Spanish in that it is declined; students who lack basic linguistic knowledge, particularly in syntax; students who are not motivated to study Latin; students who, regardless of their social situation, have access to online materials. The information was obtained through personal interviews, the analysis of academic records<sup>10</sup>, and questionnaires<sup>11</sup>. Once the general profiles had been examined, the *three basic goals* of the dictionary were established:

1. Facilitating and motivating the process of learning Latin.
2. Aiming the dictionary basically, although not exclusively, at beginning Latin students whose native language is Spanish, with little motivation, due to a large extent to the lack of basic theoretical knowledge of Spanish, which hinders learning other languages in general. Teachers who work with this kind of students are also regarded as potential users of the dictionary inasmuch as they can use it as a didactic resource.
3. Using online environments<sup>12</sup> that enable e-learning, b-learning, or m-learning, learning models with which the recent generations of students who are digital natives identify.

### 3.4. Prototype design and implementation

In the second phase of development of *DDDL*, the lexicographic, micro-, and macrostructure were configured. As an innovation, in the lexicographic model for *DDDL*, the verbal lemma, the basic core of Latin discourse, is the main axis around which the rest of units that constitute the lexicographic work revolve. The metaphor of a jigsaw puzzle is used to understand sentence formation and the dictionary as a repository of “lexical” pieces (figure 1). The verbal lemmas in the dictionary are the central pieces of the sentence and must be completed by joining the pieces that characterise basic sentence complementation. In fact, the image of jigsaw pieces is present in each of the dictionary lemmas as a visual element.



Figure 1. Metaphoric representation of the sentence as jigsaw puzzle pieces

The didactic strategy that underlies the metaphor of the jigsaw puzzle to understand sentence meaning is based on semantic knowledge of the formation of sentences in the student's native language to understand the formation of sentences in new languages. Using Dependency Grammar as a hypothesis, sentences are built or interpreted on the basis of the meaning of the main verbs, their valences, and the semantic values of their arguments. Sentences are presented as jigsaw puzzles in which the verb is the central piece (figure 1, grey piece), bringing together a number of mandatory pieces known as arguments to complete its meaning (figure 1, red and orange pentagons). The number of arguments taken by a verb depends on the *valence* of the verb. Once the student has located the verb and established the number of arguments that it takes, they can identify those arguments with the help of *semantic features*, which go from +/-animated and +/-human to +/-definite/non-definite and location. Figure 1 shows, for example, how the trivalent verb “dat” requires

<sup>10</sup> The set of questions given by Tarp (2006) to establish the profile of dictionary users was considered.

<sup>11</sup> As regards motivation, for example, the items that measure the student's motivation builder to learn a foreign language given by Gardner (2004) were adapted.

<sup>12</sup> <http://repositorios.fdi.ucm.es/DiccionarioDidacticoLatin/>



a nominative argument that must be +animate +human (shown in red), a -animate + definite accusative (orange), and a +animate +human dative. This sentence comprehension strategy is the same as that used by students in their native language, so application is direct once they understand the metaphor.

<b>LEMA:</b>	
<i>do, S1 darle algo a alguien</i>	
<i>do, as, are, dedi, datum (1ª)</i>	
<b>CATEGORÍA:</b> Verbo	
LETRA: D	
<b>• MICROESTRUCTURA VERBOS</b> * ARGUMENTOS DEL VERBO	
Numero Argumentos Significado 1: <i>TRIVALENTE</i> Significado 1: <i>darle algo a alguien</i> Primer argumento: <i>Nominativo</i> Caracterización Argumental 1: <i>+animado +humano</i> Segundo argumento: <i>Acusativo</i> Caracterización Argumental 1: <i>-animado +definido</i> Caracterización Argumental 2: <i>-animado -definido</i> Tercer argumento: <i>Dativo</i> Caracterización Argumental 1: <i>+animado +humano</i> Ejemplo: <i>Vir consilium filis dat.</i>	

Figure 2. Features and values of the verbal lemma “do” (“to give”)

Thus, the lexicographic design of the dictionary was carried out considering the following theoretical principles that underlie the didactic strategy of the jigsaw puzzle:

- Dependency Grammar (Tesnière 1959; Ágel, V. & Fischer, K. 2010): following Valence Theory, verbal lexicographic articles describe the quantitative and qualitative valence of those units. This provides a classification and search mechanisms for verbs on the basis of their argument complements, that is, the number of mandatory complements required by a verb for its semantic development to be complete. Thus, a distinction between monovalent, bivalent, and trivalent predicates is made (figure 2) - a valent cases are not included, as there are no argument complements. There are cases, such as *s.v. peto*, in which the same verb has two alternative valences (a trivalent and a bivalent one): dictionary users can see how a change in the number of arguments entails a change in meaning. The same thing happens when a verb has two quantitatively identical but qualitatively different valences; *s.v. placeo* is a bivalent verb that has two different meanings, depending on the morphology of its arguments: M1 To delight someone (nominative + dative) – M2 To decide something (nominative + infinitive).
- Ontologies: taking as a theoretical basis the distinction between first-, second-, and third-order entities established by Lyons (1977), we gave an ontological definition of the verbal arguments and nouns in the dictionary depending on their lexical characterisation: +animate +human (human beings), +animate -human (animals and living plants), -animate +definite (objects, concrete entities), -animate -definite (concepts, feelings, abstract entities) and places. Ontological differences in complementation help to distinguish meaning. Thus, *s.v. occurro* gives two different meanings for the same number of arguments (bivalent verb), with the same morphology (nominative + dative): “meeting someone” with a first argument in the nominative (+animate +human) and a second argument in the dative (+animate +human) as opposed to “facing something”, with a first argument in the nominative (+animate + human) and a second argument in the dative (-animate -definite).
- Definition of semantic functions: the dictionary gives the option - not yet public - of showing users the semantic function of verbal arguments on the basis of the semantic roles defined by Fillmore (1968:1971). *DDDL* is a lexicographic tool that is part of an inductive methodology developed for early stages of Latin learning. In this methodology, verbal argumentation is explained by means of argument valence, as well as the morphology, ontology, and semantic roles of the elements. These roles are not labelled, but explained on the basis of Fillmore’s definition of cases. However, we thought that it would be important for the dictionary to provide the label of each argument role, so that advanced users can access that information at a later stage in their learning.

Lexical units - or the jigsaw pieces - are arranged in the dictionary, grouped into lexical category, meaning, semantic feature, or valence. To do so, we followed the Hierarchical Faceted Categories model given by Hearst et al. (2006): “[...] build a set of category hierarchies each of which corresponds to a different facet (dimension or feature type) relevant to the collection to be navigated.”<sup>13</sup> (Hearst et al. 2006: 86). In the case of *DDDL*, the facets are the part of speech categories it contains: verb, noun, adjective, pronoun, adverb, preposition, conjunction, interjection, and particle. Each category is structured in a hierarchy according to the semantic and morphological features that characterise the lexical units in each category. Thus, in the case of the verb category, which is the most complex one, the first level is the feature valence of the first meaning of the verb with monovalent, bivalent, or trivalent values (figure 3). The second level is structured in turn

<sup>13</sup> A more detailed description of inductive and dynamic creation of navigation hierarchies can be found at (Fernández-Valmayor, et al., 2013)



into several subhierarchies: one for meaning (Spanish translation), another one for the first argument characterization (with all possible cases), another one for the second or third argument characterization (in the case of bivalent or trivalent verbs), and finally a subhierarchy for each predicative framework characterization for the following meanings of the verb, when it is polysemous.



Figure 3. Hierarchy of features and values for the verb category

This cognitive organisation is used to create the dictionary navigation scheme. In this way, students can not only conduct conceptual searches, but also inductively learn about semantic and morphological features and their potential values while exploring the dictionary. This learning is, in our didactic hypothesis, crucial to understanding how Latin works. In addition to this navigation mechanism, students can search by using the traditional text search systems, entering a word or part of a word in Latin or Spanish in a text box, or by means of an advanced search form.

As regards implementation, the dictionary currently comprises 720 lemmas of public use. Collaborative work continues to gradually increase the number of dictionary entries<sup>14</sup>. To select verbal lemmas, the criterion followed was the frequency of appearance in the *LatinISE historical corpus* v2.2, using Sketch Engine<sup>15</sup> to access the data. Terms that appeared between 200 and 5,000 times in the corpus - which comprises the pre-Classical, Classical, and post-Classical periods of Latin - were selected. These limits discounted working with verbs that would all too often also display a high degree of polysemy, which would hinder meaning and complementation selection (with the exception of such verbs as *habeo*, *do*, *mitto*, *sum*, *accipio*, *peto*, *pono*, and *quaero*, which are verbs with the basic semantics to create sentences in the activities proposed to work with the dictionary). The idea of working with verbs of scarce relevance in the corpus was also rejected, as they tend to have such restricted semantics that they would not serve to enrich the dictionary. Verbs that would cover all types of action, process, status, and position were selected<sup>16</sup>.

In the second stage, the *Oxford Latin Dictionary* and *A Latin Dictionary* were used to compare the definitions and complements of the selected lemmas<sup>17</sup>. Julius Caesar's *De Bello Gallico* was used as a secondary source to find examples in the specific work of an author that is usually taken as a reference in the EvAU examination (the Spanish Assessment for Access to University). This latter decision was based on the argument that students should gradually become familiar with a lexicon that they might use again in later years of study, should they choose the Latin option in EvAU.

As for nouns, selection was much simpler, finding lexical units that represented persons, animals, plants, tangible entities, abstract entities, and places on the basis of verbal complement ontologies. The sources used were again *De Bello Gallico* and *Diccionario Latino-Español, Español-Latino* by A. Blázquez (1997), specifically the Spanish-Latin dictionary part. Adjectives were selected considering the nature of the lemmatised nouns, so that they could properly perform their specifying or explanatory roles. When selecting the rest of lexical units (adverbs, pronouns, prepositions, conjunctions, interjections, and particles) Rubio & González Rolan's grammar (1990) was used.

Finally, as regards the examples that illustrate meanings and complements, they were created ad hoc by three Latin teachers, members of the Project for Educational Innovation, of which *DDDL*<sup>18</sup> is part, thus passing three filters.

### 3.5. Prototype test, revision, and improvement

In phase 3, the test stage, the educational effectiveness of prototypes was assessed using two mechanisms:

<sup>14</sup> 80 new verbal lemmas will be opened for public use in September.

<sup>15</sup> <https://www.sketchengine.eu/>

<sup>16</sup> This classification corresponds to the State of Affairs described by Dik (1997), which is what happens in a real or imaginary world, represented by a verbal predicate and its arguments.

<sup>17</sup> At this point, we would like to emphasise that these dictionaries were only used as guides, given that, as we have argued, we did not intend to create a heritage dictionary: we do not follow the usual synonymic translation procedure but explain the verbal expression by referring to the actants involved in the action, process, state, or verbal position described.

<sup>18</sup> Educational Innovation Project PIE 245: 2019\_20 funded by the Universidad Complutense de Madrid. The dictionary was completed within the framework of four sequential educational innovation projects since school year 2016-17. More information at: <https://www.ucm.es/afpc/proyecto-de-innovacion-docente-y-mejora-de-la-calidad>



- (i) Didactic quality, through the assessment tool of Spanish standard UNE 71362 “Quality of digital educational materials” (UNE71362:2020). The assessment tool UNE 71362 comprises 15 criteria with 87 items to assess three aspects: the didactic, technological, and accessibility effectiveness of digital educational materials<sup>19</sup>. The tool was applied to assess the first *DDDL* prototype in July 2017, by means of a peer review in which 3 Latin teachers and 2 students took part. In this assessment, specifically, the teacher's and student's profile for the tool were used (CTN 71/SC 36, 2017). The results made it possible to correct the prototype before its experimental application.
- (ii) Experimental assessment of didactic effectiveness. This assessment is ongoing. The research question that is being tackled is whether *DDDL* really helps or facilitates the comprehension or generation of simple sentences in Latin. To this end, a longitudinal quasi-experimental study is being conducted of measures repeated during the three academic years 2017-18, 2018-19, and 2019-20. This kind of design has been selected given the impossibility of having randomly assigned groups of students<sup>20</sup>. So far, an initial study case (Márquez & Chávez, 2016) and four experiments<sup>21</sup> have been conducted. The results of the first of the experiments indicate that the new didactic methodology improves motivation and, more specifically, the “positive attitude towards learning Latin among students” (Márquez & Fernández-Pampillón, 2019). Also, possibly, as a consequence of this positive attitude, an improvement is observed in “the stimulus of parents towards the learning of their children”. The rest of the results are currently being analysed, though this activity has been interrupted by the current health situation as the experimental materials are not available. We expect to publish them shortly.

The results of each experiment served to revise and improve *DDDL* in phase 4, reaching the current configuration, which is presented in section four below.

#### 4. *Diccionario Didáctico Digital de Latín*

*DDDL* is a bilingual lexicographic work to learn Latin hosted in a digital repository at the Universidad Complutense de Madrid, created using the *OdA* software. Access to the dictionary, which is online and free, is regulated by a *Creative Commons Attribution-ShareAlike 4.0*. According to the lexicographical model presented in the previous section, the dictionary navigation system allows varied access to the lexicographic data: in addition to simple queries (by lexical unit, based on form, meaning, conjugation, gender, part of the chain of characters that constitute the lexical chain) and complex queries (a way of accessing the data where it is possible to specify, for example, the number of mandatory arguments, their morphology and ontology, and the meaning), as an innovation, lemmas can be conceptually accessed by navigating through the dictionary by grammatical category and semantic-morphological features. Each of the nine grammatical categories has its own structure, generating the entries described below:

- Verbs: **lemma** (the first person singular of the present indicative is used); **meaning**; **classic form** (the traditional nomenclature is used: first person singular of the present indicative, present infinitive, first singular of the past perfect and the supine)<sup>22</sup>; **verb conjugation** (specification of the type of verb conjugation to which the verb belongs); **category** (grammatical category: verb); **letter** (initial letter of the lexical unit)<sup>23</sup>; **verb valence** (description of the type of verb on the basis of the number of arguments it takes: monovalent, bivalent, trivalent); **meaning** (relative to the verb valence); **argument position** specifying its **morphology** and its **ontology**; **example**.  
If a verb has two valences, each of them is described by means of the features described above.
- Nouns: **lemma** (nominative singular); **meaning**; **classic form** (nominative and genitive singular); **gender** (masculine, feminine or neuter, stated as an abbreviated form); **category** (noun); **letter**; **logical categorisation** (ontology); **meaning**.
- Adjectives: **lemma** (nominative singular of the different genders); **category** (adjective); **letter** and **meaning**.
- Pronouns: **lemma**; **category** (pronoun); **letter** and **meaning**.
- Adverbs: **lemma**; **category** (adverb); **letter** and **meaning**.
- Preposition: **lemma**; **category** (preposition); **letter** and **meaning**.
- Conjunction: **lemma**; **category** (conjunction); **letter** and **meaning**.
- Interjection: **lemma**; **category** (interjection); **letter** and **meaning**.
- Particles: **lemma**; **category** (particle); **letter** and **meaning**.

<sup>19</sup> An English version of the standard will soon be published at <https://www.une.org/>. A summary can be found in (Fernández-Pampillón, 2017).

<sup>20</sup> This limitation is in any case frequent in experiments in education contexts (Schanzenbach 2012: 221).

<sup>21</sup> Two more experiments, making a total of seven, had been scheduled, but they were cancelled due to the shutdown of educational centres in the second quarter due to the Covid-19 pandemic. The publication of the results of the last two experiments has had to be postponed for the same reason.

<sup>22</sup> To distinguish the meaning from the lemma, a different font is used, and the different meanings are numbered (S1, S2). The meaning is not reduced to a mere semantic equivalence in the target language, as is standard in bilingual dictionaries, but the arguments involved in the verbal expression are specified. As for the classic form, the font is the same as that of the lemma so that their connection is clear. The same procedure is followed for nouns.

<sup>23</sup> This information is included in all the lexical units that constitute the dictionary and makes it possible to edit the dictionary on the basis of the alphabetical order of the units it comprises.



Regarding the hyperstructure of the dictionary, in addition to the macro- and microstructure described, *DDDL* has an introduction and appendices explaining how the dictionary works by means of videos (Guidelines for Use), explaining the abbreviations used in the dictionary, giving an introduction to a Latin course whose activities are connected to the use of the dictionary (Introduction to Latin Course), providing additional materials for study of the first Latin declinations, verbs, and prepositions (Lessons), providing links to download the main scientific works published about the dictionary (Related Publications), and giving the list of participating and collaborating teachers and researchers (Working Team).

#### 4.1. Comparing the *Diccionario Didáctico Digital de Latín* microstructure to other learner's dictionaries

The microstructure of the lexical entries given in *DDDL* differs both in the form and in the basis of the treatment given to lexicographic items from other Latin learner's dictionaries, in particular in the case of verbal and nominal units. Specifically, in the case of verbal lemmas, the first innovation can be seen in the way in which the items are specified: the first person singular of the present indicative is used, followed by the meanings associated with the lemma and the conventional form given in the rest of Latin dictionaries (first and second person present, present infinitive, first person singular of the past perfect and the supine). The justification for this decision on the formalisation of verbal lemmas is a didactic argument: bookending the meanings of the lemma with the present form - as it is the first of the tenses usually taught<sup>24</sup> - and the rest of the verbal forms given according to convention, so as to comply with the standard and to allow users to become familiar with the form of the entry for verbal lemmas in other dictionaries. Regarding the rest of the information given in the lexicographic articles, we will now explain, by means of a comparative study, the differences between the treatment of entries in *DDDL* and other Latin learner's dictionaries, like *DIL-VOX* and *DL-SM*. To do so, we have selected the lemma *paveo*. This lemma has been selected because of the complexity of its lexicographical treatment, given that it is a sentiment verb - specifically, fear - a type of verb that usually poses difficulties when describing its complements with respect to its meaning.

##### *Diccionario Ilustrado Latín*

**paveo, pāvi** — 2 INTR.: to be afraid (*pavens admiratione*, disconcerted by surprise) || to take fright ¶ TR. to fear (*nec pavent numerare plagas*, and they do not fear counting their wounds).

<p><b>LEMA:</b> paveo, S1 alguien se siente aterrado/atemorizado, S2 alguien tiene miedo de alguien o de algo</p> <p>paveo, es, ere, pavi, - (2ª)</p> <p><b>CATEGORÍA:</b> Verbo</p> <p><b>LETRA:</b> p</p> <p>• <b>MICROESTRUCTURA VERBOS</b></p> <p>TIPO VERBO (según número de argumentos): MONOVALENTE, BIVALENTE</p> <p>* ARGUMENTOS DEL VERBO</p> <p>Numero Argumentos Significado 1: MONOVALENTE</p> <p>Significado 1: alguien se siente aterrado/atemorizado</p> <p>Primer argumento: Nominativo</p> <p>Caracterización Argumental 1: +animado +humano</p> <p>Caracterización Argumental 2: +animado -humano</p> <p>Ejemplo: Caesar pavet</p> <p>Numero Argumentos Significado 2 (Si Existe): BIVALENTE</p> <p>Significado 2: alguien tiene miedo de alguien o de algo</p> <p>Primer argumento: Nominativo</p> <p>Caracterización Argumental 1: +animado +humano</p> <p>Caracterización Argumental 2: +animado -humano</p> <p>Segundo argumento: Acusativo</p> <p>Caracterización Argumental 1: +animado +humano</p> <p>Caracterización Argumental 2: +animado -humano</p> <p>Caracterización Argumental 3: -animado -definido</p> <p>Ejemplo: Agnus lupos pavet</p>	<p><i>Diccionario Latín</i></p> <p><b>paveo, -es, -ēre, pavi, -</b> (2) v.tr./intr. <b>1</b> to be overcome, to be distraught, to be anguished (<i>tremet ille pavetque</i>- he trembles with fear and anguish) <b>2</b> to fear, to be afraid of (<i>pavet agna lupos</i> - the lamb fears the wolves).</p>
--	--

<sup>24</sup> Mariner (1986) argued that perhaps the past form should be taught before the perfect, as it is a tense that displays fewer formal alterations.



Figure 4. The lexicographic item *paveo* in *DDDL*

This is a verb of dubious etymology which has no supine. Both dictionaries specify the conjugation and the transitive and intransitive forms of the verb. However, while *DIL-VOX* clearly distinguishes between transitive and intransitive use, *DL-SM* does not: from the examples, it can be inferred that **1** is the intransitive, while **2** is the transitive, at least in Spanish. However, treating the meanings of “to fear” and “to be afraid of” as transitive requires a previous explanation that justifies the transitivity of the second meaning. *DIL-VOX Latin* is more consistent in the structure of meanings and complements, but the examples used - one of the strong points of learner's dictionaries - can be misleading: *pavens admiratione* is translated as “disconcerted by surprise”, where the sentiment described in the meaning “to be afraid” that precedes the example is not found. As for the translation of the example of the meaning “to fear”, *nec pavent numerare plagas*, “and they are not frightened about counting their wounds”, is again a variation with respect to the meaning given, turning the second Latin argument into the first argument in its Spanish translation. Perhaps a translation like “and they are not afraid of counting their wounds” would have been more consistent.

*DDDL* does not use labels that mark the transitivity or intransitivity of a verbal predicate, given that, as stated before, they can be misleading if they are not consistent with the translation. As for meanings, given that the goal of the dictionary is not to provide all the semantic range of the verb in question, but to provide patterns of use that users can compare to those of other lexical units with similar semantic behaviour, on the basis of the qualitative nature of the dictionary, two meanings have been selected that we understand offer enough information about the behaviour of Latin. The first one refers to the monovalent behaviour of the predicate: a verb with a single argument described as an +animate entity, a logical argument, given that sentiment is proper to animate beings. The morphology of that single argument is also described, nominative in this case. The second meaning appears when the behaviour of the verb is bivalent, taking two arguments: one first argument, in the nominative, +animate (the entity that experiences the sentiment) and a second argument, in the accusative, characterised as +animate or -animate -definite (the sentiment of fear is generated by a human, an animal, or concepts such as war, winter, loneliness, etc.) Each meaning and valence description is followed by an example. This structure is intended to ensure that users understand that both *paveo* and other similar sentiment verbs (*metuo* or *timeo*) have similar behaviour. That is, being verbs that express fear, a single +animate argument in the nominative (the source experiencing fear with no cause) is expected, or else a double argument in the nominative and in the accusative, the latter expressing the entity that causes the sentiment of fear.

We believe that dictionary queries can consolidate knowledge in the user that can be expanded at later stages with other metaphorical meanings or co-locations that extend the user's mastery of the behaviour of lexical units in Latin.

## 5. Conclusions and future work

Both the analysis conducted to establish the type of potential users of the dictionary and their needs and the theoretical linguistic frameworks used to create *DDDL* are innovative in the lexicography for the learning of Latin, at least in Spanish. Other significant contributions of the new *DDDL* dictionary are:

- The proposal of a new way to understand and learn Latin from a conceptual and constructivist point of view based on the theoretical linguistic frameworks of valence theory, Lyons' ontologies, and Fillmore's semantic frame theory.
- The jigsaw puzzle metaphor used as a reasoning and visual element in *DDDL*. In this metaphor, verbal lemmas are the core piece of the sentence meaning construction process and they are represented by triangular shapes with one to three protrusions, depending on whether they are monovalent, bivalent, or trivalent. Nouns, as complementary semantic pieces, have the shape of a pentagon that fits with the protrusion in the verb triangle. The colour of the pentagon also refers to a specific ontology (red, for example, represents an +animate +human unit), which fits the lexical characteristics required by the verbal complement.
- The mechanism of translation of the verbal lemmas which is not done by synonymic equivalent in the target language, but by a semantic-argumental description, based on the users' lexical knowledge of their native language (in this case, Spanish): e.g. the translation of the Latin word “do” is not “to give”, but “someone gives something to someone”.
- The description of the verbal complements, which starts at the semantic level and reaches the morphological level with no syntactic labels when establishing the complementation relationship.
- Its conception as an educational resource that can be easily integrated in e-learning, b-learning, and m-learning training and self-training proposals.

According to the results obtained so far in the experiments, these contributions seem to help learners to understand the new language on the basis of their knowledge of their native language. The learner realizes that on the semantic level, Latin works like his or her mother tongue, on the morphological level it differs in the use of marks to indicate cases and on the syntactic level the learner realizes that the order of the constituents is different, but that does not hinder him or her from being able to understand, and even with a little observation, construct sentences that are not always syntactically correct, but correct and comprehensible from the morpho-syntactic point of view. We have found that this learning path empowers the student and prepares him/her for deeper language learning.



The next step in our work will be to not only continue to test and improve didactic effectiveness of the DDDL, completing the longitudinal experiments that are being conducted, but also to consider how to complete the dictionary so that it can also be useful for learners at higher levels. Finally, we are interested in determining whether the hypotheses on which this dictionary is based apply to other languages.

## 6. Bibliography

- Ágel, V. & Fischer, K. (2010). Dependency Grammar and Valency Theory. In B. Heine & H. Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, Oxford: Oxford University Press, pp. 225-257.
- Bergenholtz, H. & Tarp, S. (2002). Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen. In *Lexicographica*, 18, Tübingen: Max Niemeyer Verlag, pp. 253-263.
- Bergenholtz, H. & Tarp, S. (2003). On H.E. Wiegand's recent discovery of lexicographic functions. In *Hermes*, 31, pp. 171-196.
- Blázquez, A. (1997). *Diccionario Latino-Español, Español-Latino*. Barcelona: Ramón Sopena.
- Boehm, B. (1986) A Spiral Model of Software Development and Enhancement". In *ACM SIGSOFT Software Engineering Notes*, 11(4), pp.14-24.
- Cowie, A. P. (1999). *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.
- Dik, S. C. (1997). *The Theory of the Functional Grammar. Part 1. The Structure of the Clause*. Berlin: Mouton de Gruyter.
- CTN 71/SC 36. (2017). Fernández-Pampillón Cesteros, Ana María y Porras Guardo, Arturo de y González Serrano, Ángel Luis y García Villalobos, Julián y Moreno López, Lourdes y Domínguez Romero, Elena y Armas Ranero, Isabel de y Rodrigo, Covadonga y Sarasa Cabezuolo, Antonio y Arús Hita, Jorge y Sierra Rodríguez, Jose Luis y Cabanilles Gomar, Juan Pedro y Vizoso Martín, Clara María y González Maroto, Yolanda y Camacho Fernández, Patricia y Iglesias Vázquez, Pedro Luis y Pons Betrián, Daniel y Castro Soriano, Luis de y Delgado Leal, Jose Luis (2017) *Herramienta de evaluación de la calidad de los Materiales Educativos Digitales: perfiles de aplicación del profesor y del alumno*. In UNE 71362 Calidad de los materiales educativos digitales. 35.240.90 / Aplicaciones de las tecnologías de la información en educación. AENOR, Madrid. España, pp. 114-131. (English version at: [https://eprints.ucm.es/45338/13/ANEXOF\\_english.pdf](https://eprints.ucm.es/45338/13/ANEXOF_english.pdf))
- Favarin, S. (1979). Per un vocabolario valenziale dei verbi latini. In *Revista di Studi Classici*, 27, pp. 454-470.
- Fernández-Pampillón, A. (2017). UNE71362, calidad de materiales educativos digitales. <https://portal.aenormas.aenor.com/revista/329/une-71362.html>. English draft version at: [https://eprints.ucm.es/45088/7/articulo\\_presentacion\\_UNE71362\\_eng.pdf](https://eprints.ucm.es/45088/7/articulo_presentacion_UNE71362_eng.pdf)
- Fernández-Valmayor, A. - Fernández-Pampillón, A. M.<sup>a</sup> - Varadero Software Factory, VSF. (2013). *Guía de Gestión Oda 2.0*. (English version also available: "Oda 2.0 Administration Guide") Documento Técnico. Available at: [https://eprints.ucm.es/20263/12/GUIA\\_ODA\\_v2\\_5%20%282%29.pdf](https://eprints.ucm.es/20263/12/GUIA_ODA_v2_5%20%282%29.pdf)
- Fillmore, Ch. J. (1968). The case for case. In E. Bach & R. Harms (eds.) *Universals in Linguistic Theory*. New York: Holt, Rinehart & Winston, pp. 1-91.
- Fillmore, Ch. J. (1971). Some Problems for Case Grammar. In R.J. O'Brien (ed.) *Report of the Twenty-Second Annual Round Table Meeting on Linguistics and Language Studies*, Washington, D.C.: Georgetown University Press, pp. 35-56.
- Gardner, R. C. (2004). *Attitude/Motivation Test Battery: International AMTB Research Project*. London: University of Western Ontario.
- García Ferrer, M. (2013). Análisis contrastivo de las herramientas lexicográficas para enseñar y aprender latín. In E. Casanova & C. Calvo. *Actas del XXVI Congreso Internacional de Lingüística y Filología Románicas. (Valencia, 6-11 de septiembre de 2010)*. Berlin/Boston: De Gruyter, pp. 171-182.
- Gouws, R. H. (2011). Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries. In P.A. Fuenes-Olivera & H. Bergenholtz (eds.) *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London: Continuum, pp. 17-29.
- Happ, H. (1976). *Grundfragen ciner Dependenz-Grammatik des Lateinischen*. Göttingen: Vandenhoeck & Ruprecht.
- Heuberger, R. (2016). Learner's Dictionaries. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp.45-23.
- Heuberger, R. (2018). Dictionaries to assist teaching and learning. In P. A. Fuertes-Oliveira (ed.) *The Routledge Handbook of Lexicography*. London and New York: Routledge, pp. 300-316.
- Jackson, H. (2002). *Lexicography: An Introduction*. London and New York: Routledge.
- Lyons, J. (1977). *Semantics*. London: Cambridge University Press.
- Mariner, S. (1986). Fundamentos científicos para una enseñanza no compartimentada de las lenguas clásicas (Apéndice: Bachillerato unificado y polivalente. Iniciación a la lengua latina: metodología; programa). In A. Alvar (coord.) *Minerva Restituta: 9 lecciones de Filología Clásica*. Madrid: Universidad de Alcalá de Henares.
- Márquez, M. (under review). Algunas consideraciones sobre la lexicografía de aprendizaje de latín en España.
- Miller, J. (2018). Learner's dictionaries of English. In P. A. Fuertes-Oliveira (ed.) *The Routledge Handbook of Lexicography*. London and New York: Routledge, pp. 353-366.
- Rubio, L. & González Rolán, T. (1990). *Nueva Gramática Latina*. Madrid: Coloquio Editorial.
- Tarp, S. (2006), Lexicografía de aprendizaje, *Cuadernos de Tradução*, 18(2), pp. 295-317.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*, Paris: Klincksieck.



UNE 71362:2020 Calidad de los materiales educativos digitales/Quality of digital educational materials. CTN 71/SC 36 - TECNOLOGÍAS DE LA INFORMACIÓN PARA EL APRENDIZAJE. <https://www.aenor.com/normas-y-libros/buscador-de-normas/une/?c=N0063263>

Wiegand, H. E. & Fuentes, M. T. (2010). *Estructuras lexicográficas. Aspectos centrales de una teoría de la forma del diccionario*. Granada: Ediciones Traganto.

### Acknowledgements

*DDDL* was created thanks to four Educational Innovation Projects (PIE 193:2016, PIE 269:2017, PIE 164:2018 and PIE 245:2019) funded by the Universidad Complutense de Madrid. It is part of the Computational Linguistics projects in the Department of General Linguistics and the ILSA research group (<http://ilsa.fdi.ucm.es/>) at Universidad Complutense of Madrid. We thank the members of the various Innovation Projects and the ILSA Group for their invaluable collaboration. We also thank the Spanish Ministry of Science and Innovation for funding the project “Creation, exploitation, and transformation of educational object repositories in specialised domains”, with reference TIN2017-88092-R.



# Ενδογλωσσική και διαγλωσσική προσέγγιση της συνωνυμίας. Συγκριτική μελέτη λογοτεχνικών μεταφράσεων με δίγλωσσα και μονόγλωσσα λεξικά.

Ροντογιάννη Α.

Πανεπιστήμιο Θεσσαλίας, Ελλάδα

## Abstract

Η παρούσα μελέτη προσεγγίζει το επίθετο σύμφωνα με τη ψυχομηχανική / ψυχοσυστηματική θεωρία του G. Guillaume και εστιάζει στις ιδιαιτερότητες τόσο της φύσης του όσο και της λειτουργίας του. Με βάση το ίδιο θεωρητικό πλαίσιο της ψυχομηχανικής, όπως διατυπώθηκε στη σημασιολογική και λεξικολογική της διάσταση, από τους Pottier, Martin και Picoche, προσεγγίζονται και τα ζητήματα που παρατηρούνται κατά τη μελέτη της συνωνυμίας του. Στόχος μας είναι να ερευνήσουμε τη συμβολή της μετάφρασης στη διαβάθμιση και οριοθέτηση της συνωνυμίας του επιθέτου. Συγκρίνοντας και αντιπαραβάλλοντας τις λογοτεχνικές με τις λεξικογραφικές μεταφράσεις, αλλά και τους ορισμούς των μονόγλωσσων λεξικών, ερευνάται η δυνατότητα άντλησης παραδειγμάτων από έργα έγκριτων μεταφραστών ως μέσο εμπλουτισμού των λημμάτων των δίγλωσσων λεξικών.

**Keywords:** Ψυχομηχανική, Λεξικολογία, Σημασιολογία, Λεξικογραφία, Μετάφραση

## 1 Εισαγωγή

Η εργασία αυτή αναπτύσσεται γύρω από τρεις κεντρικούς άξονες. Στο πρώτο μέρος μελετάμε το επίθετο σύμφωνα με τη θεωρία του Gustave Guillaume. Τα επίθετα, αν και εντάσσονται στις τρεις κύριες λεξικές κατηγορίες που αναγνωρίζει ο Chomsky (1975), αφού θεωρούνται μαζί με τα ονόματα ως οι περισσότερο λεξικές λέξεις, παρουσιάζουν κάποιες ιδιαιτερότητες τόσο από τη φύση και τη λειτουργία τους όσο και ως προς τη μελέτη των σχέσεων συνωνυμίας που αναπτύσσουν. Σε αυτές τις ιδιαιτερότητες θα εστιάσουμε στο δεύτερο μέρος μελετώντας τη συνωνυμία του επιθέτου υπό το πρίσμα της ψυχομηχανικής / ψυχοσυστηματικής θεωρίας του Guillaume.<sup>1</sup> Στο τρίτο μέρος επιχειρούμε μια διαγλωσσική μελέτη του φαινομένου αντιπαραβάλλοντας λογοτεχνικές μεταφράσεις με δίγλωσσα γαλλοελληνικά λεξικά. Έχοντας θέσει ως στόχο τη συμβολή της μετάφρασης στη μελέτη της συνωνυμίας, αντλούμε τα παραδείγματά μας από το έργο του Victor Hugo, *Les Misérables* και πέντε από τις αντιπροσωπευτικότερες μεταφράσεις του, ξεκινώντας από την πρώτη του Σκυλίτση (1863) έως και αυτή του Λυκούδη (2000). Παράλληλα με τη διαχρονικότητα του έργου του μεγάλου συγγραφέα που αναδεικνύεται και μέσω της περιγραφικής του δεινότητας, αναδύεται και η διαχρονικότητα των γλωσσικών φαινομένων.

## 2 Ψυχομηχανική προσέγγιση του επιθέτου

O Guillaume (1883-1960) είναι ένας από τους σημαντικότερους γάλλους γλωσσολόγους, με πλούσιο συγγραφικό έργο. Η πρώτη του μονογραφία εκδόθηκε το 1911, πριν τα *Μαθήματα Γενικής Γλωσσολογίας* του Saussure και το τελευταίο του άρθρο το 1958, μετά τις *Συντακτικές Δομές* του Chomsky. Ο ίδιος φημίζεται για τον ερμητισμό της ορολογίας που εισήγαγε, κάτι που συχνά συνδέεται με τον βαθμό δυσκολίας των θεωριών που διατύπωσε. Στην πραγματικότητα η δυσκολία οφείλεται –εν μέρει– στο γεγονός ότι, ενώ η θεωρία του τοποθετείται στον «αντίποδα της παραδοσιακής γλωσσολογίας» (Αναστασιάδη-Συμεωνίδη 1978: 311), ανατρέχει στην παραδοσιακή ορολογία για να εκφράσει νέες ιδέες. Στο έργο του οι μελετητές του βρίσκουν στοχασμούς που βασίζονται, εμβαθύνουν και αναδιατυπώνουν στοιχεία της αριστοτελικής θεωρίας (Stefanini 1992). Σύμφωνα με την ψυχομηχανική / ψυχοσυστηματική θεωρία του Guillaume,

η αληθινή πραγματικότητα ενός γλωσσικού σημείου δεν είναι οι ποικίλες και φευγαλέες σημασίες και αποχρώσεις που προκύπτουν από τη χρήση του, αλλά η αναλλοίωτη πράξη της σκέψης που υπάρχει πριν από αυτήν. Ο Guillaume προχώρησε έτσι στη γνώση των φαινομένων που βρίσκονται στο επίπεδο βάθους, όπου στηρίζεται η κατασκευή (και όχι δομή) της γλώσσας (Αναστασιάδη-Συμεωνίδη 1978: 311).

Στην ψυχομηχανική, η βασική έννοια της αναφοράς σχετίζεται με την απολύτως γενική κίνηση στο λόγο, σύμφωνα με την οποία παντού και πάντα υπάρχει φορέας σημασίας και αναφορά αυτού του φορέα σε μια βάση. Αυτό ερμηνεύεται ως εξής: η λέξη, σε κάθε περίπτωση, είναι φορέας σημασίας και αναζητά μια βάση στην οποία θα αναφερθεί (*DTSL*<sup>2</sup>, 1996: 229). Αυτός ο μηχανισμός που της αποδίδεται και στον οποίο οφείλεται, είναι κατά ένα μεγάλο μέρος ο προορισμός της, καθώς μέσω αυτού κατηγοριοποιείται και γίνεται αυτό που ονομάζουμε μέρος του λόγου. Όταν μιλάμε γίνεται πάντα λόγος για κάτι, που είναι αυτό για το οποίο μιλάμε, δηλαδή το υποκείμενο, στο οποίο αναφέρεται το κατηγορήμα, αυτό που λέμε. Το υποκείμενο είναι η βάση της σκέψης. Είναι αυτό στο οποίο στηρίζεται ο φορέας της σημασίας (το κατηγορήμα), όπως τον έχει καθορίσει η ίδια η σκέψη στη γλώσσα. Το κατηγορήμα προσδιορίζεται σε ένα πρώιμο στάδιο, στο επίπεδο της γλώσσας και όχι στο επίπεδο του λόγου, ενώ η αναφορά του φορέα στη βάση, ή αλλιώς του

<sup>1</sup> Για τη θεωρία του Guillaume, βλ. Αναστασιάδη –Συμεωνίδη, Α. (1978).

<sup>2</sup> Boone, A. et Joly, A. (1996) *Dictionnaire Terminologique de la Systématique du Langage*.



κατηγορήματος στο υποκείμενο, αποτελεί αντιθέτως όψιμο στάδιο, στο επίπεδο του λόγου (Vassant 1998 & 2005).

Ως συνέπεια αυτής της διάκρισης ανάμεσα στον προσδιορισμό του υποκειμένου, με το οποίο συνδέεται το κατηγορήμα στο επίπεδο της γλώσσας, και στην αναφορά του κατηγορήματος στο υποκείμενο στο επίπεδο του λόγου, το ίδιο του επιθέτου είναι ότι δεν αναφέρεται στον εαυτό του. Είναι προορισμένο να βρει μια βάση, ένα υποκείμενο έξω από το ίδιο, εκτός του πεδίου της σημασίας του. Όπως το ουσιαστικό, έτσι και το επίθετο φέρει μια σημασία, αλλά ενώ στο ουσιαστικό ο φορέας της σημασίας, το κατηγορήμα, δηλώνει ήδη από τη γλώσσα τη φύση του υποκειμένου – γεγονός το οποίο αποτελεί αρχή της αναφοράς – στο επίθετο το κατηγορήμα δεν δηλώνει τη φύση του υποκειμένου στο επίπεδο της γλώσσας: η επιλογή του υποκειμένου πραγματοποιείται κατά την πραγμάτωση του λόγου. Έτσι, το επίθετο π.χ. *ωραίος* δε φανερώνει σε καμία περίπτωση από μόνο του σε τι υποκείμενο αναφέρεται. Η αναφορά του δεν ξεκινά, όπως αυτή του ουσιαστικού, από τη γλώσσα και έτσι μπορούμε να χαρακτηρίσουμε ως *ωραία* τόσο έμψυχα όντα όσο και άψυχα αντικείμενα ή αφηρημένες έννοιες. Αυτός ο μηχανισμός, σύμφωνα με τον οποίο η φύση του υποκειμένου δεν είναι προβλέψιμη από το ίδιο το κατηγορήμα στο επίπεδο της γλώσσας, ονομάζεται εξωτερική αναφορά. Το επίθετο είναι κατηγορηματικό μέρος του λόγου που χαρακτηρίζεται από εξωτερική αναφορά πρώτου βαθμού, καθώς το ουσιαστικό στο οποίο αναφέρεται, χαρακτηρίζεται από εσωτερική αναφορά. Αυτή ακριβώς η φύση του επιθέτου και η λειτουργία του στο επίπεδο της πραγμάτωσης του λόγου είναι που καθιστά ιδιαίτερη την εξέτασή του, ειδικά στο πλαίσιο της εννοιακής σχέσης της συνωνυμίας.

### 3 Η εννοιακή σχέση της συνωνυμίας

Η σύνδεση των θεωριών του Guillaume με θέματα που άπτονται της λεξικής σημασιολογίας είναι έμμεση. Στα μαθήματά του δήλωνε πως «δεν υπάρχουν απόλυτα συνώνυμα στη γλώσσα» (Guillaume, 2018). Βασική διάκριση της θεωρίας του, που εστιάζει στο επίπεδο βάθους της γλώσσας, είναι ο διαχωρισμός του *δυνάμει σημαινόμενου* στη γλώσσα, και του *ενεργεία σημαινόμενου*, στο λόγο. Η σημασία οργανώνεται στη γλώσσα, στο επίπεδο του *δυνάμει σημαινόμενου* το οποίο αποτελεί σταθερό σημείο αναφοράς και από το οποίο αντλεί το εκφώνημα σε συνάρτηση με τα εκάστοτε περιβάλλοντα του περικειμένου. (Thavaud-Piton, 2016: 143). Ο όρος *δυνάμει σημαινόμενο* εμφανίζεται σπανιότερα στα κείμενα του Guillaume, αλλά συναντάται πολύ συχνά στα έργα των γλωσσολόγων που εμπνεύστηκαν από τις θεωρίες του. Ο ίδιος ο Guillaume εξηγεί ότι η σχέση του *δυνάμει σημαινόμενου* με το σημείο είναι μόνιμη και το καθιστά σημαίνον, σε αντίθεση με το *ενεργεία σημαινόμενο* το οποίο αντιστοιχεί στην στιγμιαία αξία της χρήσης του σημείου κατά την πραγμάτωση του λόγου (DTSL, 1996: 382). Έτσι, το *δυνάμει σημαινόμενο* αποτελείται από το σύνολο των *ενεργεία σημαινόμενων*, όπως αυτά εκδηλώνονται στον λόγο. Στην ίδια κατεύθυνση, η Picoche (1992β:8) ορίζει το ρόλο του σημαίνοντος ως διαμεσολαβητικό ανάμεσα στο *δυνάμει* και στο *ενεργεία σημαινόμενο* με το τελευταίο να δηλώνει «κάθε έννοια ή κάθε εννοιακή κατασκευή που επιτρέπει μια λογική ταξινόμηση αναδεικνύοντας διαφορετικές σημασίες» (Picoche, 1992α:78). Η συνωνυμία αποτελεί φαινόμενο που εκδηλώνεται στο επίπεδο του λόγου, αφού αφορά δύο σημασιακά αποτελέσματα που απορρέουν από δυο *δυνάμει γλωσσικά στοιχεία*.

#### 3.1 Ορισμοί

Όσον αφορά τους ορισμούς της συνωνυμίας, παρατηρούμε πως οι διαφορές τους περιορίζονται στο επίπεδο της ορολογίας: σχετική ή μερική συνωνυμία, λογικοπροτασιακή συνωνυμία, παρασυνωνυμία, πλησιωνυμία, είναι κάποιοι από τους όρους που αποδίδουν τις διαβαθμίσεις στο εσωτερικό της, καθώς τα διαφοροποιητικά κριτήρια δεν παρουσιάζουν σημαντικές αποκλίσεις από τη μια γλωσσολογική προσέγγιση στην άλλη. Το ίδιο παρατηρούμε και με την εφαρμογή των κριτηρίων ανίχνευσης, τόσο της αμοιβαίας λογικής συνεπαγωγής, όσο και –κυρίως– του βαθμού φυσικότητας, μέσω της μεθόδου της αλληλοϋποκαταστασιμότητας (Ξυδόπουλος, 2016: 129) ή αμφίδρομης υποκαταστασιμότητας (Μότσιου, 1994: 185).

Με τον όρο συνωνυμία αναφερόμαστε στην εννοιακή σχέση δυο λέξεων οι οποίες θεωρούνται σημασιολογικά ισοδύναμες μεταξύ τους. Πιο συγκεκριμένα, ως συνώνυμα ορίζονται οι σημασιολογικά ισοδύναμες λεξικές μονάδες, με διαφορετικά σημαίνοντα, τα οποία ανακαλούν στο μυαλό μας το ίδιο σημαίνόμενο (Μότσιου, 1994:188). Απαραίτητη προϋπόθεση για τη μελέτη του φαινομένου της συνωνυμίας αποτελεί το περικείμενο και ο ρόλος που επιτελεί στην αποσαφήνιση της έννοιας της λέξης και στην ενίσχυση ή απόρριψη της υπόθεσης της συνωνυμίας καθώς και στην οριοθέτηση και διαβάθμισή της. Ως εκ τούτου, η συνωνυμία εξετάζεται κυρίως τόσο σε φραστικό και σε προτασιακό επίπεδο, όσο και σε κειμενικά είδη μεγαλύτερης έκτασης γενικότερα.<sup>3</sup> Στην παρούσα μελέτη θα εστιάσουμε στη σχέση της συνωνυμίας των περιγραφικών ή ποιοτικών επιθέτων, των επιθέτων δηλαδή που περιγράφουν ή αξιολογούν αυτό που προσδιορίζουν. Τα επίθετα σημασιολογικά δηλώνουν την ποιότητα ή την ιδιότητα ενός ουσιαστικού. Στην πρόταση το επίθετο καταλαμβάνει το συντακτικό ρόλο του προσδιορισμού και δίνει πρόσθετες πληροφορίες για τους όρους με τους οποίους βρίσκεται σε σχέση. Στα παραδείγματα που έχουμε καταγράψει και κατηγοριοποιήσει συναντάται τόσο ως επιθετικός όσο και ως κατηγορηματικός (ονοματικός) ομοιόπρωτος προσδιορισμός αναλόγως της ιδιότητας (μόνιμης ή προσωρινής) που προσδίδει στο ουσιαστικό που χαρακτηρίζει.

#### 3.2 Κατανομική ανάλυση και οργάνωση πεδίων

Με αφετηρία τον τίτλο του μυθιστορήματος (*Les Misérables* / *Οι Αθλιοί*) συγκεντρώσαμε, τόσο από το πρωτότυπο κείμενο, όσο και από τις μεταφράσεις του τα επίθετα και τα συνώνυμά τους και δημιουργήσαμε το λεξικό -

<sup>3</sup> Στη γλώσσα της λογοτεχνίας συναντούμε και συνώνυμα με την ευρύτερη έννοια του όρου. «Paris est synonyme de Cosmos» (*Les Misérables*, 2000: 605), «το Παρίσι είναι συνώνυμο του κόσμου» (Λυκούδης: 802), «Aube et resurrection sont synonymes» (*ibid.*, 1266), «Αυγή και επανάσταση είναι συνώνυμα», (Λυκούδης: 1681), «Boue est synonyme de honte» (*ibid.*, 1319), «Η λάσπη είναι συνώνυμο της ντροπής» (Λυκούδης: 1748).



σημασιολογικό πεδίο της αθλιότητας (*champ lexical de la misère*). Η εξέταση των κατανομών, των γλωσσικών δηλαδή περιβαλλόντων μέσα στα οποία μπορεί να εμφανιστεί μια λεξική μονάδα, πραγματοποιείται τόσο στον παραδειγματικό όσο και στο συνταγματικό άξονα. Κατά τη σημασιολογική προσέγγιση, η λέξη εξετάζεται σε συνάρτηση με την κατανομή, τα περιβάλλοντα και τα παραδείγματα μέσα στα οποία εμφανίζεται, προκειμένου να καταγραφεί στο εννοιολογικό πεδίο. Σύμφωνα με την ονομασιολογική<sup>4</sup> μέθοδο, οριοθετούμε ένα εννοιολογικό πεδίο στο οποίο αντιστοιχεί ένα δομημένο σύνολο επιθέτων (Soutet, 1995: 269) και θα αποτελέσει το λεξιλογικό μας πεδίο. Η προσέγγιση αυτή βασίζεται στη θεματική ταξινόμηση, ξεκινάει από την ουσία του περιεχομένου για να περιγράψει το γλωσσικό σημείο, σε αντίθεση με τη σημασιολογική που ξεκινάει με το γλωσσικό σημείο και το συνδέει με την έννοια (την ουσία του περιεχομένου). Η ιεραρχική οργάνωση του λεξιλογίου σε πεδία συμβάλλει στη διαμόρφωση μιας σφαιρικής εικόνας των διαφορετικών κατανομών των ισοδύναμων σημασιολογικών χαρακτηριστικών και προσφέρει το υλικό που χρειαζόμαστε για να προχωρήσουμε στην ανάλυση σε ελάχιστες σημασιακές μονάδες, αλλά έχει δεχτεί πολλές κριτικές όσον αφορά στην αποτελεσματικότητά της. Οι συνώνυμες λεξικές μονάδες θεωρούνται μορφήματα των οποίων η κατανομή μπορεί να είναι ταυτόχρονα ισοδύναμη και συμπληρωματική, άρα επικαλυπτική καθώς πληρούνται τα κριτήρια της υποκαταστασιμότητας τους σε κάποια, αλλά όχι σε όλα τα περιβάλλοντα. Ενώ οι σχέσεις της συνωνυμίας με την κατανομική ανάλυση δεν αμφισβητούνται, η φύση και τα όρια των σχέσεων αυτών έχουν αμφισβητηθεί. Ο Slatka παρατηρεί πως « η κατανομική ανάλυση είναι αποτελεσματική για την οργάνωση των λεξικών μονάδων σε πεδία, αλλά όχι πέρα από αυτό. Η ανάλυση σε ελάχιστα σημασιακά συστατικά είναι το επόμενο στάδιο » (D. Slatka, 1971: 97). Η αποτελεσματικότητα της μεθόδου αυτής είναι αποδεδειγμένη, παρά τις κριτικές που έχει δεχτεί. Ο Martin, αξιολογεί θετικά τον συνδυασμό των τεχνικών αυτών :

*η δομική σημασιολογία αξιοποιώντας τις θεωρίες των Trier, Weisgerber, Migliorini και άλλων γύρω από την έννοια του πεδίου, συνδυάζοντας τις τεχνικές της κατανομικής ανάλυσης με τα δεδομένα της ανάλυσης σε ελάχιστα σημασιακά συστατικά, την οποία πρώτος πρότεινε ο Hjelmslev, κατάφερε να λύσει πολλά προβλήματα της σημασιολογίας. Οι λεξικές μονάδες με συγγενική σημασιολογική σχέση συγκεντρώνονται στον παραδειγματικό άξονα και δημιουργούν ένα σημασιολογικό πεδίο. Ακολουθεί η ανάλυσή τους σε ελάχιστα σημασιακά συστατικά (R. Martin, 1976: 9).*

### 3.3 Το συνωνυμικό πεδίο των Αθλίων

Στο μυθιστόρημα, στην κατηγορία των αθλίων συγκαταλέγονται όχι μόνο οι οικονομικά αδύναμοι, αλλά κυρίως οι ηθικοί αυτουργοί των κοινωνικών ανισοτήτων και αυτοί που εκμεταλλεύονται όσους έχουν στερηθεί το δικαίωμα στη μόρφωση : « Τα λάθη των γυναικών, των παιδιών, των υπηρετών, των αδυνάτων, των ενδεών και των αμαθών είναι λάθη των συζύγων, των πατέρων, των δασκάλων, των αφεντικών, των δυνατών, των πλουσίων και των σοφών» (Λυκούδης:31), « Η πραγματική ανθρώπινη διαίρεση είναι τούτη: οι φτωχοί και οι σκοτεινοί. [...] Μαθαίνω να διαβάζω σημαίνει ανάβω φως» (Λυκούδης:1337). Στο σκοτάδι που επικρατεί στον κόσμο, ο Β.Ουγκώ βλέπει φως μόνο μέσα από την ψυχική καλλιέργεια, την πνευματική και διανοητική εξύψωση. Αλλά και οι αναφορές στην ανέχεια, η οποία οδηγεί τον άνθρωπο σε ακραίες καταστάσεις, είναι πολλές και εκτενείς. Ο συγγραφέας περιγράφει λεπτομερώς πώς εκδηλώνονται όλα τα προβλήματα που απορρέουν από αυτή, ξεκινώντας από τα ίδια τα άτομα, την εξωτερική τους εμφάνιση, τις ενδυματολογικές και διατροφικές τους συνθήκες, τις συνθήκες διαβίωσής τους, τους χώρους κατοικίας τους και το εξωτερικό τους περιβάλλον, τις καθημερινές τους ανάγκες, τόσο σε υλικά όσο και σε πνευματικά αγαθά, τον χαρακτήρα τους και τις πνευματικές τους αναζητήσεις, αλλά και τις (δια)νοητικές τους δυνατότητες, από την επιθυμία τους για μόρφωση ως τα επικοινωνιακά τους μέσα και την χρήση της γλώσσας.

#### 3.3.1 Το αρχισήμημα : «αρνητική συναισθηματική κατάσταση»

Προκειμένου να περιγράψει όλες τις εκφάνσεις της αθλιότητας, ο Β.Ουγκώ κάνει χρήση μιας μεγάλης γκάμας επιθέτων των οποίων ο «σημικός πυρήνας» είναι φορέας αρνητικού περιεχομένου καθώς κατά την ανάλυση των επιθέτων παρατηρούμε ότι αυτά απαρτίζονται από αρνητικά σημασιακά χαρακτηριστικά. Ο «σημικός πυρήνας [...] παρουσιάζεται ως ένα διαρκές σημικό ελάχιστο, ως μια σταθερά [...] ο συνδυασμός του σημικού πυρήνα με τα συμφραστικά σήματα προκαλεί [...] το σημασιακό αποτέλεσμα που ονομάσαμε σήμημα» (Greimas, 1995: 45). Ο Greimas, όπως και οι Pottier και Picoche χρησιμοποιούν τους όρους σήμα – σήμημα – αρχισήμημα για να δηλώσουν τα σημασιακά χαρακτηριστικά. Δύο ή περισσότερες λέξεις που ανήκουν στο ίδιο μέρος του λόγου θεωρούνται συνώνυμες όταν έχουν το ίδιο σήμημα (Picoche, 1992-α: 99). Το αρχισήμημα «αρνητική συναισθηματική κατάσταση», το οποίο συγκεντρώνει το σύνολο των σημασιακών χαρακτηριστικών, και διατηρείται αμετάβλητο σε περίπτωση ουδετεροποίησης, δεν εξετάζεται σε συνάρτηση μόνο με την ίδια την κατάσταση ή το άτομο που νοιώθει έτσι, αλλά και με αυτό που μπορεί να προκαλέσει τέτοια κατάσταση. Τα επίθετα που οργανώνονται γύρω από το αρχισήμημά μας είναι, με αλφαβητική σειρά:

*« abject, abominable, affreux, difforme, fétide, hideux, horrible, ignoble, immonde, infâme, infect, laid, malpropre, malsain, mauvais, méchant, mesquin, minable, misérable, odieux, sinistre, sordide, terrible, véneneux, venimeux, vil, vilain ».*

Λαμβάνοντας υπόψη τα παραπάνω, είμαστε σε θέση να προχωρήσουμε σε υποδιαιρέσεις στο εσωτερικό του αρχισήμηματος. Κατά την ανάλυση των επιθέτων σε ελάχιστα σημασιακά χαρακτηριστικά, εστιάζουμε αρχικά στα διαφοροποιητικά χαρακτηριστικά τους και στις διαφορετικές κατανομές τους, όπως αυτές προέκυψαν από την

<sup>4</sup> Βλ. Μότσιου, 1994: 175: «Ονομασιολογικό πεδίο ή εννοιολογική κατηγορία», και Ξυδόπουλος, 2016: 303 : «ονομασιοκεντρική ή ονομαστική προσέγγιση»



κατανομική ανάλυση και τα λεξικά σημασιολογικά πεδία. Το κριτήριο της υποκαταστασιμότητας σύμφωνα με το οποίο κάποια επίθετα μπορούν να αντικατασταθούν από κάποια άλλα με τα οποία μοιράζονται κοινά σημασιακά συστατικά (ή σήματα), θα επικυρώσει στη συνέχεια την αρχική μας υπόθεση περί συνωνυμίας.

### 3.3.2 Ανάλυση σε σημασιακά χαρακτηριστικά

Ήδη από την παραπάνω περιγραφή της εσωτερικής οργάνωσης των πεδίων διακρίνουμε πως το αρχισήμημα *αρνητική συναισθηματική κατάσταση* χαρακτηρίζει τρία σημασιακά πεδία: [ΑΙΣΘΗΣΕΙΣ], [ΣΥΝΑΙΣΘΗΜΑ] και [ΗΘΟΣ]. Για τη διάκριση των εννοιών στο εσωτερικό των πεδίων ακολουθήσαμε τις σημασιολογικές υποδιαίρεσεις του *Θησαυρού της Γαλλικής Γλώσσας* (*Trésor de la Langue Française*). Μέσω της λεξικής αποσύνθεσης των λέξεων σε σύνολα σημασιακών χαρακτηριστικών (Ξυδόπουλος, 2017: 153) ξεχωρίσαμε τα διαφοροποιητικά χαρακτηριστικά, τα οποία αποτελούν «σημάδι» της αντίθεσης και, με βάση τις κοινές ιδιότητες που εντοπίσαμε, χωρίσαμε τα επίθετα του αρχισήμηματος στα διαφορετικά σημασιακά πεδία. Έτσι, στο ένα σημασιακό πεδίο συγκεντρώνονται τα επίθετα των οποίων ο σημικός πυρήνας, οργανώνεται γύρω από το ευρύτερο πεδίο των αισθήσεων [ΑΙΣΘΗΣΕΙΣ] και περιγράφουν εξωτερικά χαρακτηριστικά και καταστάσεις που προκαλούν *αρνητική συναισθηματική κατάσταση*, όπως *ασχήμια*, *βρώμα* κ.α. Στο δεύτερο, τα ίδια επίθετα χρησιμοποιούνται σε άλλα παραδείγματα για να περιγράψουν συναισθήματα [ΣΥΝΑΙΣΘΗΜΑ] όπως *δυσaréσκεια*, *απέχθεια*, *μίσος*, *στεναχώρια*, *θλίψη*, *δυστυχία*, *αγωνία*, *φόβο*, *τρόμο*, ενώ στο τρίτο [ΗΘΟΣ], περιγράφουν άτομα, καταστάσεις, ιδιότητες, σκέψεις, με γνώμονα ηθικά και πνευματικά χαρακτηριστικά.

## 4 Οι μεταφράσεις των συνώνυμων επιθέτων

« Le traducteur est un peseur perpétuel d'acceptions et d'équivalents. Pas de balance plus délicate que celle où on met en équilibre les synonymes ». (V. Hugo, [*Les Traducteurs*], p. 632)

Μελετώντας τα είκοσι επτά αυτά επίθετα στα διάφορα περιβάλλοντα όπου εμφανίζονται, παρατηρούμε ότι στο εσωτερικό των σημασιακών πεδίων μπορούμε να εφαρμόσουμε το κριτήριο της υποκαταστασιμότητας ενός επιθέτου με άλλα χωρίς να επηρεάζονται οι συνθήκες αληθείας της έκφρασης. Η σύγκριση του πρωτότυπου κειμένου με τις μεταφράσεις του επιβεβαιώνει τη συνωνυμία των όρων και στις δυο γλώσσες, με ποικίλους τρόπους:

1) αρχικά παρατηρούμε, τόσο σε διαχρονική όσο και σε συγχρονική διάσταση, απόλυτη ομοφωνία των μεταφραστών ως προς την επιλογή συγκεκριμένων όρων κατά την απόδοση όρων στα ελληνικά:

Hugo	1457	ce qui est <i>affreux</i> , c'est de mourir sans la voir
Σκυλίτσης	1248	<i>φρικτόν</i> είνε
Αυγέρης s	368 / 536	τό <i>φρικτό</i> είναι πού
Σκουλούδης	1436	τό <i>φριχτό</i> είναι πού
Λυκούδης	1926	Το <i>φριχτό</i> είναι να

Παράδειγμα 1.

2) Μεγαλύτερο όμως ενδιαφέρον παρουσιάζουν οι αποκλίσεις στην επιλογή όρου μεταξύ των μεταφραστών. Αυτά τα παραδείγματα είναι εξίσου πολλά και αποτελούν ένα επιπλέον επιχείρημα ως προς την ύπαρξη της συνωνυμίας. Κάποια από αυτά τα παραδείγματα εντάσσονται στην κατηγορία των πλησιώνυμων (Ξυδόπουλος, 2017:132), αφού παρουσιάζουν: α) γειτνίαση των σημασιών δυο ή και περισσότερων λέξεων σε μια κλίμακα διαβάθμισης:

Hugo	683	L'homme est mauvais, l'homme est <i>difforme</i>
Αυγέρης	133 / 146	Ό άνθρωπος είναι κακός, [είναι] <i>άσ[χ]κημος</i>
Σκουλούδης	677	Ό άνθρωπος είναι κακός [...] είναι <i>κακομούτσουνος</i>
Λυκούδης	1926	Κακός ο άνθρωπος καί <i>ασουλούπωτος</i>

Παράδειγμα 2.

Στο ίδιο επίθετο, σε άλλα παραδείγματα συναντούμε και τα σύνθετα *άμορφος*, *δύσμορφος*, *κακόμορφος*, *παραμορφωμένος*, *κακόσχημος*, *ασ(κ)χημομούρης*, *ασ(κ)χημόμουτρο*, αλλά και *κακοσουλούπωτος*, *κακοφτιαγμένος*, *κακοτράχαλος*, *ακατασκεύαστος*, *άγαρμος* και *σακατεμένος*. Εκτός από το *άσχημος*, κανένα από τα υπόλοιπα επίθετα δεν περιλαμβάνεται στις μεταφράσεις των λεξικών:

- Ηπίτης: *δύσμορφος*, *δυσειδής*, *ειδεχθής*, *στρεβλός*

- Βαρβάτης: *δύσμορφος*, *δυσειδής*, *άσχημος*

- Kauffmann: *δύσμορφος*, *άσχημος*

- Πατάκης – Larousse: *δύσμορφος*.

β) ή και διαφορά στα πρωτοτυπικά χαρακτηριστικά:

(i) *affreux* + N/ +An/[ + Hum]:

Hugo	46 600	α) Maillard est <i>affreux</i> β) L' <i>affreux</i> Dautun
Σκυλίτσης	51 536	α) Ό Μαιλάρδος υπήρξε <i>φρικώδης</i> β) τόν <i>άποτρόπαιον</i> Δοτέν
Αυγέρης	75 / 64 19/19	α) Ό Μαγιάρ είναι <i>φρι[κ]χτός</i> β) τό <i>φριχτό</i> Ντωτέν
Σκουλούδης	46 590	α) Ό Μαγιάρ είναι <i>φρικτός</i> β) τόν <i>άπαισιο</i> Ντωτέν



Λυκούδης	69 793	α) Ο Μπαγιάρ ήταν <i>φριχτός</i> β) τον <i>απαίσιο</i> Ντοτέν
----------	-----------	--

## Παράδειγμα 3.

Στο παράδειγμα αυτό παρατηρούμε και διαφορά στην κλίμακα διαβάθμισης: *απαίσιο, άποτρόπαιο, φριχτό, φρικώδης*  
(ii) *affreux + N[-An]/[+concr]*:

Hugo	1440 1479	α) <i>affreuse</i> vieille cave moisie β) <i>affreuse</i> maison
Σκυλίτσης	1231 1271	α) παλιό υπόγειο <i>έλεινόν</i> β) την <i>οίκτρ</i> άν κατοικίαν
Αυγέρης	344 / 516 403/556	α) <i>άσχημο</i> , παλιό υπόγειο β) τό <i>φρικτό</i> σπίτι
Σκουλούδης	1419 1454	α) <i>φριχτό</i> παλιό υπόγειο β) τό <i>παλιό</i> σπιτο
Λυκούδης	69 1954	α) <i>φριχτό</i> κελάρι μουχλιασμένο β) στο <i>απαίσιο</i> αυτό σπίτι

## Παράδειγμα 4.

Και στο παράδειγμα αυτό παρατηρούμε και διαφορά στην κλίμακα διαβάθμισης: *παλιό-, άσχημο, απαίσιο, φριχτό, οίκτρο*.

(iii) *affreux + N[-An]/[+abstr]*:

Hugo	368 1002	α) La guerre a d' <i>affreuses</i> beautés β) Mais l'argot est <i>affreux</i> !
Σκυλίτσης	338 851	α) Ο πόλεμος έχει <i>φρικτά</i> κάλλη β) Είνε <i>φρικώδες</i> τό ιδίωμα τούτο
Αυγέρης	88 / 82 199/40	α) Ο πόλεμος έχει <i>φρι/κ/χτές</i> όμορφιές β) Μά τό άργκό είναι <i>φριχτό</i> !
Σκουλούδης	354 992	α) Ο πόλεμος έχει όμορφιές <i>φρικαλέες</i> β) Τό άργκό είναι <i>φριχτό</i> !
Λυκούδης	491 1329	α) Ο πόλεμος έχει <i>απαίσιες</i> όμορφιές β) Μα η αργκό είναι <i>φριχτή</i>

## Παράδειγμα 5.

Εδώ, όπως και στο παράδειγμα 1, υπάρχει ομοφωνία μεταξύ των μεταφραστών ως προς την απόδοση του όρου *affreux*. Το επίθετο *φρικτός* που επικρατεί έναντι των υπολοίπων όρων που αποδίδουν στα ελληνικά το συγκεκριμένο επίθετο εμφανίζεται πρώτο στα λήμματα μόνο των πιο πρόσφατων λεξικών :

- Ηπίτης: *δεινός, φοβερός, φρικτός, φρικαλέος, φρικώδης, έκπληκτικός [...]* // *Είδεχθής, άπεχθής, δυσειδής, μυσαρός, άποτρόπαιος*. // *Υπερβολικός*.

- Βαρβάτης : *δεινός, φοβερός, φρικτός*. // *άπεχθής, δυσειδής, άποτρόπαιος*.

- Kauffmann : *φρικτός, απαίσιος*

- Πατάκης – Larousse : *φρικτός, αποκρουστικός, απαίσιος*.

3) Τα περισσότερα παραδείγματά μας όμως αποτελούν παραδείγματα λογικοπροτασιακών συνωνύμων καθώς «μπορούν να αλληλοϋποκατασταθούν σε οποιαδήποτε έκφραση χωρίς να επηρεαστούν οι συνθήκες αληθείας της έκφρασης αυτής» (Ξυδόπουλος, 2016:129) :

Hugo	1420 1474	α) un fourbe <i>abominable</i> ! β) <i>abominable</i> menteur!
Σκυλίτσης	1210 1266	α) ένας <i>κατάπτυστος</i> ύποκριτής β) <i>ψεύτη έπονείδιστε</i> !
Αυγέρης	314 / 496 396/552	α) ένας <i>συχαιμένος</i> άπατεώνας! β) <i>άχρειε</i> ψεύτη !
Σκουλούδης	1400 1451	α) ένας <i>άπαίσιος</i> καί <i>δόλιος</i> ύποκριτής ! β) <i>άπαίσιε</i> ψεύτη !
Λυκούδης	1881 1947	α) ένας <i>σιχαμερός</i> αγύρτης ! β) <i>παλιο</i> ψεύταρε !

## Παράδειγμα 6.

Οι παρατηρήσεις που εγείρει το παράδειγμα 6 είναι οι εξής : α) η διαφορά στην κλίμακα διαβάθμισης μεταξύ των επιθέτων *παλιοψεύταρε, άπαίσιε, άχρειε, έπονείδιστε, συχαιμένος / σιχαμερός, άπαίσιος, κατάπτυστος*. β) Η θέση του επιθέτου επηρεάζει τη σημασία του ουσιαστικού και κατά συνέπεια τη μετάφρασή του. Ο ρόλος της φύσης του υποκειμένου είναι επίσης καθοριστικός στη διαδικασία προσδιορισμού της σημασίας. Αυτή η προσέγγιση του επιθέτου σε βαθιά δομή εγείρει παρατηρήσεις σχετικές με τη σύνταξή του. Σύμφωνα με τους γραμματικούς κανόνες που διακρίνουν δυο λογικές λειτουργίες του επιθέτου, το ταξινομικό επίθετο βρίσκεται πριν το ουσιαστικό και φανερώνει μια αντικειμενική και αναγνωρίσιμη ιδιότητα. Το μη ταξινομικό επίθετο, από την άλλη, βρίσκεται δεξιά του ουσιαστικού και το χαρακτηρίζει πιο υποκειμενικά, φανερώνει την στάση του (συν)ομιλητή είτε σε συναισθηματικό επίπεδο είτε σε επίπεδο αξιολογικής κρίσης. Σύμφωνα με αυτή την διαπίστωση που αφορά στους επιθετικούς προσδιορισμούς, η θέση του επιθέτου καθορίζει την αντικειμενικότητα ή όχι της κρίσης και η θέση του δεν είναι σταθερή καθώς παρεμβάλλουν και άλλοι παράγοντες τυπικοί, σημασίας ή/και ύφους. γ) παρατηρούμε την ομοιότητα στην απόδοση των όρων σε σχέση με τις μεταφράσεις τόσο των προηγούμενων παραδειγμάτων, όσο και των λεξικών. Εκτός από το *απαίσιος*, καμία άλλη



από τις μεταφράσεις των λεξικών δεν χρησιμοποιήθηκε από τους μεταφραστές κατά την απόδοση των επιθέτων του συγκεκριμένου παραδείγματος:

- Ηπίτης: *Μυσαρός, βδελυρός, στυγερός, άνόσιος, άπεχθής, άποτρόπαιος* // *Κατ' έκτ. Κάκιστος* // (*οίκ.*), *έλεεινός, άθλιος*
- Βαρβάτης: *άποτρόπαιος, βδελυρός, μισαρός* // *έλεεινός, άθλιος*
- Kauffmann: *φρικτός, άποτρόπαιος, απαίσιος*
- Πατάκης – Larousse: *απαίσιος, άποτρόπαιος*

4) η σύγκριση και η αντιπαράβολή των όρων, τόσο μεταξύ τους, μέσω της κατανομικής ανάλυσης, όσο και με τους όρους που προτείνουν τα αντίστοιχα μονόγλωσσα λεξικά, ενισχύεται από τη σύγκριση και την αντιπαράβολή όλων των παραπάνω με τα αντίστοιχα δίγλωσσα λεξικά. Στο επίπεδο αυτό παρατηρούνται τα εξής φαινόμενα:

(i) σε κάποιες περιπτώσεις παρατηρείται ο πρώτος όρος που δίνεται στη γλώσσα -στόχο να είναι κοινός στα διαφορετικά λεξικά και επικρατέστερος στα παραδείγματα των λογοτεχνικών μας μεταφράσεων:

Hugo	109	Le monde moral n'a pas de plus grand spectacle que celui-là : une conscience troublée et inquiète, parvenue au bord d'une <i>mauvaise</i> action
Σκυλίτσης	114	εις τήν άκμήν του νά πράξη τό <i>κακόν</i>
Αυγέρης	173/137	έτοιμάζεται νά [κάμη]κάνει τήν <i>κακή</i> πράξη
Σκουλούδης	106	έτοιμη γιά μιιά <i>κακή</i> πράξη
Λυκούδης	148	που φτάνει στο <i>χειλός</i> μιας <i>κακής</i> πράξης

#### Παράδειγμα 7.

- Ηπίτης: *κακός, φαῦλος, μοχθηρός, σαθρός*
- Βαρβάτης: *κακός, άθλιος, έλαττωματικός* // *Αδέξιος, άπειρος* // *Βλαβερός, επίζημιος*. // *Απαίσιος, όλέθριος*. /// *Κακότροπος, κακεντρεχής, πονηρός, φαῦλος*
- Kauffmann: *κακός, άθλιος, άσχημος, ασθενικός*
- Πατάκης – Larousse: *κακός, άσχημος, αηδιαστικός* // *λανθασμένος, λάθος* // *βλαβερός*.

Το συγκεκριμένο λήμμα του Πατάκης – Larousse παρουσιάζει μια πολύ ενδιαφέρουσα και πρωτοφανή για τη δίγλωσση γαλλοελληνική λεξικογραφία πρωτοτυπία, καθώς διαχωρίζει στο εσωτερικό του λήμματος τις σημασιολογικές υποδιαίρεσεις σε διακριτές κατηγορίες, με υπολήμματα αντίστοιχα του *Trésor de la Langue Française*. Σε κάθε υποδιαίρεση το λήμμα εντάσσεται στο αντίστοιχο περιεχόμενο του καθιστώντας τις διαφορετικές μεταφράσεις πιο κατανοητές. Εκτός από το προσληπτικό επίπεδο, το λήμμα γίνεται πιο εύχρηστο και σε επίπεδο παραγωγής λόγου.

(ii) ο πρώτος όρος που δίνεται στη γλώσσα στόχο να χρησιμοποιείται από κάποιους από τους μεταφραστές μας αλλά όχι από όλους

Hugo	76	il se pourrait qu'il y eût de <i>méchantes</i> rencontres
Σκουλούδης	74	<i>κακό</i> συναπάντημα
Λυκούδης	107	<i>κακό</i> συναπάντημα

#### Παράδειγμα 8.

Η προτίμηση άλλου επιθέτου αντί του επικρατέστερου στα λεξικά δεν περιορίζεται, βέβαια, στο συγκεκριμένο παράδειγμα. Είναι μάλλον συνηθισμένη πρακτική, ιδίως κατά τη μετάφραση των επιθέτων ευρείας χρήσης, ο μεταφραστής να μην χρειάζεται να προστρέξει στη χρήση του λεξικού, παρά μόνο προς αναζήτηση συνώνυμου όρου.

- Ηπίτης: *Κακός, πάγκακος, πικρολόγος, ό έχων έλλειψιν καλοσύνης έπιρρεπής εις τό νά πράττη τό κακόν* // *διαβολικός, σατανικός*.
- Βαρβάτης: *Κακός*. // *Κακεντρεχής, μοχθηρός*. // *Άπεχθής, επικίνδυνος*. // *άθλιος, ούτιδανός, έλεεινός*. *Un homme méchant, βλάσφημος, κακολόγος*. // *Un méchant homme, πονηρός*. [...] // *Θορυβοποιός, un mauvais enfant, παιδίον θορυβοποιόν*. // *Σκυθρωπός, κατηφής*

- Kauffmann: *κακός, μοχθηρός, άσχημος, ελεεινός*. *Un homme méchant. Un méchant homme*. Ένας *κακός άνθρωπος*

Πατάκης – Larousse: *κακός, άγριος, που δαγκώνει, μοχθηρός* // *απαίσιος, άσχημος*.

Βλέπουμε δυο παραδείγματα χρήσης του επιθέτου από το λεξικό του Βαρβάτη και τα ίδια παραδείγματα από το λεξικό του Kauffmann. Η απόκλιση στην απόδοση είναι σύνηθες φαινόμενο. Όμως εδώ παρατηρούμε, όπως και στο παράδειγμα 6, πώς η θέση του επιθέτου επηρεάζει τη σημασία της φράσης στο λεξικό του Βαρβάτη και κατά συνέπεια τη μετάφρασή της.

(iii) Αντί του πρώτου, προτιμώνται κάποιοι από τους υπόλοιπους όρους που περιέχονται στο λήμμα:

Hugo	1478	le vilain père !
Σκυλίτσης	1270	τόν <i>κακόν</i> αὐτόν πατέρα!
Αυγέρης	400/555	ό <i>κακός</i> πατέρας
Σκουλούδης	1453	τόν <i>κακόν</i> μου τόν πατέρα !
Λυκούδης	1952	Τι <i>κακός</i> !

#### Παράδειγμα 9.

- Ηπίτης: *Δύσμορφος, άσχημος*. // *Ένοχλητικός, άηδής, άσχημος* // *Φιλάργυρος*. // *Αίσχρος, μυσαρός, φαῦλος, άχρεϊος, άνθρωπος οὔτινος ή διαγωγή ή οι λόγοι έχουσι τό άναίσχυντον, τό φαῦλον*.

- Βαρβάτης: *Δυσειδής, άσχημος*. // *άηδής, άπειρόκαλος*. // *όκληρός, δυσάρεστος*. // *Αίσχρος, ρυπαρός, κακοήθης* (έπί προσώπων, λόγων καί πράξεων) // *επικίνδυνος, κακοήθης*

- Kauffmann: *άσχημος, κακός, πονηρός*

- Πατάκης – Larousse: *αποκρουστικός, αντιαισθητικός, άσχημος, λέρος, άθλιος, κακός*

Στο συγκεκριμένο παράδειγμα η εκφραστικά μαρκαρισμένη χρήση του επιθέτου δεν γίνεται αντιληπτή παρά μόνο από το περιεχόμενο των λογοτεχνικών μας μεταφράσεων. Παραθέτουμε ενδεικτικά τη μετάφραση του Λυκούδη από το συγκεκριμένο απόσπασμα: «Από πότε έχετε έρθει; Γιατί δεν μας ειδοποιήσατε; Ξέρετε πως έχετε πολύ αλλάξει; *Τι κακός!*



Αρρώστησε και μεις δεν το μάθαμε!». Στο Πατάκης – Larousse η αναφορά δεν επισημαίνεται ως συναισθηματική / συγκινησιακή συνυποδήλωση (Μότσιου, 1994: 187) : « 2. [méchant] κακός, *tu es un vilain garçon ! είσαι κακό παιδί!*» και δεν γίνεται αντιληπτό ότι η χρήση του επιθέτου στο συγκεκριμένο περιεχόμενο απαλύνει την αρχική αρνητική του δήλωση (Γούτσος 2015 : 89).

(iv) Γνωρίζοντας ότι το λήμμα ενός δίγλωσσου λεξικού παρέχει επίσης ισοδύναμα του λήμματος στη γλώσσα-στόχο αλλά και αριθμό αποδόσεων που μπορούν να γίνουν εσφαλμένα αντιληπτές ως συνώνυμα, ο μεταφραστής λαμβάνει υπόψη τις πληροφορίες, τόσο για το συγκεκριμένο όσο και τις σχετικές με τα συμφραζόμενα που περιέχονται στο λήμμα (Βλαχόπουλος 2015: 94). Οι πληροφορίες αυτές ορίζουν και τη ή τις σημασίες κάποιας λέξης. Αυτό επιβεβαιώνεται και από τα παραδείγματά μας :

Hugo	105 368	α) méditation hideuse β) contresens hideux!
Σκυλίτσης	110 354	α) Έν τῷ μέσω τῆς κακεντρεχοῦς ταύτης μελέτης β) εἰδεχθῆς ὀξύμωρον!
Αυγέρης	167 / 133 110/101	α) μέσα στήν ἀπαίσια σκέψη του β) Ἀπεχθής ἀντινομία
Σκουλούδης	101 101	α) Μέσα σ' αὐτήν τήν ἀπαίσια περισυλλογή β) Ἀντίθεση φρικαλέα
Λυκούδης	144 508	α) στον απαίσιο αυτό διαλογισμό β) Αποκρουστική αντίφαση

*Παράδειγμα 10.*

- Ηπίτης: Εἰδεχθέστατος, εἰδεχθής, ἀσχημότητος. // Ἀποτρόπαιος, βδελυρός
- Βαρβάτης Δυσειδής, εἰδεχθής, ἀπεχθής, ἀποτρόπαιος.
- Kauffmann : αποκρουστικός, βδελυρός, αποτρόπαιος
- Πατάκης – Larousse : οἰκτρός, βδελυρός, αποκρουστικός, ἀπεχθής, εἰδεχθής, ἀποτρόπαιος.

#### 4.1 Συσσώρευση επιθέτων

Η περιγραφή, με την υποκειμενικότητα που της αναλογεί, αποτελεί μια επιπλέον δυσκολία για τον μεταφραστή κατά την απόδοσή της, καθώς η επιλογή του αντίστοιχου – ισοδύναμου επιθέτου απαιτεί ιδιαίτερη προσοχή και δεξιότητα. Το μυθιστόρημα βρίθκει περιγραφών και ο μεταφραστής καλείται, μεταξύ άλλων δυσκολιών, να αποδώσει και την πληθώρα των επιθέτων που χρησιμοποιεί ο συγγραφέας με την περίφημη δεξιότητά του, χωρίς να υπερφορτώνει το κείμενο.

Hugo	73	Tout cet ensemble était hideux, petit, lugubre et borné
Σκυλίτσης	78	Τά πάντα δέ πένθιμα, θλιβερά καί ἀποτρόπαια
Αυγέρης	117 94	Όλο αυτό τό σύνολο ἦταν ἄσχημο, μικρό πένθιμο καί στενόχωρο Τό σύνολο ἔδειχνε ἄσχημο, μικρό πένθιμο καί στενόχωρο
Σκουλούδης	70	Όλα αὐτά μαζί ἦταν ἀπαίσια, λειψά, πένθιμα, πληχτικά
Λυκούδης	103	Όλο τοῦτο το σύνολο ἦταν ἀπαίσιο, λειψό κακορίζικο καί πένθιμο

*Παράδειγμα 11.*

Hugo	758	Sans doute ils paraissaient bien dépravés, bien corrompus, bien avilis, bien odieux même
Σκυλίτσης	674	τά πλάσματα ταῦτα ἐφαίνοντο διεφθαρμένα, ἐπονεϊδιστα, μυσάρα μάλιστα τά πιό ἄχρεϊα ὄντα
Αυγέρης	254/251	Βέβαια φαίνονταν πολύ χαλασμένοι, [πολύ] πεσμένοι, πολύ προστυχωμένοι, μισητοί μάλιστα
Σκουλούδης	756	Βέβαια φαίνονταν πολύ ἐξαχρειωμένοι, πολύ χαλασμένοι, πολύ ξεπεσμένοι, μπορεῖ καί ἀπαίσιοι
Λυκούδης	1018	Φαίνονταν βέβαια ὄντα διαλυμένα, διεφθαρμένα, εκφυλισμένα, ἀξιομίσητα ἴσως

*Παράδειγμα 12.*

Hugo	759	le taudis [...] était abject, sale, fétide, infect, ténébreux, sordide
Σκυλίτσης	675	τό κατάλυμα [...] ὑπῆρχε βδελυρόν, ρυπαρόν, δυσῶδες, σκοτεινόν, γλίσχρον
Αυγέρης	256/253	Ἡ βρωμοκέλα [...] ἦταν [ἀθλία] ἄθλια, [ἀκάθαρτη], βρωμοῦσε, ἦταν σκοτεινή, κ' ἔδειχνε φοβερή φτώχεια
Σκουλούδης	757	Τό ἀχούρι [...] ἦταν ἐλεεινό, βρώμικο, μολυσμένο, ἄθλιο, νοσηρό
Λυκούδης	1020	Ἡ φωλιά [...] ἦταν ἀπαίσια, βρωμερή, κάκοσμη, μολυσμένη, σκοτεινή καί τρισάθλια

*Παράδειγμα 13.*

Hugo	593	cette cale étroite fétide, obscure, sordide, malsaine, hideuse, abominable
Σκυλίτσης	528	τήν στενήν, τήν δυσώδη, σκοτεινήν, ἐκκωφαντικήν, μiasματώδη, εἰδεχθῇ, καί ἀποτρόπαιον ἐκείνην τροπίδα
Αυγέρης	8/10	τό ἀμπάρι αὐτό τό στενό, τό βρώμικο, τό σκοτεινό, τό ἄσ[χ]κημο
Σκουλούδης	582	γιά νά γίνει αὐτό τό ἀμπάρι τό στενό, τό μολυσμένο, τό σκοτεινό, τό νοσηρό, το ἀνθυγιεινό, τό ἀπαίσιο, τό φριχτό
Λυκούδης	784	το στενόχωρο, ρυπαρό, σκοτεινό, νοσηρό, κάκοσμο, φριχτό καί ἀπαίσιο αὐτό ἀμπάρι

*Παράδειγμα 14.*

Hugo	1005	la langue laide, inquiète, sournoise, traître, venimeuse, cruelle, louche, vile, profonde, fatale, de la misère
Σκυλίτσης	853	ή δύσμορφος, ή ἀνήσυχος, ή συνθηματική, ή προδότις, ή δηλητηριώδης, ή τραχεῖα καί πλήρης ὀμότητος, ή ἀπαίσια, ή χαμερπής γλώσσα τῆς πενίας
Αυγέρης	199/43	γλώσσα ἄσχημη, ἀνήσυχη, [σκυθρωπή], προδότρα, φαρμακερή, σκληρή, διφορούμενη, πρόστυχη



		βαθ[ε]ιά, άπαισία, είναι ή γλώσσα της δυστυχίας
Σκουλούδης	994	γλώσσα άσχημη, άνήσυχη, ύποκριτική, προδοτική, φαρμακερή, σκληρή, ύποπτη, ποταπή, βαθειά και μοιραία γλώσσα της αθλιότητας
Λυκούδης	1332	η άσχημη, κρυψίβουλη, ύπουλη, φαρμακερή, προδοτική, ωμή, χαμόσυρτη, βαθιά και μοιραία γλώσσα της αθλιότητας

## 4.2 Επίθετα με μεγαλύτερη συχνότητα χρήσης

Η αναζήτηση του όρου αρχικά μέσα από τις διαφορετικές μεταφράσεις, και έπειτα στα λεξικά, τόσο στα δίγλωσσα όσο και στα μονόγλωσσα της γαλλικής και της ελληνικής γλώσσας, εξασφαλίζει στον μεταφραστή τη δυνατότητα επιλογής ανάμεσα σε μια γκάμα συνώνυμων επιθέτων. Οι ομοιότητες στους ορισμούς, στα μεταφράσματα, αλλά και μεταξύ τους, δεν επιβεβαιώνουν απλά την ύπαρξη της συνωνυμίας, αλλά αποκαλύπτουν και μια άλλη διάστασή της, η οποία αναδύεται μέσα από την διαγλωσσική προσέγγισή της. Η καταγραφή των παραδειγμάτων με τα περικείμενά τους μας επιτρέπει να έχουμε μια σφαιρική άποψη των διαφορετικών κατανομών των συνώνυμων επιθέτων και των αντίστοιχων μεταφράσεών τους. Τόσο από τις λογοτεχνικές μας μεταφράσεις όσο και από αυτές των λεξικών παρατηρούμε ότι συνώνυμοι δεν είναι μόνο οι όροι που παρατάσσονται ανά σημασιακό πεδίο, αλλά και οι περισσότεροι από τους όρους των διαφορετικών σημασιακών πεδίων μεταξύ τους. Από τα είκοσι επτά επίθετα του αρχισημειώματός μας, τα δέκα που σημειώνουν τη μεγαλύτερη συχνότητα χρήσης στο πρωτότυπο είναι, κατά φθίνουσα σειρά εμφάνισης, τα εξής:

*mauvais, terrible, misérable, affreux, hideux, horrible, sinistre, laid, abominable* και *difforme*. Συγκρίνοντας τους όρους με τους οποίους τα επίθετα αυτά αποδίδονται στα δίγλωσσα γαλλοελληνικά λεξικά που εξετάζουμε, παρατηρούμε:

α) ότι βρίσκονται σε σχέση ελεύθερης εναλλαγής μεταξύ τους (Μότσιου, 1994: 186):

β) ότι οι λογοτεχνικές μεταφράσεις (Λ.Μ.) έρχονται να συμπληρώσουν τη λίστα αυτών των μεταφράσεων με νέες προτάσεις:

1) *mauvais*

- Ηπίτης : Κακός, φαύλος, μοχθηρός, σαθρός

- Βαρβάτης : Κακός, άθλιος, ελαττωματικός Βλαβερός, επίζημιος. Απαίσιος, όλέθριος. Κακότροπος, κακεντρεχής, πονηρός, φαύλος

- Kauffmann : κακός, άθλιος, άσχημος, ασθενικός

- Πατάκης – Larousse : κακός, άσχημος, [...] βλαβερός

- Λ.Μ.: αλ[ι]ητήριος, άχρεϊος, , καταχθόνιος, κολασμένος, σατανικός, τιποτένιος, ύποπτος [*mauvaise mine*: ύποπτη φάτσα] χαμένος

2) *terrible*

- Ηπίτης : Τρομερός, φοβερός, δεινός, βλοσυρός, ίκανός νά ένσπείρη τόν τρόμον. Σφοδρός, ίσχυρότατος Παράδοξος, έκπληκτικός, άλλόκοτος, παράξενος άσχημος

- Βαρβάτης : Τρομερός, φοβερός// (μεταφ.) Δεινός, σφοδρός, καταπληκτικός,

// Άθλιος, άνιαρός. // Παράδοξος, άλλόκοτος, ιδιότροπος. // άπαίσιος

- Kauffmann : φοβερός, συνταρακτικός, καταπληκτικός,, ανυπόφορος

- Πατάκης – Larousse : τρομερός, φοβερός, δεινός, ανυπόφορος, φρικτός, // καταπληκτικός,

Θαυμάσιος

Λ.Μ. άβάστατος, άγριος, άθλιος, άποτρόπαιος, ειδεχθής, επίφοβος, εφιαλτικός, κακός, οργίλος, παγερός τρομα[κ]χτικός, φρικαλέος, φρικώδης, ώμός

3) *misérable*

- Ηπίτης : άθλιος, δύσμοιρος, ταλαίπωρος, κακοδαίμων, δυστυχής, κακορροϊζικός, οϊκτρός, έλεεινός, πανάθλιος, τρισάθλιος, [...] μίζερος, πτωχός, // περιφρονητέος // άχρείος, φαύλος, μηδαμινός

- Βαρβάτης : δύστηνος, ταλαίπωρος, δυστυχής, αζιολύπητος, άθλιος/ μοχθηρός, πονηρός, απαίσιος / ευτελής, μηδαμινός

- Kauffmann : Άθλιος, εξαθλιωμένος, ασήμαντος, πανάθλιος

- Πατάκης – Larousse : άθλιος, εξαθλιωμένος, μηδαμινός, ασήμαντος, παλιο--, άχρείος

Λ.Μ. -

4) *affreux*

- Ηπίτης : Δεινός, φοβερός, φρικτός, φρικαλέος, φρικώδης, // Είδεχθής, άπεχθής, δυσειδής, μυσαρός, άποτρόπαιος. // Υπερβολικός

- Βαρβάτης : Δεινός, φοβερός, φρικτός. // άπεχθής, δυσειδής, άποτρόπαιος

- Kauffmann : φρικτός, απαίσιος

- Πατάκης – Larousse : φρικτός, αποκρουστικός, απαίσιος, τρομερός

Λ.Μ. άγωνιώδεστατος, άθλιος, άηδής, ανατριχιαστικός, άσχημος, άχρεϊος, βαρύς, βρώμικος, δεινός, , έλεεινός, έφιαλτικός, θανάσιμος, μαύρος, οϊκτρός, τρομαχτικός, φρικιαστικός, , χείριστος

5) *hideux*

- Ηπίτης : Είδεχθέστατος, ειδεχθής, άσημότατος. // Άποτρόπαιος, βδελυρός

- Βαρβάτης : Δυσειδής, ειδεχθής, άπεχθής, άποτρόπαιος, άσημότατος

- Kauffmann : αποκρουστικός, βδελυρός, άποτρόπαιος

- Πατάκης – Larousse : οϊκτρός, βδελυρός, αποκρουστικός, άπεχθής, ειδεχθής, άποτρόπαιος

Λ.Μ. [ά]άγριος, [ά]αηδής, [ά]άθλιος, [ά]ακάθατος, αντιπαθητικός, [ά]απαίσιος, [ά]άσχημος, άχρεϊος, βρωμερός, γελοιός, δυσειδής, έλεεινός, κακεντρεχής, κακόμορφος, κακός, καταχθόνιος, μιάρός, μοχθηρός, μυσαρός, πένθιμος, σιχαμερός, σιχαμένος, στυγερός, τερατόμορφος, τερατώδης, τρομερός, φοβερός, φονικός, φρικαλέος, φρι[κ]χτός

6) *horrible*



- Ηπίτης : Φρικτός καί φρικώδης, ό προξενών φρίκην, φρικαλέος, φοβερός, τρομακτικός, στυγερός. [...] // Κάκιστος, έλεεινός
- Βαρβάτης : Φρικώδης, άποτρόπαιος, βδελυρός, άθλιος ύπέρμετρος
- Kauffmann : φρικτός, απαίσιος, τρομερός, φοβερός
- Πατάκης – Larousse : φρικτός, φρικιαστικός, τρομακτικός, απαίσιος, φρικαλέος, αποτρόπαιος / αποκρουστικός, πανάσχημος, κακόσχημος, αηδιαστικός
- Α.Μ. απεγνωσμένος, βλοσυρός, δεινός, είδεχθής, ζοφώδης, μυσαρός, οδυνηρός, τερατώδης
- 7) *sinistre*
- Ηπίτης : Απαίσιος, σκαιός, άποτρόπαιος. // Δυσοίωνος, όλέθριος, // Δυστυχεστάτος
- Βαρβάτης : Απαίσιος, δυσοίωνος. // Φοβερός, στυγερός, άποτρόπαιος. // Κατηφής, σκαιός, σκυθρωπός // Μοχθηρός, κακός, όλέθριος, άπενκταϊός
- Kauffmann : απαίσιος, άσχημος, θλιβερός
- Πατάκης – Larousse : απειλητικός, τρομακτικός, απαίσιος, διαβολικός, κακός / θλιβερός, απαίσιος
- Α.Μ. άνατριχιαστικός, , [ά]πόκοσμος, αποκρουστικός, βλαβερός, επίβουλος, επικίνδυνος, ζοφερός, κακορίζικος, , κατάμαυρος, μακάβριος, πένθιμος, πονηρός, πυκνός, σκοτεινός, στυ[ι]γνός, τραγικός, τρομερός, φρικαλέος, φρικιαστικός, φρι[κ]χτός
- 8) *abominable*
- Ηπίτης : Μυσαρός, βδελυρός, στυγερός, άνόσιος, άπεχθής, άποτρόπαιος // Κάκιστος. έλεεινός, άθλιος. άχρειέστατος
- Βαρβάτης : άποτρόπαιος, βδελυρός, μισαρός. // έλεεινός, άθλιος
- Kauffmann : φρικτός, αποτρόπαιος, απαίσιος
- Πατάκης – Larousse : απαίσιος, αποτρόπαιος
- Α.Μ. αίσχος, ακατονόμαστος, άνόσιος, άσχημος, άφόρητος, άχρειός, βρώμικος, δόλιος, έπονείδιστος, κατάπτυστος, καταραμένος, οίκτηρος, συχαμένος, σιχαμερός, φοβερός, φρικαλέος
- 9) *laid*
- Ηπίτης : Είδεχθής, δυσειδής, άσχημος, κοινώς, κακομούτσουνος // Άπρεπής
- Βαρβάτης : Δυσειδής, κακόμορφος, άσχημος, // (μεταφ.) Άπρεπής
- Kauffmann : άσχημος
- Πατάκης – Larousse : άσχημος, κακάσχημος, κακός
- Α.Μ. άκαλαίσθητος, άχαρις, δύσμορφος
- 10) *difforme*
- Ηπίτης : Δύσμορφος, δυσειδής, είδεχθής, στρεβλός
- Βαρβάτης : Δύσμορφος, δυσειδής, άσχημος
- Kauffmann : δύσμορφος, άσχημος
- Πατάκης – Larousse : δύσμορφος
- Α.Μ. [ά]άγαρμος, άκατασκενάστος, [ά]άμορφος, [ά]απαίσιος

## 5 Συμπεράσματα

Τόσο από τα παραδείγματα που παραθέσαμε παραπάνω όσο και από την τελευταία αυτή λίστα των επιθέτων, παρατηρούμε: 1) τα περισσότερα από τα επίθετα που αποτελούν το αρχισιμήμά μας μπορούν να εναλλάσσονται στο ίδιο περικείμενο, χωρίς να διαταράσσονται οι συνθήκες αληθείας, ούτε ο βαθμός φυσικότητας. 2) Οι διαφορές που συναντούμε στα λογικοπροτασιακά συνώνυμα είναι κυρίως στο επίπεδο ύφους, ενώ στα πλησιώνυμα είναι τόσο στην κλίμακα διαβάθμισης όσο και στο ποιο ενεργείας και στα πρωτοτυπικά χαρακτηριστικά. 3) Τα παλαιότερα λεξικά περιλαμβάνουν περισσότερα παραδείγματα και προτείνουν περισσότερες μεταφράσεις. Στα σύγχρονα λεξικά παρατηρούμε άνιση έκταση στην ανάλυση και παράθεση παραδειγμάτων, ειδικά αν αντιπαραβάλουμε το λήμμα *mauvais* που καταλαμβάνει μια ολόκληρη σελίδα, με το λήμμα *difforme* που αποτελείται μόνο από το επίθετο *δύσμορφος*. 4) Το επίθετο *misérable* - ίσως και λόγω της ιδιαιτερότητάς του, καθώς αποτελεί τον τίτλο του μυθιστορήματος- είναι το μόνο από τα δέκα επίθετα του οποίου οι μεταφράσεις καλύπτονται από τις προτάσεις των λεξικών. Ωστόσο, το ελληνικό του αντίστοιχο *άθλιος* συναντάται ως μετάφραση και πολλών άλλων επιθέτων. 5) Στα υπόλοιπα επίθετα παρατηρούμε ότι οι λογοτεχνικές μεταφράσεις συμπληρώνουν την εικόνα του επιθέτου διευρύνοντας το σημασιακό του πεδίο μέσω των διαφορετικών περικειμένων. Στις λογοτεχνικές μεταφράσεις συναντούμε και τις μεταφράσεις των λεξικών και τις μεταφράσεις που έχουμε συγκεντρώσει στη λίστα παραπάνω. Είναι αξιοσημείωτο ότι υπάρχουν παραδείγματα που συγκεντρώνουν έναν μεγάλο αριθμό επιθέτων που εντοπίζονται μόνο στις λογοτεχνικές μεταφράσεις, όπως τα : *hideux* (30), *sinistre* (20), *affreux* (17), *abominable* (16) και *terrible* (14). 6) Τέλος, σε μια αντίστροφη προσέγγιση, τα επίθετα που εμφανίζονται πιο συχνά κατά την απόδοση των συγκεκριμένων επιθέτων στα ελληνικά είναι τα: *κακός, απαίσιος, φρικτός, τρομερός, άσχημος* και *άθλιος*. Η μελέτη τόσο της συνωνυμίας τους όσο και της απόδοσής τους στη γαλλική γλώσσα αναδεικνύει μέσα από μια ανατόφευκτη κυκλικότητα τη σχέση που συνδέει τα συγκεκριμένα επίθετα μεταξύ τους. Ο Martin προτείνει ως μέσο αποφυγής της κυκλικότητας του λεξικογραφικού έργου την αναζήτηση των πιθανών μεταφράσεών του. Το νόημα ενός εκφωνήματος μπορεί να οριστεί από το σύνολο των πιθανών μεταφράσεών του τόσο ενδογλωσσικών όσο και διαγλωσσικών. Συνεχίζοντας, αναφέρεται στα οφέλη της περικειμενικής ανάλυσης, στην αναζήτηση δηλαδή της σημασίας του όρου ενός εκφωνήματος μέσα από ένα σύνολο γλωσσικών περικειμένων στα οποία εμφανίζεται (Martin, 1976). Η αναφορά των διαφοροποιητικών σημασιολογικών χαρακτηριστικών ως αποτέλεσμα αυτής της μεθόδου μπορεί να συμβάλλει στην αποσαφήνιση της αμφισημίας και της πολυσημίας των όρων καθώς και στην ιεράρχηση της δομής του λήμματος.



Με βάση τα αποτελέσματα που προέκυψαν από τη σύγκριση των όρων με αυτούς που χρησιμοποιούνται στις μεταφράσεις τόσο τις λογοτεχνικές όσο και των διγλωσσών λεξικών, αλλά και τους ορισμούς που παρατίθενται στα μονόγλωσσα λεξικά, παρατηρούμε ότι το ίδιο επίθετο μπορεί να χρησιμοποιηθεί σε πολλά και διαφορετικά περιεχόμενα καθώς και ότι, στις περισσότερες περιπτώσεις, πληροί το κριτήριο της αλληλοϋποκαταστασιμότητας. Αυτό οφείλεται στη φύση του και στη δυνατότητα αναφοράς του, ως κατηγορήμα, σε πολλά και διαφορετικά υποκείμενα. Η διαγλωσσική μελέτη της συνωνυμίας του επιθέτου μέσα από τις μεταφράσεις του και η αντιπαραβολή τους με τις μεταφράσεις των διγλωσσών λεξικών διαχρονικά αποδεικνύει πως στα παλαιότερα λεξικά γινόταν μεγαλύτερη αναφορά σε παραδείγματα. Η παράθεση παραδειγμάτων από έγκριτες μεταφράσεις, σε διαχρονική και συγχρονική κλίμακα, μπορεί να συμβάλλει στον εμπλουτισμό του λήμματος και κατ' επέκταση στην αποσαφήνιση περιπτώσεων συνωνυμίας αλλά και πολυσημίας, συνεισφέροντας σημαντικά στην αποτελεσματικότερη χρήση του λεξικού.

## 6 Βιβλιογραφία

### I.

- Hugo, V. (2000). *Les Misérables*. Édition établie et annotée par Maurice Allem. Paris : Bibliothèque de la Pléiade, nrf, éditions Gallimard, Premier dépôt légal : 1951 :
- Victor Hugo. *Οι Άθλιοι*. Κατά την μετάφραση 'Ι. Σκυλίτση, 2 τόμοι, (Πλήρεις καί ἄνευ οὐδεμιᾶς περικοπῆς τοῦ κειμένου, μετὰ πολλῶν καί καλλιτεχνικῶν εἰκόνων). Ἀθῆναι: Ἐκδότης Οικονόμου, (χ.χ.).
- Βίκτωρ Ουγγώ. *Οἱ Ἀθλοῖ*. Μετάφρασις Μ. Αὐγέρη. Βιβλιοθήκη «Ἐκλεκτά Ἔργα», Ἀρ. 51-55, 5 τόμοι. Ἀθῆναι: Βιβλιοπωλεῖον Γεωργίου Ι. Βασιλείου, 1922 (1ος, 2ος, 3ος τ.), 1925 (4ος, 5ος τ.).
- Βίκτωρ Ουγγώ. *Οἱ Ἀθλοῖ*. Μετάφραση ἀναθεωρημένη – συμπληρωμένη καί προλογική κριτική Μ. Αὐγέρη, 2 τόμοι. Ἐπίτιμος γενική ἐποπτεία Οκταβίου Μερλιέ, Διευθυντοῦ τοῦ Γαλλικοῦ Ἰνστιτούτου. Ἀθῆναι: Ἐκδοτικὸς ἐπιμορφωτικὸς ὀργανισμὸς, 1958.
- Βίκτωρ Ουγγώ. *Οἱ Ἀθλοῖ*. Πλήρης μετάφρασις ἀπὸ τὸ πρωτότυπον τῆς Edition Nationale, ὑπὸ Μ. Σκουλούδη, 2 τόμοι. Ἀθῆναι: Ἐκδόσεις «Περγαμηνά», Δημοτράκος, 1958.
- Βίκτωρ Ουγγώ. *Οἱ Ἀθλοῖ*. Μετάφραση Μ. Λυκούδης, 5 τόμοι. Ἀθήνα: Νεανική Βιβλιοθήκη, Κλασική λογοτεχνία, Ἐκδόσεις Καστανιώτη, (1988), 2000.

### II.

- Αναστασιάδη-Συμεωνίδη, Α. (1978). Ψυχοσυστηματική της μετοχής στην κοινή νεοελληνική, *Φιλολόγος* 13, Θεσσαλονίκη, σσ. 311-319.
- Βαλετόπουλος, Φ. (2014). Décrire l'état psychologique de peur. In : Γαβριηλίδου, Ζ., Ρεβυθιάδου, Α. *Μελέτες αφιερωμένες στην ομότιμη καθ. Α. Αναστασιάδη-Συμεωνίδη*. Καβάλα: Σαῖτα, σσ. 165-178.
- Βαρβάτης, Κ. *Λεξικὸν Γαλλοελληνικόν, Dictionnaire franco-grec*. Ἐκδοσις νέα ἀναθεωρηθεῖσα, διορθωθεῖσα καί συμπληρωθεῖσα ὑπὸ Ε.Γ. Καραθάνου. Ἀθήνα : Ἐκδ. οἶκος Ι. Σιδέρης, 1918.
- Βλαχόπουλος, Σ. (2015). *Η μετάφραση οικονομικών κειμένων ως διαπολιτισμική επικοινωνία γνώσεων, Διαπολιτισμική επικοινωνία στην οικονομία*, σσ. 75-106 [https://repository.kallipos.gr/bitstream/11419/212/1/chapter03\\_15149.pdf](https://repository.kallipos.gr/bitstream/11419/212/1/chapter03_15149.pdf).
- Βλάχος, Ά. *Ἑλληνογαλλικὸν λεξικόν*. Ἀθήνα : Ἐκδ. οἶκος Ι. Σιδέρης, 4<sup>η</sup> ἔκδοσις, 1963.
- Boone, A., Joly, A. (1996). *Dictionnaire terminologique de la systématique du langage*. Paris : L'Harmattan, coll. « sémantiques ».
- Γαλλοελληνικὸ Λεξικόν, *Dictionnaire français-grec moderne*. (2018) Ἀθήνα: Πατάκης - σε συνεργασία με τον εκδοτικό οἶκο Larousse, responsable éditorial Georgios Galanes.
- Γούτσος, Δ., Φραγκάκη, Γ. (2015) Εισαγωγή στη γλωσσολογία σωμάτων κειμένων. Ἀθήνα : ΣΕΑΒ, [https://repository.kallipos.gr/bitstream/11419/1932/1/00\\_master\\_document\\_Goutsos.pdf](https://repository.kallipos.gr/bitstream/11419/1932/1/00_master_document_Goutsos.pdf)
- Chevalier, J.-Cl., Launay, M., Molho, M., Sur la nature et la fonction de l'homonymie, de la synonymie et de la paronymie. In : Fuchs, C. (dir.), *L'ambiguïté et la paraphrase*, Caen, Centre de Publications de l'Université de Caen, p.45-52.
- Chomsky, N. (1975). *Questions de sémantique*. Trad. de l'anglais par B. Cerquiglini. Paris : Seuil, L'ordre philosophique.
- Crystal, D. (2003). *Λεξικό Γλωσσολογίας και Φωνητικής*. Μετάφραση Γ. Ξυδόπουλος. Ἀθήνα: Πατάκης.
- Douay ; C. ; Roulland, D : (1990). *Les mots de Gustave Guillaume : Vocabulaire technique de la psychomécanique du langage* : Rennes : PUR :
- Fuchs, C., (1988). L'ambiguïté et la paraphrase en linguistique. In : Fuchs, C. (dir.), *L'ambiguïté et la paraphrase*, Caen, Centre de Publications de l'Université de Caen, pp.15-20.
- Greimas ; A.J.(1995). *Sémantique structurale : Recherche de méthode*. Paris : PUF ; Coll. Formes Sémiotiques.
- Greimas ; A.J.(2005). *Δομική Σημασιολογία. Αναζήτηση μεθόδου*. Μετάφραση Παρίσης Γ., Επιμέλεια Κατωμένους Ε. Ἀθήνα : Ἐκδόσεις Πατάκης.
- Guillaume, G. (1994). *Langage et science du langage*. Paris-Québec : A.-G. Nizet, Presses de l'Université Laval.
- Guillaume, G. (2018). *Leçons de linguistique de Gustave Guillaume. Leçons de l'année 1940-1941 (vol. 26)*. Laval : PUL.
- Guimier, Cl. (1988), Incidence, ambiguïté et paraphrase. Approche psychomécanique. In : Fuchs, C. (dir.), *L'ambiguïté et la paraphrase*, Caen, Centre de Publications de l'Université de Caen, pp. 77-81.
- Guimier, Cl. (1991), Sur la fonction d'attribut du sujet : approche psychomécanique. In : Gaulmyn, M.-M. et Rémi-Giraud, S. (dirs), *À la recherche de l'attribut*. Lyon, Presses de l'Université de Lyon, p. 209-235.
- Ηπίτης, Α. *Μέγα Γαλλοελληνικὸν Λεξικόν*. Προσθήκαι εἰς τα στοιχεία Α-Β-Γ-Δ. Ἀθήναι: Ἐκδόσεις Ἀ/φῶν Τολιδή, (1977), 2000.
- Λεξικό της Κοινῆς Νεοελληνικῆς*, (2013). Θεσσαλονίκη: ΑΠΘ, Ἰνστιτούτο Νεοελληνικῶν Σπουδῶν, Ἰδρυμα Μανόλη Τριανταφυλλίδη.
- Martin, R. (1976). *Inférence, antonymie et paraphrase, éléments pour une théorie sémantique*. Paris : Librairie C.



- Klincksieck, Bibliothèque française et romane.
- Μότσιου, Β. (1994). *Στοιχεία Λεξικολογίας, Εισαγωγή στη νεοελληνική λεξικολογία*. Αθήνα: Νεφέλη / Γλωσσολογία, 9.
- Μπαμπινιώτης, Γ. (1998). *Λεξικό της νέας ελληνικής γλώσσας*. Αθήνα: Κέντρο Λεξικολογίας Ε.Π.Ε..
- Ξυδόπουλος, Γ. (2017). *Λεξικολογία, Εισαγωγή στην ανάλυση της λέξης και του λεξικού*, Αθήνα: Εκδόσεις Πατάκη.
- Παντελοδήμος, Δ., Lust, C. (1996). *Γαλλο-Ελληνικό Λεξικό, Dictionnaire français grec moderne*. Athènes : Librairie Kauffmann, Distribution Internationale Hatier.
- Picoche, J. (1992-α). *Précis de lexicologie française, L'étude et l'enseignement du vocabulaire*. Nathan Université, coll.fac. linguistique, Série « Linguistique ».
- Picoche, J. (1992-β). *Structures sémantiques du lexique français*. Paris : Nathan -Recherche, Linguistique Française.
- Pottier, B.(1976). *Sémantique et logique*. Paris : J. P. Delarge, coll. Univers sémiotiques, sous la direction de A.-J. Greimas, Editions Universitaires.
- Pottier, B.(1992). *Sémantique générale*. Paris : P.U.F., coll. Linguistique Nouvelle.
- Robert, P. (1985). *Le Grand Robert de la langue française, dictionnaire alphabétique et analogique de la langue française*, 2<sup>e</sup> édition revue et enrichie par Alain REY, 9 vol. Montréal-Paris : Dictionnaires Le Robert.
- Rontogianni, A. (2006). *Sémantique et traduction. Problèmes de polysémie et de synonymie à travers les traductions grecques des Misérables de V.Hugo*. Thèse de Doctorat, Université Paris-Sorbonne, Paris IV (υπό έκδοση).
- Slatka D. (1971) Esquisse d'une théorie lexico-sémantique: pour une analyse d'un texte politique (Cahiers de doléances). In: *Langages*, 6<sup>e</sup> année, n°23, Le discours politique, sous la direction de Louis Guespin, Jean-Baptiste Marcellesi, Denise Maldidier et Denis Slatka. pp. 87-134.
- Stefanini ; J. (1992). Quelques remarques sur la notion d'incidence. In : *Linguistique et Langue Française*. Paris : éditions du CNRS, pp. 193-201.
- Σύγχρονο Ελληνογαλλικό Λεξικό, *Dictionnaire grec-français*. (1998) Συγγραφική ομάδα : Βράτσου, Ε., Γαλάνης, Γ.Φ., Καραντζόλα, Ε., Παπάζογλου, Χ., Tonnet, H., Τσαμαδού- Jacobberger, Ει. Αθήνα: Εκδόσεις Πατάκη.
- Thavaud-Piton ;S. (2016). *Sémantique lexicale et psychomécanique guillaumienne*. Limoges :Lambert-Lucas.
- Trésor de la langue française, dictionnaire de la langue du XIX<sup>e</sup> et du XX<sup>e</sup> siècles (1789 –1960)*, publié sous la direction de Paul Imbs, tomes 1 à 14. Paris : éd. du C.N.R.S., 1971, tomes 15 et 16, sous la direction de Bernard Quemada, INaLF. Paris : Gallimard, 1994.
- Valette, M. (2007). Remarques sur la genèse du concept d'effectation chez Gustave Guillaume. In *Psychomécanique du langage et linguistiques cognitives*, Actes du XI<sup>e</sup> Colloque International de l'AIPL, Montpellier (France), les 8, 9 et 10 juin 2006. Limoges :Lambert-Lucas, pp. 99-108.
- Vassant, A. (1998). De la théorie de l'incidence, encore. In : *La ligne claire: De la linguistique à la grammaire : mélanges offerts à Marc Wilmet à l'occasion de son 60<sup>e</sup> anniversaire*, Paris-Bruxelles, Duculot, coll : Champs linguistiques, pp. 355-366.
- Vassant, A. (2005), Dire quelque chose de quelque chose ou de quelqu'un et la théorie de l'incidence de Gustave Guillaume. In : *Langue française*, n°147, La langue française au prisme de la psychomécanique du langage. Héritages, hypothèses et controverses. Paris : Larousse, pp. 40-67.







# Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations

Tavast A., Koppel K., Langemets M., Kallas J.

*Institute of the Estonian Language, Tallinn, Estonia*

## Abstract

We report on the ongoing project of developing the Ekilex dictionary writing system and joining existing dictionaries into the EKI Combined Dictionary. To facilitate the joining, several tools have been developed to solve data quality issues and turn textual data into structured entities. The resulting superdictionary thus contains various sets of information, which we call layers, either transformed from existing dictionaries or authored already in Ekilex. Our current focus is on the layers for synonyms and equivalents, which we describe in terms of their data model, lexicographic processes and lexicographer feedback from the first six months of Ekilex in production. As it turns out, the layer system may need expanding to accommodate an ever-growing list of requirements. The unidirectional data model for synonyms fully conforms to its design specification and received favourable first impressions, but extended use has started to cast doubt on the optimality of the model. We describe the pros and cons of this model and possible alternatives.

**Keywords:** synonyms; equivalents; data modelling; unified dictionary

## 1 Introduction

The goal of the Ekilex project (Koppel et al. 2019; Tavast et al. 2018) is to join dictionaries into a single superdictionary, the EKI Combined Dictionary (EKI ühendsõnastik, CombiDic), as opposed to linking between dictionaries or aggregated search across dictionaries (Boelhouwer, Dykstra & Sijens 2017). The underlying assumption is that users look for information about words, not about dictionaries, which means that the current system of multiple dictionaries with duplicated and conflicting information is not desirable.

Timing of the project also coincides with the rise of automated, corpus-based processes to replace introspective lexicography (Gantar, Kosem & Krek 2016; Kallas et al. 2019) as well as training lexicographers to pay more attention to the modelling of lexicographic data. Continued development of the superdictionary is an integral part of the project, so the goals are to: 1) join existing dictionaries, 2) create technical and administrative incentives for authors to cooperate, 3) improve the superdictionary to provide a radically better lexical resource for the user.

Despite a consensus about user benefits, these ideological and process-related changes are difficult for lexicographers, due to four interconnected reasons:

- Bringing a legacy dictionary into a structured database exposes its internal conflicts, previously hidden in disconnected entries. Doing the same with a number of dictionaries additionally exposes duplication and conflicts across the dictionaries. The result looks hideous, especially in a traditionally compiled bilingual dictionary trying to fulfil the needs of all conceivable users, where the target language equivalents have been a long list of (partial) translational equivalents fitting many different specific translation contexts. Gathering such occurrences to form a *word* entity mercilessly displays them side by side, which is not a pleasant sight for the authors.
- While specialised tools (described below) can be developed to assist in resolving these data quality issues, it is still largely manual lexicographic work. Given the decades that have gone into compiling the original dictionaries, the volume of this work looks daunting if not unrealistic. Lexicographers also rightly feel that their previous work is not sufficiently respected, and they are forced to start over from scratch.
- Especially in combination with the descriptivism of corpus-based lexicography, this necessitates a shift in thinking. Even if one would ideologically still prefer the old system, it is simply not feasible due to the workload involved. Responsibility gets transferred from the lexicographer, announcing the truth, to the reader, making sense of messy empirical data. While agreeing theoretically that it is better to be broadly right than precisely wrong, in practice authors feel uncomfortable with allowing uncertainty in a dictionary and trusting readers to draw their own conclusions.
- Previously autonomous dictionary working groups, now united into a large group working on layers of a single central dictionary, trust each other to varying degrees. There may also be differences in the lexicographic principles followed by each group. Unifying those principles and achieving trust is an administrative challenge.

Ekilex aims to make the shift easier by delivering tangible benefits for lexicographers, moving processes towards more automation on the continuum between manual authoring and fully automated generation of dictionaries. Development is ongoing and iterative, meaning that tasks are continually adjusted to lessons learned and insights discovered. At the time of writing, we are able to report on two relatively straightforward batch processes (word joiner and meaning joiner), but the main focus of this paper is a specialised tool for synonyms and translational equivalents. After listing some prerequisites in section 2, the specialised tool will be described in the rest of the sections.



## 2 Prerequisites

### 2.1 The Ekilex Data Model

Let us first briefly describe the Ekilex data model (Tavast et al. 2018). The central idea is that *word* and *meaning* are connected through *lexeme*, to express a many-to-many relation between words and meanings, see Figure 1. In the text, we use monospaced font to refer to database entities.

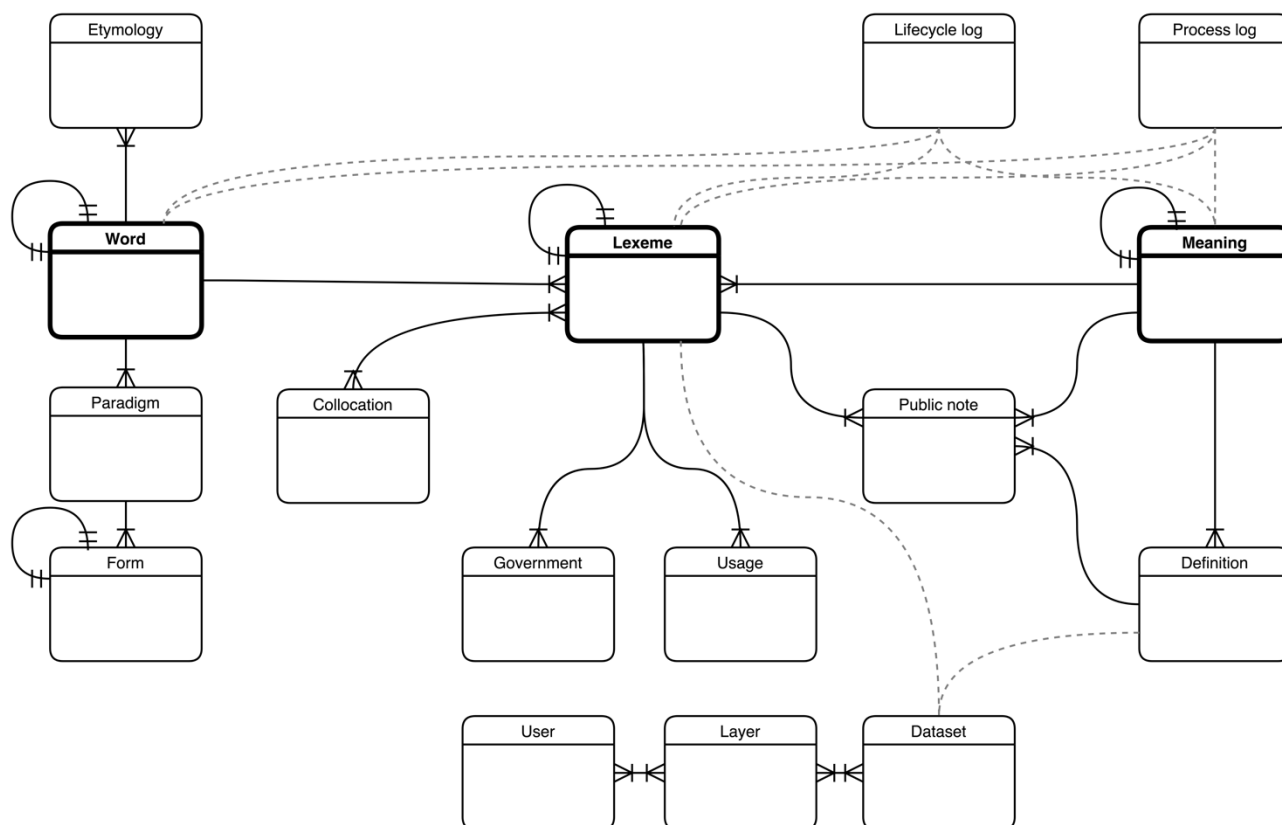


Figure 1: Simplified logical data model of Ekilex. Central entities are highlighted in bold.

- The *word* is an object in language, characterised mainly by its character composition, morphology and etymology. We make a distinction between homonyms (separate *words* with identical character composition) and polysemes (a single word with multiple meanings).
- The *meaning* is an object in cognition, characterised in the database by its domain(s), definition(s) and any related notes.
- The *lexeme* is an object in the dictionary, expressing the connection between *word* and *meaning*. It could be defined as "this word in this meaning as described in this dictionary". It contains information peculiar to the word-meaning combination like part of speech and example sentences, as well as dictionary-specific information like administrative status of the entry.

To refer to the user perspective as opposed to the data model, we also use traditional lexicographic terminology like *entry* (a record in a semasiologically organised dictionary, for us a word entity plus all related entities) and *headword* (a word that has such an entry). Similarly, since Ekilex is also used for terminology work, terminologists have their own *entries* (a record in an onomasiologically organised termbase, for us a meaning entity plus all related entities), *concepts* (a meaning having such a record) and *terms* (a word in a termbase). Viewpoints of the database and the user are distinguished elsewhere too: a *synset* for a user is a set of synonyms (words), which in database terms is described as a meaning that all these words are connected to. A headword may have several *senses*, each represented by a combination of lexeme and meaning in the database. A *lexical resource*, which may be a dictionary or a termbase for the user, is a dataset entity in the database.

### 2.2 Tools for Data Quality Improvement

Ekilex first obtained its data from importing dictionaries from previous dictionary writing systems, each with its own words and meanings, resulting in massive duplication of both. We used batch joiners to help mitigate this data quality issue.

Words were mostly character strings in the imported dictionaries, so the importer had no way of distinguishing between yet another occurrence of a previously found word and a new legitimate homonym. Only the Dictionary of Estonian 2019



(DicEst 2019; Langemets et al. 2018) had treated homonyms systematically. We assumed that a large general language dictionary like DicEst would have found all homonyms. Or conversely, if a word is not homonymous in DicEst, we could safely assume that all such character strings are occurrences of a single word and can be combined into a single *word* entity. Based on this assumption, the word joiner took care of 87,013 duplicated but really non-homonymous Estonian word types imported from multiple dictionaries.

Legitimate homonyms needed manual disambiguation, most of which was done before importing using specialised tools built by Indrek Hein, a senior developer at the Institute of the Estonian Language.<sup>1</sup> Manual joining of homonyms is also possible in Ekilex, and this is being done as part of normal dictionary compilation or editing. A total of 1,080 homonymous word types have been disambiguated manually, and it has taken about 15 person-days.

A similar approach was used for mapping meanings across component dictionaries of the CombiDic (see Koppel et al. 2019 for details). If a word was monosemous in DicEst, its meaning was joined with meanings of the same word from other component dictionaries where it was also monosemous. We found 57,461 such monosemes in DicEst and connected to them 76,845 meanings from other component dictionaries.

Meanings have also been joined manually, both before import using the same specialised tools, and already in Ekilex as part of the normal lexicographic workflow. Unlike homonyms with clear (even if theory-dependent) criteria for deciding whether two words are the same or not, meanings are completely open to human judgement, therefore also disagreements between authors of different dictionaries. This process is ongoing, much more time-consuming than the 15 person-days of homonyms, and can cause fundamental problems as we will show below.

The last batch tool so far, also with the smallest effect, joined homonyms for other languages. We don't have a similar gold standard for homonymy in other languages as DicEst is in Estonian, therefore we can only guess based on various hints. One such hint has been used, namely that if foreign words with the same form have the same Estonian equivalent, then they are most probably one and the same word, and can be combined. This took care of 12,882 foreign word types.

### 3 Layers

Uniting previously separate dictionaries into a single CombiDic does not entail, at least not initially, a lack of distinction between types of lexicographic information originating from the component datasets, or even a consolidation of the groups of authors. Lexicographers are still working on separate or at best partially overlapping tasks in their respective projects and entering their own data elements (e.g. synonyms, equivalents, usage examples, normative recommendations). The difference is that since September 2019 all this information now ends up in the same headword entry of CombiDic (i.e. data connected to a single word in the data model, see Figure 1), together with other data types imported from existing dictionaries like etymology, morphology and collocations.

To manage this agglomeration, we use the concept of layers. A layer belongs to one or more datasets, provides a coherent set of data elements, is accessible to and authored by a specific team of lexicographers, and has its own process status to track the team's progress. The idea is to allow multiple teams to contribute their expertise to entries of CombiDic, seeing each other's work, but not being overly disturbed by changes made by other teams.

The following layers are being actively authored in 2020:

- Partial synonyms. We will describe their data sources, data model and authoring process in section 4.
- Russian equivalents. While the synonym process is also applicable to equivalents, it is not used in the particular case of Russian. The reason is that rich information (meaning divisions of the equivalents) is already available from component datasets, which makes it easier to simply join existing meanings across the components in Ekilex.
- Normative recommendations, which will allow a specifically filtered view of CombiDic to replace the revered normative Dictionary of Standard Estonian ÕS (*Eesti õigekeelsussõnaraamat* 2018).<sup>2</sup>

At the time of writing, the current challenge with layers is that both their nomenclature and expectations towards them are growing rapidly. As the latest development, it has become evident that new ad hoc layers need to be created on the fly. The reason is that the lexicographic process is usually not random but organised by distinct tasks even within one team of authors. Each task starts from some kind of search result or list of entries that the lexicographer needs to check. As work progresses, items on the list need to be checked off one by one. The problem is that there may be any number of such lists, and both the lists and their progress status need to be managed somehow. It is also not known in advance which teams want to see the status of which (sub)layers, as the work of a neighbouring team may or may not be relevant for the task at hand. A practical example of when it does become relevant: when the CombiDic core team changes the meaning distribution of a word, then this should reopen the headword for several other teams to update their information accordingly. This is still work in progress without an agreed solution so far.

### 4 (Partial) Identity of Meaning

Words and meanings, the two central data elements of a lexical resource, differ in how well established their representation in lexical resources is. Words are straightforward to write down as a character sequence, and there is very

<sup>1</sup> <http://www.eki.ee/dict/selgroog/> [30/05/2020]

<sup>2</sup> The centre of the Estonian dictionary publishing tradition has been formed by two large, competing, partially duplicating and partially conflicting general dictionaries, the descriptive Dictionary of Estonian 2019 (DicEst) and the normative Dictionary of Standard Estonian ÕS 2018. In an effort enabled by and parallel to the Ekilex project, both will be merged into the CombiDic. Special treatment of the normative layer is needed due to the legal status of the ÕS in normative situations like exam grading.



little room for disagreement about how to do that in most languages. The common practice to organise dictionaries alphabetically also provides a widely accepted (even if arbitrary) similarity metric: words sharing initial characters are treated as belonging together. Meanings, on the other hand, lack a physical form that would simultaneously be human-readable,<sup>3</sup> sufficiently debate-proof to be usable in practice, and capture which other meanings this meaning is similar to.

When designing Ekilex, the objective was to be able to represent both full and partial identity of meaning, both within a language and across languages. Full identity, a notoriously debatable concept, is here defined as a function of the particular dataset and lexicographic judgement: two meanings are identical (i.e. they are really one meaning) if the lexicographer decides not to distinguish between different shades of meanings, but to enter their words as full synonyms in one language or exact equivalents across languages.<sup>4</sup> This judgement can change in time and vary across datasets of different sizes or objectives, but within the process of authoring a particular headword entry it can be treated as constant. The design objective also included the ability to represent partial identity or similarity of meanings, which is needed for partial synonyms and non-exact equivalents.

In the following, we first describe requirements of representing meaning similarity from the lexicographer's standpoint, then discuss conceivable solutions in a lexical database, and finally the approach(es) taken in Ekilex.

## 4.1 Requirements

As the defining characteristic of full synonyms and exact equivalents is that of having the same meaning, the proper way to represent them is to connect those *word* entities to the same meaning entity. As far as we have the data from existing dictionaries, this has already been done, and can further be done in the Ekilex user interface. Being connected to the same meaning, such synonyms and equivalents are direction-agnostic from the lexicographer's viewpoint: if  $a = b$ , then inevitably  $b = a$ .

For several reasons however, practical lexicography has a strong tradition of directionality. Lexicographers want to express that  $a = b$  without necessarily taking a stand on whether  $b = a$  or not. The whole concept of reversing bilingual dictionaries (Krek, Šorli & Kocjančič 2008) is based on the premise that equivalence is directional. Collocation dictionaries (e.g. Kallas et al. 2015) are another directional example, listing collocations according to their frequency or salience relative to the headword, not to the collocate. Earlier dictionary projects may even have been planned to remain unidirectional. E.g. if the objective was to present synonyms for 10,000 most frequent words but the synonyms were not restricted to the same frequency class, then there would have been tens of thousands of words in the synonym dictionary without synonyms of their own. Removing such restrictions has only been made possible by including synonyms as a layer in CombiDic and semi-automatic compilation.

The preference for directionality is amplified by an aspect of using empirical data from corpora, namely quantification. Word-level alignment of parallel corpora (for equivalents) and distributional semantics (for synonyms and equivalents) yield quantitative measures of how close the meanings are. Exact matches occur rarely, if at all, which complicates the picture to a level where the feasibility of ideal directionlessness is no longer beyond doubt. Especially as dictionaries become more comprehensive (and CombiDic is an attempt to maximise comprehensiveness), it is increasingly the norm that meaning identity is not exact, but some subtle differences need to be explicated.

Regarding synonymy and equivalence between polysemes, there is also a pragmatic workflow consideration. When starting to design the Ekilex module for synonyms, lexicographers expressed a strong preference to avoid the rabbit hole of chained relations, and instead complete the compiling of one headword before moving on to the next. Given the sense distribution of the current headword, they wanted to be able to add synonyms and equivalents to each of its meanings, without (yet) taking a stand on the sense distributions of the words added. For example,<sup>5</sup> when finding synonyms for the (sub)senses of the headword *board*, the lexicographer wants to connect the word *plank* to the 'piece of timber' meaning and the word *management* to the 'governing body' meaning, but not to select the correct sense for *plank* or *management*, even if these have other senses totally unrelated to *board*.

This preference entails the need to visit each similarity relation twice, entering *plank* as a synonym for *board*, and then separately entering *board* as a synonym for *plank*. To generalise, the number of required visits to a set of synonyms (a synset, to use the Wordnet term) equals the number of members in the set. Suppose we decide to consider *board*, *committee*, *management* and *directoriate* synonymous in our dictionary, this synset needs to be visited four times, each time adding three synonyms. In the design phase, lexicographers were confident that this is an acceptable trade-off for keeping their habitual headword-based process, as opposed to meaning- or synset-based (like in Wordnets or termbases).

## 4.2 Data Sources

Compilation of the synonym layer of CombiDic follows the semi-automatic method where lexicographers post-edit automatically generated lists of synonym candidates. The candidates were extracted from existing dictionaries, including component datasets of CombiDic itself, taking advantage of the tradition to include synonyms in the definition and other fields, as well as semantic mirroring (Dyvik 1998, 2004). Distributional similarity (Turney & Pantel 2010) has so far only

<sup>3</sup> As opposed to machine-readable. Since Ekilex is an information system, all of its contents, including any representations of meanings, are machine-readable by definition.

<sup>4</sup> Other lexicographically relevant aspects of synonymy and equivalence, like style, register or frequency (see e.g. Yong & Peng 2007: 129–131), are properties of the *lexeme* entity in Ekilex, which makes them a separate discussion. Here we concentrate on identity or similarity of meanings as characterised by their definitions and domains.

<sup>5</sup> In the following, we use simplified examples in English to improve readability. CombiDic, including its synonym layer, starts from Estonian. English is not yet among the languages of CombiDic, but adding it is near the top of the wish list.



been calculated from the multilingual FastText model (Grave et al. 2018), with the new Estonian National Corpus 2019 (Kallas & Koppel 2019) being in the queue.

Since equivalents are like synonyms, only in another language, this synonym process can be extended to bilingual dictionaries with practically no modifications. We take candidate equivalents from wherever they can be found, including corpora and existing lexical resources, and make them available for the lexicographer to connect to meanings as described above.

### 4.3 Representation in the Data Model

In the simplest case, if synonyms and equivalents are deemed to have identical meanings, the corresponding words can literally be connected to the same meaning entity. Figure 2 shows the situation where *board* and *management* share the meaning of 'a governing body', while each word has other meanings too. Examples of the other meanings are shown greyed out, and any further synonyms in those meanings are omitted completely.

This is the pervasive approach in termbases, where it is known as concept-based or onomasiological<sup>6</sup> (Wüster 1979; Felber 1984; see also Tavast 2008). For general language, it is used in Wordnets (Fellbaum 1998). The obvious benefit is simplicity, both technically and conceptually: meanings are identical or not, there is no third option or gradation. However, synonymy and equivalence are necessarily directionless in this model, which is contrary to the lexicographic understanding of language. Especially bilingual dictionaries need to represent partial equivalents, which is not possible using this simple model. By forcing lexicographers to take a stand about the meaning distribution of both words simultaneously, it is also in conflict with the design objectives described in section 4.1 above. Finally, it is difficult, although not impossible, to quantify the degree to which each word denotes the meaning, but long-term goals of Ekilex include empirical quantification of as many pieces of information as possible.

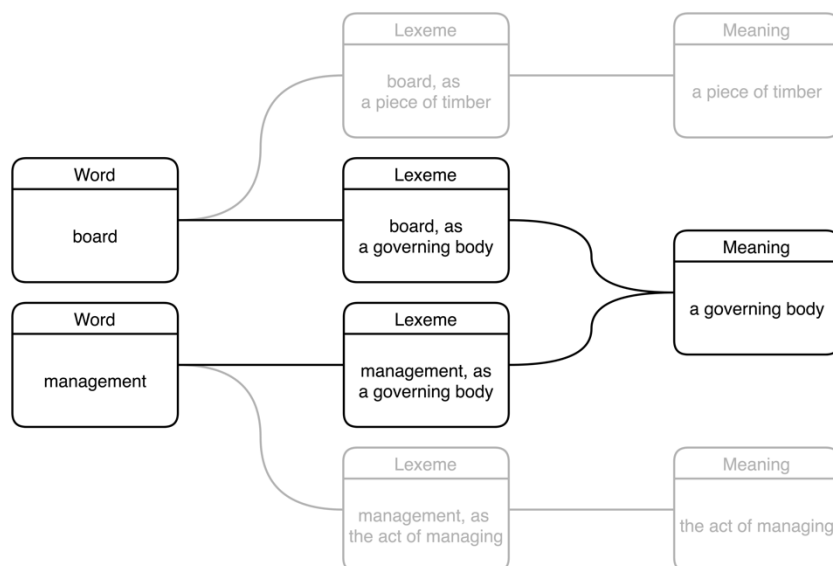


Figure 2: A single meaning: if *board* and *management* are considered full synonyms, they can be connected to the same meaning.

Therefore, this simplest model is only sufficient for full synonyms and exact equivalents, for which it is also used in Ekilex. It alone cannot represent partial synonyms or equivalents; neither does it accommodate lexicographers' preference for a directional, headword-based working process.

The next step is to have a separate meaning for each word, and link those meanings with (possibly weighted)<sup>7</sup> similarity relations (cf. Rudnicka et al. 2019). This is shown on Figure 3, where the 'governing body' meaning has been split in two and then reconnected with a similarity relation with a high similarity value. Taking this approach to the extreme by *never* allowing a meaning to have more than one word would make the lexeme entity redundant and reduce the data model to semasiology, which would be unacceptable for terminological users of Ekilex. This approach does, however, work seamlessly together with the single-meaning model above, so that full synonyms share a meaning (and terms share a concept), while providing the additional capacity to represent partial synonyms as meaning relations.

The similarity relations could further be made directional, which would allow describing situations where the similarity of *a* to *b* is different from the similarity of *b* to *a*, or one of the directions is absent altogether. This is a step in the right

<sup>6</sup> Pure onomasiology would treat all words as homonyms rather than polysemes, i.e. there would be two *boards* and two *managements* in the figure. This distinction is omitted here for simplicity. Incidentally, since Ekilex is used for both general language dictionaries and termbases, Ekilex users are also shielded from this distinction. The same words can be shown as polysemes to lexicographers and as homonyms to terminologists.

<sup>7</sup> How to obtain the weights is a separate topic. They could be based on the lexicographer's introspection, distributional similarity measures from a corpus, a function of which previous dictionaries have listed the relation, etc., or any combination thereof. The point here is that weighting is possible, should it be desired.



direction but does still not address the main point of the directionality requirement of section 4.1 above, because the correct meaning still needs to be specified on both sides of the relation at the same time.

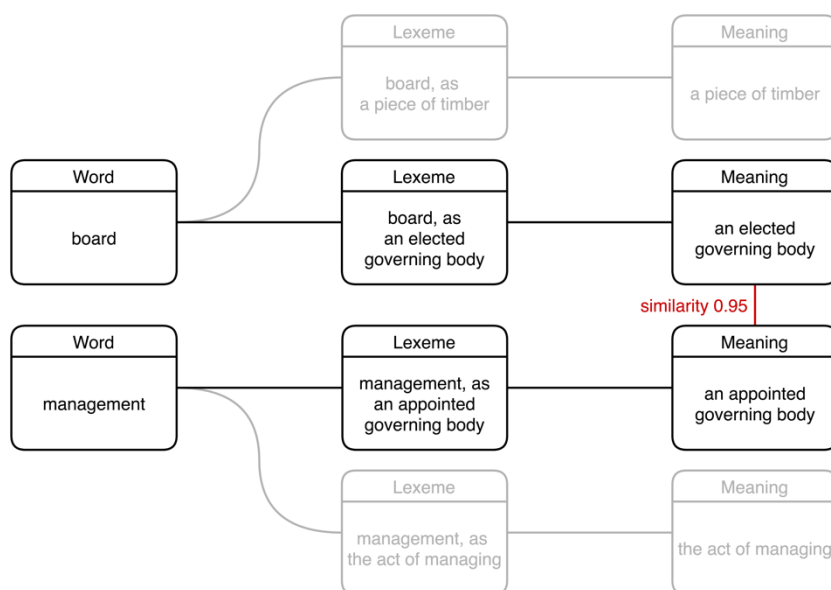


Figure 3: Related meanings: each word has its own meaning, and the meanings are related, with a similarity measure.

To cater for a completely directional, headword-based authoring process, Ekilex uses a special type of *lexeme*, the *secondary lexeme*, defined as "this meaning can *also* be expressed with that word". Figure 4 shows the result of a lexicographer working on the entry for *management* and adding the word *board* as a partial synonym to its first sense 'an appointed governing body'. The secondary lexemes can be weighted, so that a meaning can have stronger or weaker relations to many words in the same language (partial synonyms) or other languages (partial equivalents), like the weight 0.95 on Figure 4.

Note that *board* does not at this stage get *management* as a synonym. Theoretically it could get a new, third meaning through the secondary lexeme, but this behaviour was quickly ruled out based on feedback from lexicographers. To recap, adding *board* as a synonym in another entry leaves the entry for *board* itself completely unchanged. This is exactly the result that lexicographers requested: they only have to specify the meaning on one side of the relation.

Whether or not a corresponding partial synonymy relation needs to be added in the opposite direction, i.e. if *board* is a synonym for *management*, then whether *management* is also a synonym for *board*, will be decided only when the lexicographer reaches the other headword in the authoring process. This unidirectionality part of the requirements differs markedly from the habitual process of describing synonymy and equivalence used by other teams in EKI and elsewhere, and was intended by the synonyms team as a means of optimising their workflow. However, as discussed in section 5, using this solution for practical work has surfaced negative side effects that may motivate returning to the related meanings model.



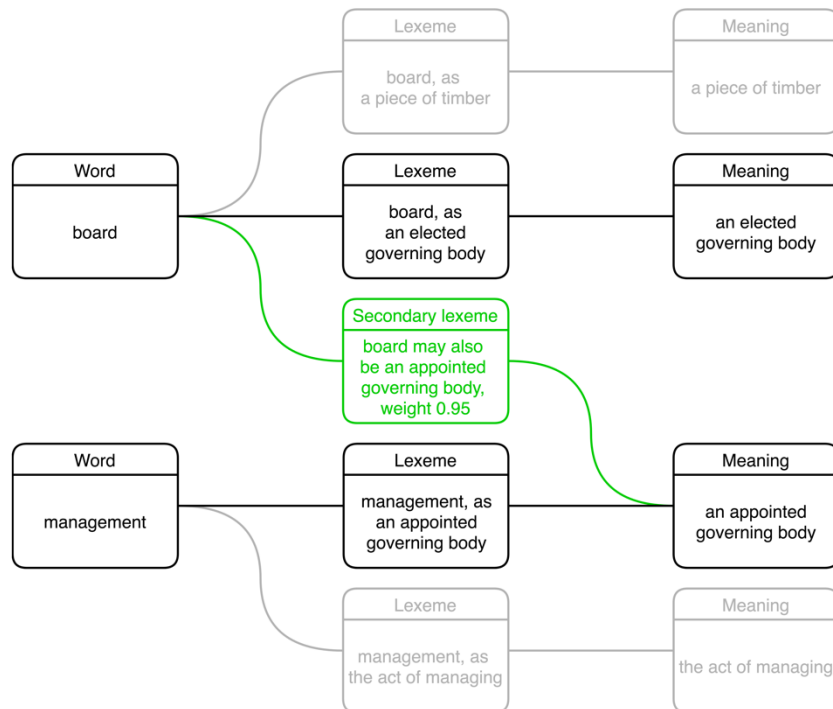


Figure 4: Unidirectional relation: *management* in its first meaning has a partial synonym.

Figure 5 shows the result after the lexicographer does decide to add the same synonym in the opposite direction, only with a different weight. There are now two independent and unrelated secondary lexemes, one for each direction, which again conforms exactly to initial requirements specification.



Figure 5: Several unidirectional relations: *management* in its first meaning and *board* in its second meaning each have a partial synonym.



## 5 Discussion and Conclusions

The experience of the first six months of compiling the synonym layer has shown that providing the lexicographer with an automatically generated list of synonym candidates makes the task of compiling an entry less time-consuming. The candidates are already there in the database, preventing the lexicographer from leaving the dictionary writing system to look up possible synonym candidates from existing dictionaries, thesauri, corpora, etc. What makes the specialised tool especially easy to use is the option of working on a keyboard instead of dragging and dropping the candidates to corresponding senses, as well as tooltips that display the definition of the words when hovering the mouse cursor over the candidates.

On the other hand, feedback from lexicographers over this extended period has provided valuable insights into the design choices, in some cases even casting doubt on the initial requirements.

- In an ideal world, layers as currently conceived would be sufficient to soften the transition from separate dictionaries to CombiDic, allow specialised teams to work on different aspects of the same entry and prevent conflicts. Reality has proven to be different in two ways: current layers are not really independent or even separated clearly enough, and lexicographers need a growing nomenclature of new (sub)layers. This necessitates a reconceptualisation of the layer system.
- The view used for synonym and equivalent layers in Ekilex is narrowly specialised for the simple repetitive task of connecting target words to source meanings. There are or will be other views for other tasks, including the clean-up task described above. Lexicographers, however, prefer to organise their work by headword, not by task, which necessitates either jumping between the specialised views, or adding more and more ad hoc functions to the views, thereby losing the ergonomics benefits of specialisation. We don't have a solution for this at the time of writing.
- While the described unidirectional approach of secondary lexemes exactly conforms to initial requirements and received favourable first impressions from lexicographers, doubts have started to emerge. Especially for large synsets, the need to enter all synonyms again for each member of the set has proven to be a significant drawback. Lexicographers have even submitted bug reports on the grounds that they remember having added a synonym, but the synonym is not there (admittedly, this confusion was amplified by deficiencies of the logging system of Ekilex at the time). Investigation then showed that indeed, the synonym was added, but to another member of the synset. This may mean that the conceptually and procedurally complicated approach of secondary lexemes is not justified after all, and it may be necessary to fall back on the related meanings approach.
- Another indication in the same direction is that for lexicographers, synonymy, antonymy and cohyponymy belong to the same category of semantic relations and should receive similar treatment. The current Ekilex data model differs from this categorisation by treating synonyms in a completely different way. Falling back on the related meanings approach would also even out this difference.
- In bilingual dictionaries, it has been the norm to allow sense distributions of the source headword to be influenced by the target language. In a central dictionary like the CombiDic, this is not sustainable, as there will eventually be many languages. The solution, again, has been agreed to be the related meanings approach described above: each language has its own meanings, and there are links of (partial) equivalence between meanings.
- Since we have limited information about homonyms in other languages, there are massive data quality issues in the target languages. Although some semi-automatic tools can be conceived, achieving a quality level comparable to that of Estonian words (a task that would traditionally be called "reversing" a dictionary) will employ lexicographers for a long period.

## 6 References

- Boelhouwer, B., Dykstra, A., & Sijens, H. (2017). Dictionary portals. In P. A. Fuertes-Olivera (ed.), *The Routledge handbook of lexicography*. London and New York: Routledge, pp. 754–766.
- Dyvik, H. (1998). A translational basis for semantics. *Language and Computers*, 24, 51–86.
- Dyvik, H. (2004). Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers*, 49(1), 311–326.
- Eesti keele sõnaraamat 2019 [The Dictionary of Estonian 2019, DicEst]. (2019). Tallinn: Eesti Keele Instituut. Retrieved from <https://doi.org/10.1515/3-00-0000-0000-08240L> [30.07.2020]
- Eesti õigekeelsussõnaraamat 2018 [Dictionary of Standard Estonian 2018, ÕS]. (2018). Tallinn: Eesti Keele Sihtasutus.
- EKI ühend sõnastik 2020 [EKI Combined Dictionary 2020, CombiDic]. (2020). Tallinn: Eesti Keele Instituut, Sõnaveeb. Retrieved from <https://sonaveeb.ee/> [30.07.2020]
- Felber, H. (1984). *Terminology Manual*. Paris: UNESCO.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Gantar, P., Kosem, I., & Krek, S. (2016). Discovering automated lexicography: The case of the Slovene lexical database. *International Journal of Lexicography*, 29(2), 200–225.
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. *ArXiv Preprint ArXiv:1802.06893*.
- Kallas, J., Kilgariff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., & Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*. Ljubljana; Brighton: Trojina, Institute for Applied Slovene Studies; Lexical Computing Ltd.



- Kallas, J., Koeva, S., Langemets, M., Tiberius, C., & Kosem, I. (2019). Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Lexical Computing, pp. 519–536.
- Kallas, J., & Koppel, K. (2019). Eesti keele ühendkorpus 2019 [Estonian National Corpus 2019]. Retrieved July 30, 2020, from <https://doi.org/10.15155/3-00-0000-0000-08489L> [30.07.2020]
- Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal*. Lexical Computing, pp. 1–3.
- Krek, S., Šorli, M., & Kocjančič, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. In *Proceedings of the XII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra*, pp. 535–542.
- Langemets, M., Tiits, M., Uibo, U., Valdre, T., & Voll, P. (2018). Eesti keel uues kuues. Eesti keele sõnaraamat 2018. *Keel ja Kirjandus*, 12, 942–958.
- Rudnicka, E., Piasecki, M., Bond, F., Grabowski, L., & Piotrowski, T. (2019). Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography*, 32(3), 296–325.
- Tavast, A. (2008). *The Translator is Human Too: A Case for Instrumentalism in Multilingual Specialised Communication*. Tartu: Tartu Ülikooli Kirjastus.
- Tavast, A., Langemets, M., Kallas, J., & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data : The Case of EKILEX.
- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188.
- Wüster, E. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Dordrecht: Springer.
- Yong, H., & Peng, J. (2007). *Bilingual Lexicography from a Communicative Perspective*. John Benjamins Publishing.

### Acknowledgements

The creation of the dictionary and terminology database Ekilex was funded by EKI-ASTRA program (2016–2022). The creation and development of the portal Sõnaveeb was funded by the Digital Focus Program of the Ministry of Education and Research (2018–2021) and by EKI-ASTRA program (2016–2022). Technical support is provided by OÜ TripleDev. The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicography for Specialised Languages,  
Terminology and Terminography**







# Lemma Selection and Microstructure: Definitions and Semantic Relations of a Domain-Specific e-Dictionary of the Mathematical Field of Graph Theory

Kruse T., Heid U.

*University of Hildesheim, Germany*

## Abstract

We design a bilingual electronic dictionary for the mathematical domain of graph theory. The target group of the dictionary are students in the field, and the dictionary should support them in both cognitive and communicative situations. Therefore, it will not only provide equivalents but also an ontology of the terminology. The dictionary is based on a corpus and the lemmas are selected by combining results of automatic extraction tools with the work of expert raters. For the microstructure, a domain-specific scheme is developed and presented. The lemmas are divided into nine categories (one for adjectives, one for verbs and seven for nouns). In addition, we introduce thirteen semantic relations for which information can be given in the microstructure, depending on the category of the lemma. The microstructure items for each semantic relation are introduced by means of a specific indicator phrase, as the target group might not be acquainted with the linguistic terminology.

**Keywords:** LSP-dictionary; microstructure; lemma selection; mathematics; pedagogical lexicography

## 1 Introduction

We plan to develop a bilingual e-dictionary for the mathematical domain of graph theory. Besides the equivalents in German and English, information on the relations between the concepts of the domain will be given, as an ontology forms the backbone of the dictionary. The aim of the dictionary is to meet cognitive as well as communicative needs. Therefore, aspects of domain-specific and pedagogical lexicography have to be combined in this project. Later on, we plan to determine if and how the dictionary influences the LSP-skills of the students compared to usual aids like Wikipedia.

Of course, one may wonder why we do not put our effort into the digitization of one of the already existing mathematics dictionaries. This might be sensible if the aim of the project simply was to have the dictionary as a product. Our project, however, also includes the development of a method to find the lemmas and the conceptual and/or semantic relations using linguistic patterns which are typical for the language of the domain (cf. Kruse & Giacomini 2019). The aim is to develop a generalizable method which makes it easier to create electronic corpus-based dictionaries for other sub-domains of mathematics as well.

In this paper, we will present the current state of the dictionary development regarding lemma selection and microstructure. The results may be applied to future mathematics LSP dictionary projects as well. In Section 2 we give an outline regarding the target group of the pedagogical dictionary. Section 3 gives an overview of mathematical lexicography by introducing the work of Eisenreich (2008) on printed dictionaries and presenting already existing electronic dictionaries for the domain of mathematics. In Section 4, we present our corpus. Section 5 introduces the process of lemma selection and a category system for classifying the lemmas. Based on that, we outline in Section 6 the planned microstructure of the dictionary focusing on the presentation of definitions and of conceptual relations. A conclusion and an outlook towards future developments is given in Section 7.

## 2 The User Group

The intended user group of the dictionary are mathematics students attending lectures on graph theory. Therefore, we consider our planned dictionary as a pedagogical dictionary. As Tarp (2011) has pointed out, there is some discussion on the question under which conditions a dictionary might be considered as pedagogical. Nevertheless, we use the terminology introduced by Gouws (2010) and regard the target group of our dictionary as semi-experts, as they already have basic mathematical knowledge. We classify them as advanced learners in the specialized language of the domain. The dictionary should help them in cognitive and communicative situations (cf. Fuertes-Olivera & Tarp 2014; Tarp 2008): They have to read and understand papers in English, which is generally their L2; and they have to give presentations or write theses in German, which is generally their L1. The target group as well as the functions of the dictionary were already described in detail in Kruse & Giacomini (2019).

## 3 Lexicography and Mathematics

Eisenreich (2008) gives an overview of mathematics lexicography with a focus on printed dictionaries. Nevertheless, some of his results are also relevant for electronic dictionaries. He states that mathematics dictionaries tend to be out of



date rather soon after publication as constantly new terminology comes up. An electronic dictionary seems to be appropriate to deal with this obstacle, as it can be updated much more easily than a printed one. Further, Eisenreich (2008) recommends to focus only on a sub-domain of mathematics for writing a dictionary, since there exist several terms which have multiple meanings depending on the particular sub-domain. Recognizing this problem, our project is restricted to the sub-domain of graph theory.

It is difficult to make a clear separation between encyclopaedic and terminological works when dealing with LSP-dictionaries (Adamska-Salaciak 2012). There are overlaps between the two areas and they are not clearly distinguishable. Nevertheless, we try to give an overview of existing mathematics lexicographic works (see Table 1). The main focus is on German and English resources. The list is by far not complete, especially as there exist a lot of private projects. Eisenreich (2008) has divided his overview into the following categories: (1) monolingual dictionaries in the narrower sense, (2) overall mathematics reference works, (3) elementary mathematics for the general public, (4) multilingual dictionaries. The same categorization may be applied for electronic dictionaries, but it seems reasonable to divide category (3) into didactic resources for pupils or the lay public vs. works for an academic audience. Furthermore, we will merge categories (1) and (2), as they are difficult to distinguish. Additionally, we want to look at content and form as two different dimensions. Therefore, we first distinguish between monolingual, bilingual and multilingual resources. The second dimension concerns the lemma selection and the purpose: school, academic, general public. Thirdly, we checked whether the dictionaries cover our topic of graph theory. So, in all of the resources, we looked up the term *graph* to see to which degree graph theory is considered in the particular work. In the context of scientific textbooks about the domain of graph theory we expect to find this word describing a graph in the discrete sense, consisting of edges and nodes; whereas in school mathematics it will rather refer to the graph of a function, meaning its representation in the plane like the parabola for  $f(x)=x^2$ , because graph theory is not part of school education at the moment. This assumption turned out to be true: Graph theory is, if at all, only covered in dictionaries for academic purposes.

The category *purpose* is based on the self-portrayal of the dictionaries. Of course, there are a lot of reference books for mathematics, such as collections of formulas, which at present appear either in print or with increasing frequency in digital form (Schmidt-Thieme & Weigand 2015). But following the terms of Wiegand (1998) these are non-lexicographic resources. Therefore, they are left out of this overview. Private publishers are not named, companies are. Some resources also combine properties of a simple dictionary and a general learning tool.

Furthermore, this overview only contains dictionaries with a proper user interface. For example, PDF documents such as online word lists, are not part of this overview, as they cannot really count as electronic dictionaries. The considered dictionaries either offer monolingual definitions or a list of terms, but not both.

Another dictionary or rather an encyclopaedia not mentioned here is Wikipedia, as we only list works where the author was named. Wikipedia does not fulfil this criterion. As it is an open collaborative resource, it might be difficult to trust the information from an academic perspective.

## 4 Corpora

To compile our dictionary, we built two comparable corpora consisting of academic texts the students use during their studies of graph theory. Therefore, the selection was based on the bibliography used in the course as well as on a survey we carried out with mathematics students. In the survey we asked them which sources they use. The result was that most of them consult Wikipedia (Kruse & Giacomini 2019). However, in order to maintain the quality of the dictionary we only included scientific publications in our corpora.

The English corpus contains eight books and 26 scientific papers (about one million tokens) and the German corpus consists of the lecture notes as well as of (parts of) nine textbooks (about 700.000 tokens). Each corpus comprises about 30.000 word types.

One obstacle in the creation process of the corpus was to deal with mathematical formulas. Due to different source file formats it was not possible to use a single workflow. Therefore, one has to keep in mind that as a result of these differences, the number of tokens for the same formula may vary in different texts. Yet, this is of no concern as the focus of this project is on the terminology and not on the formulas or the corpus itself.

## 5 Lemma Selection and Semantic Categorization

Our process of corpus-based lemma selection consists of different steps. We first extract definition patterns from the corpus which are typical for the mathematics language (Pagel & Schubotz 2014). Each of these patterns expresses a certain semantic relation which can be used in the further development of the dictionary (Kruse & Giacomini 2019). This pattern-based approach will be combined with data produced by other term extraction tools (e.g. Rösiger et al. 2016). The merged results are assessed by three expert raters (inter-rater reliability to be computed). This will lead to the final lemma list.

The selected lemmas will be classified according to nine different categories. The microstructure for the entry of each lemma will depend on the category of the lemma. The categories are: PARTS OF GRAPHS, TYPES OF GRAPHS, PROPERTIES OF GRAPHS, ALGORITHMS, MAPPINGS, THEOREMS, PROBLEMS, ACTIVITIES and PERSONS.

The dictionary will cover nouns, verbs and adjectives. The latter two are each assigned a single category, according to their respective function. Adjectives are used to express PROPERTIES OF GRAPHS, typically in the form of adjective+noun combinations. In the entries, we will distinguish cases where objects always have a given property (indicated to the user by the key phrase *X is always ADJ*) from those where an object may or may not have a given



property (*X can be ADJ*). There are rather few verbs with a terminological meaning in the domain of graph theory. They express ACTIVITIES (or states) and will be presented like in a valency dictionary, with an indication of the possible (categories of) subjects and complements. For example, the verb *inzidieren* ("be a neighbour of") has *Kante* ("arc") as a typical subject.

Nouns are classified by the categories TYPES, PARTS, ALGORITHMS, MAPPINGS, THEOREMS, PROBLEMS and PERSONS. Examples for TYPES OF GRAPHS are *tree* or *Petersen graph*. Our notion of TYPE OF GRAPH is based on the structure of the graphs (with/without circles, bridges, etc.). PARTS OF GRAPHS are lemmas such as *node*, *edge*, *path* – so all the objects of which a graph consists or rather all the terms being in a meronymic relationship with the term *graph* or with another lemma from the category TYPES OF GRAPHS.

The categories ALGORITHMS, MAPPINGS, THEOREMS and PROBLEMS should be self-explanatory. For example, they apply in cases that a theorem is given a proper name, such as the *Handshaking-Lemma*. Thus, not all theorems found in the corpus will have an entry in the dictionary.

PERSON NAMES are part of the dictionary, in case that a category, e.g. a THEOREM or a TYPE OF GRAPH, is named after a person. Probably, there will not be a lot of information on the persons available in the corpus. Therefore, we plan to link these entries to other databases dealing with mathematicians.

Name	Purpose	Form	Graph theory	Publisher	URL
Encyclopedia of Matheamtics	a	m EN	covered	Springer / European Mathematical Society	<a href="https://www.encyclopediaofmath.org/index.php/Main_Page">https://www.encyclopediaofmath.org/index.php/Main_Page</a>
Encyclopedia of Triangle Centers	a	m EN	no, other focus	private	<a href="https://faculty.evansville.edu/ck6/encyclopedia/glossary.htm">https://faculty.evansville.edu/ck6/encyclopedia/glossary.htm</a>
epi Wörterbuch	?	b DE-EN	no	private / spirito GmbH	<a href="http://www.informatik.oelinger.de/dictionary/index.html">http://www.informatik.oelinger.de/dictionary/index.html</a>
Illustrated Mathematics Dictionary	s	m EN	no	private	<a href="https://www.mathsisfun.com/definitions/index.htm">https://www.mathsisfun.com/definitions/index.htm</a>
Lexikon der Mathematik	a	m DE	covered	Guido Walz / Springer Spektrum	<a href="https://www.spektrum.de/lexikon/mathematik/">https://www.spektrum.de/lexikon/mathematik/</a>
Mathematik online Lexikon	a	m DE, m EN	covered	Universitäten Stuttgart und Ulm	<a href="https://mo.mathematik.uni-stuttgart.de/lexikon/">https://mo.mathematik.uni-stuttgart.de/lexikon/</a>
Mathematisches Lexikon	s, a	m DE	no	Universität Wien	<a href="https://www.mathe-online.at/mathint/lexikon">https://www.mathe-online.at/mathint/lexikon</a>
Mathematisches Wörterbuch / Math Dictionary	?	b DE-EN	no	private	<a href="https://www.henkede.de/maple/woerterbuch.htm">https://www.henkede.de/maple/woerterbuch.htm</a>
Math Glossary, Math Terms	?	m EN	covered	private	<a href="https://www.cut-the-knot.org/glossary/atop.shtml">https://www.cut-the-knot.org/glossary/atop.shtml</a>
Math spoken here	?	m EN	no	private	<a href="http://www.mathnstuff.com/math/spoken/here/1words/words.htm">http://www.mathnstuff.com/math/spoken/here/1words/words.htm</a>
Mathworld Wolfram	a	m EN	covered	Wolfram Research	<a href="https://mathworld.wolfram.com/">https://mathworld.wolfram.com/</a>
SchulMatheLexikon	s	m DE	no	Vorhilfe.de e.V.	<a href="https://www.matheraum.de/wissen/SchulMatheLexikon">https://www.matheraum.de/wissen/SchulMatheLexikon</a>
UniMatheLexikon	a	m DE	no	Vorhilfe.de e.V.	<a href="https://matheraum.de/wissen/UniMatheLexikon?mrsessionid=aa46eb31c21ae22eb45e2930f26a487c24689235">https://matheraum.de/wissen/UniMatheLexikon?mrsessionid=aa46eb31c21ae22eb45e2930f26a487c24689235</a>

Table 1: Electronic mathematics dictionaries. In the purpose column, academic is abbreviated to a, school to s; ? means that the purpose is not given. The form is described as either monolingual (m) or bilingual (b); the particular languages are indicated.

## 6 Microstructure

Our intended microstructure consists of two main parts: definitions and relations. Before we present their role in our dictionary, we give an overview of different types of definitions considered in lexicography based on the work of Lew and Dziemianko (2006).



## 6.1 Definitions

Lew and Dziemianko (2006) discuss three types of definitions which are used in lexicography: single clause *when*-definitions, contextual definitions and analytic definitions. We go through them and see how far they apply for our case and with which advantages and disadvantages they come.

### 6.1.1 Analytic Definitions

First, we examine analytic definitions (or logical definitions). They are the most classical ones following the Aristotelian scheme. Mathematical definitions in textbooks are generally written in the following defining format, cf.:

A graph  $G$  is an ordered pair  $(V(G), E(G))$  consisting of a set  $V(G)$  of *vertices* and a set  $E(G)$ , disjoint from  $V(G)$ , of *edges*, together with an *incidence function*  $\psi_G$  that associates with each edge of  $G$  an unordered pair of (not necessarily distinct) vertices of  $G$ . (Bondy & Murty 2008: 2; *emph. in original*)

This definition provides the genus proximum of *graph*, namely *ordered pair*. This definition style is almost always used for nouns.

Adamska-Sałaciak (2012) had a closer look at this kind of definitions and describes some downsides coming with their usage. The first problem she investigates is circularity because the genus proximum might be defined itself at some other place in the dictionary and in the end becomes the definiens. This is especially a problem in general language lexicography because not every word has a clear definition, e.g. due to connotation, collocational meaning, etc. But in terminology, especially in mathematics, this is not very likely to happen, as mathematics typically relies on definitions of the objects it works with, and on logical relationships between such objects. So, in our case there is no need to worry about this issue from a lexicographer's perspective.

Secondly, Adamska-Sałaciak (2012) deals with obscurity which occurs when the words in the definitions are even harder or less common in texts than the lemma itself. That might also apply for our dictionary since the user may have to look up words used in the definition, but as they are a prerequisite to understand the subject from a cognitive perspective, this is a risk we absolutely have to take.

Similarly, a third issue addressed by Adamska-Sałaciak (2012) will not be very likely to happen in mathematics for most of the lemmas: gaps in hierarchy resulting in missing hypernyms. If we go back to the basic definitions of a mathematical domain, the words used are taken from the general language, as is the case above with *pair*, of which the user should have an intuitive understanding. In general, most of the mathematical definitions rely on set theory which is basically an idea of the cognitive concept of being inside or outside something.<sup>1</sup> Nevertheless, not all definitions can be based on the indication of hypernyms: Adamska-Sałaciak (2012) suggests to use hyponyms in these cases. We will come back to this proposal in Section 6.2.

Another point of criticism are the abbreviations used by lexicographers which might not be always understandable to the user. Most of them date back to printed dictionaries which had a notorious lack of space. As we create an electronic dictionary, space is not a problem and such abbreviations will not be used.

### 6.1.2 Single Clause when-Definitions

With single clause *when*-definitions and full sentence definitions (FSD) a new format was established which is closer to the general language than analytic definitions. The beginning of this development might not date back to Aristotle, but even 30 years ago the following was stated:

Lexicographic definitions have a curious tendency not to stick in the mind, whereas the immediacy, the accessibility and the vividness of folk definitions often make them more memorable and consequently more likely to be of help in both decoding and encoding. (Stock 1986: 86f.)

What Stock (1986) here refers to as folk definitions were the bases for the development of FSD and single clause *when*-definitions.

According to Dziemianko and Lew (2006) single clause *when*-definitions are mostly used for the definition of nouns. In mathematical texts however, definitions with the use of *when* do not really appear. It is more common to use *if*, as in "A graph is *simple* if it has no loops or parallel edges" (Bondy & Murty 2008: 3; *emph. in original*). This definition style is mostly used to define properties of mathematical objects, expressed by means of adjectives. Thus, the actual definiendum is often an adjective+noun combination that denotes a subtype of a mathematical object, e.g. a certain type of graph.

Similarly, definitions of this form also appear to define verbs, e.g. in "If  $e$  is an edge and  $u$  and  $v$  are vertices such that  $\psi_G(e) = \{u, v\}$ , then  $e$  is said to *join*  $u$  and  $v$ " (Bondy & Murty, 2008: 2; *emph. in original*). Dziemianko and Lew (2013) and Lew and Dziemianko (2012, 2006) did several experiments on the usage of single clause *when*-definitions in pedagogical dictionaries and concluded:

One way in which dictionary users confronted with a single-clause definition might recognize that the definition defines a noun would be through their familiarity with the convention of using this definition type to explain nouns. The question is, however, to what extent this actually is a convention: can we be sure, for example, that such definitions are never used to define adjectives or verbs? There is no evidence to tell us this. (Lew and Dziemianko, 2012)

<sup>1</sup> For a comprehensive account of that idea see Lakoff & Núñez (2000).



As pointed out before, in mathematics the single-clause definitions are indeed used for adjectives and verbs. Therefore, we can conclude that mathematical definitions are somehow different (Vanetik et al. 2020). In our dictionary we will use the definition scheme established in mathematics.

### 6.1.3 Full Sentence Definitions (FSD)

The third type of definition are the FSD, which came up in mid 1990s when the Cobuild dictionary was published. The research carried out along with it mainly focuses on the acquisition of a foreign language (e.g. Allen 1996; Bogaards 1996; Herbst 1996). Though this definition form was highly praised, it did not really find its way directly into more dictionaries. Rundell (2006) tries to explain this fact as he is actually in favour of them: “They provide a much fuller picture of the target lexical items, yet without making unreasonable demands on users or requiring them to know any special conventions” (Rundell 2006: 326). This statement fully applies to our case, as our user group is not familiar with linguistic terminology or definition styles; thus, FSD may be a reasonable option. Nevertheless, Rundell (2006) also gives three major disadvantages of FSD: length, overspecification, and new conventions for old.

As our dictionary will be (only) electronic, length is not as important as for a printed dictionary because a clearly arranged layout can be used without any loss of space. Nevertheless, sentence length and sentence complexity should be kept in mind. Therefore, we will use indicator phrases in the microstructure which paraphrase the semantic relations by using expressions of general language. They are presented in detail in the next section.

Rundell (2006) also mentions anaphora resolution but as the target group will be familiar with either German or English or at least the grammar of an Indo-European language this can be ignored. Further arguments of Rundell (2006) against FSD address the general language and are not really relevant for the case of LSP. Additionally, as the mathematical definitions always include a specific meaning, overspecification is not an issue.

## 6.2 Relations

Having all this in mind, we will now take a look at the second part of the microstructure, the relations between mathematical objects. As stated above, in mathematics, semantic and logical relations tend to be equivalent. In other domains it might be necessary to distinguish these two levels carefully.

In our dictionary we will use a kind of FSD when we explain the relations, since the intended user group of the dictionary is not familiar with linguistic terminology. We paraphrase the relations using expressions of general language: synonyms (*is also called*), hypernyms (*is always a*) / hyponyms (*examples are*), meronyms (*is part of*) / holonyms (*is composed of*), eponyms (*is named after*), pertainyms (*linguistically related*), mapped to (*is usually mapped to* / *is canonically mapped to*), alternatives, attributes (*possible properties*), analogies (*is analogous to*). Which relations apply for each category is shown in Table 2.

Most of these relations are known from lexical semantics and used in our dictionary in the standard way, but there are also some domain-specific ones: *mapped to*, *alternatives* and *analogous to*. *Mapped to* means a mapping in the mathematical sense. For example, to each *edge* a *weight* can be assigned. We differentiate between *usually* and *canonically mapped to*. A *canonical mapping* is one that occurs because it follows from the way how graphs (or other objects of the domain) are defined. For example, each edge *is canonically mapped to* two nodes because this is how graphs are defined. In contrast, weights are only *usually mapped to* edges because not every edge needs to have a weight. The mappings can be defined between GRAPHS or their PARTS.

There are two other relations, *alternative* and *medium*, which are both related to ALGORITHMS. An alternative can only exist for ALGORITHMS: there can be two ALGORITHMS to reach the same goal. For example, both, *Fleury's algorithm* and *Hierholzer's algorithm* can be used to compute an *Euler tour*. But the terms are not synonymous as the ALGORITHMS apply different techniques to reach their goal.

Usually, the texts of our corpus contain textual definitions following the established scheme of mathematics for lemmas from the categories PARTS OF GRAPHS, TYPES OF GRAPHS, PROPERTIES OF GRAPHS, MAPPINGS and ACTIONS. ALGORITHMS, PERSONS, PROBLEMS and THEOREMS will not be defined (see above). As the adjectives always have a noun they refer to, they will be given as a lemma together with this noun. This uniqueness applies within a certain field: In German, for example, *vollständiger Graph* and the proof technique *vollständige Induktion* would be regarded as two different lemmas.

The underlying ontology structure is implemented in the Web Ontology Language (OWL)<sup>2</sup> using the editor Protégé (Musen 2015). Thereby, the categories are used as classes and the relations are used as object properties. Therefore, we will use some of the terminology from OWL in the remainder of this section. For each relation we can indicate a possible source category (domain) and a target category (range). For example, if we have a look at the eponymic relation, indicated by *is named after* only PERSONS are a possible range, whereas members of all the other categories can serve as the domain.

In addition, we can differentiate between symmetric and non-symmetric relations. In our case, equivalents, synonyms, pertainyms, antonyms, analogies, alternatives and mappings are symmetric relations, the others are not. Symmetric relations can only be established between members of the same category.

Not for each lemma from each category all relations are relevant and thus described in the dictionary. The equivalents are always given, as they constitute an essential part of the dictionary. Synonyms are given wherever possible. Hyper- and hyponyms are given for the defined lemmas. For the others, ALGORITHMS, PROBLEMS and PERSONS, they do not

<sup>2</sup> OWL is a W3C-Standard. More information can be found on their web page <https://www.w3.org/TR/owl2-overview/>



really exist in a way which is relevant for the project: each member of the category would have the same hypernym, namely the name of the category. Of course, the category itself will be visible within the microstructure (see Figure 1).

	ALGORITHMS	MAPPINGS	PARTS	PERSONS	PROBLEMS	THEOREMS	TYPES	PROPERTIES	ACTIVITIES
isEquivalentOf	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>
isSynonymOf	<u>DR</u>	<u>DR</u>	<u>DR</u>		<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>	<u>DR</u>
isHypernymOf		<u>DR</u>	<u>DR</u>				<u>DR</u>	<u>DR</u>	
isHyponymOf		<u>DR</u>	<u>DR</u>				<u>DR</u>	<u>DR</u>	
isHolonymOf			DR				D		
isMeronymOf			DR				R		
isPertonymOf	DR	DR	DR	DR	DR	DR	DR	DR	DR
isAntonymOf			<u>DR</u>				<u>DR</u>	<u>DR</u>	<u>DR</u>
isMediumTo	D	R	R				R	R	
isAnalogueTo		<u>DR</u>	<u>DR</u>				<u>DR</u>	<u>DR</u>	
isAlternativeTo	<u>DR</u>								
isAttributeTo	R	R	R		R		R	D	
isMappedTo			DR						
isEponymOf	R	R	R	D	R	R	R	R	R

Table 2: Categories and Relations. It is indicated whether the particular category serves as domain *D* or range *R* for each relation.

Underlined entries denote that the relation can only exist between members of the same category.

Another question is the order in which the relations should be presented in the dictionary. It might be useful to give the equivalent first or even visually marked, as the user often either wants an explanation or a translation. Next, it is useful to give synonyms as the users might recognize terms which they are already familiar with and therefore do not need any further explanations. The synonyms can be followed by the hypernyms in order for the user to learn that the concept looked up is a subtype or an example of another given concept. Mathematics language is structured in a strongly hierarchical way. Therefore, the given hypernym will always be the direct hypernym on the next higher level. Further information can be arranged in blocks which the user can open on demand. One block contains holonyms /meronyms, pertonyms and antonyms as they are linguistically related with the term. The other block provides information on domain-specific relations as it contains terms related as mediums, analogies, alternatives, attributes and mappings. The equivalents, synonyms, pertonyms and eponyms may be there for all the terms independently from the category.

### Euler-Tour - *Category: Parts*

**Definition** A closed trail in a graph which contains each edge exactly once.

**German equivalent** Eulertour

**Synonyms** Eulerian path, Eulerian trail

**...is always a** trail, path

**Examples** Eulerian circuit, Eulerian cycle

**...is composed of** edges, nodes

**...is part of** Eulerian graph

**...can be computed with** Hierholzer's algorithm, Fleury's algorithm

**...is analogous to** Hamiltonian path

**...is named after** Leonhard Euler

**...is linguistically related to** Eulerian graph, Euler

Figure 1: Showcase article. The items shown would be linked to the corresponding article.

## 7 Conclusion and Future Work

We have now shown a possible microstructure for an electronic LSP-dictionary for mathematics. This microstructure is



based on a category system we developed for classifying our lemmas. Our next step is to implement this structure and to fill it mostly automatically. The category system can be applied to other mathematical domains as well.

## 8 References

- Adamska-Salaciak, A. (2012). Dictionary definitions: problems and solutions. In *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 129(41), pp. 323-339.
- Allen, R. (1996). The big four: The year of the dictionaries. In *English Today*, 12(2), pp. 41-55.
- Bogaards, P. (1996). Dictionaries for Learners of English. In *International Journal of Lexicography*, 9(4), pp. 277-320.
- Bondy, A. & Murty, M. (2008). *Graph Theory*. London: Springer 2008.
- Dziemianko, A. & Lew, R. (2006). When you are Explaining the Meaning of a Word: The Effect of Abstract Noun Definition Format on Syntactic Class Identification. In E. Corino, C. Marelllo, C. Onesti (eds.) *Proceedings of the 12th EURALEX International Congress*, Torino, 6-9 September 2006. Torino: Edizioni dell'Orso, pp. 857-863.
- Dziemianko, A. & Lew, R. (2013). When-definitions revisited. In *International Journal of Lexicography*, 26(2), pp.154-175.
- Eisenreich, G. (2008). Die Fachlexikographie der Mathematik: eine Übersicht. In L. Hoffman & H. Kalverkämper & H. Wiegand (eds.) *Fachsprachen*. Berlin / New York: de Gruyter, pp. 1959-1966.
- Fuertes-Olivera, P. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminology*. Berlin / Boston: de Gruyter.
- Gouws, R. (2010). The Monolingual Specialised Dictionary for Learners. In P. Fuertes-Olivera (ed.) *Specialised Dictionaries for Learners*. Berlin / Boston: de Gruyter, pp. 55-68.
- Herbst, T. (1996). On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. In *International Journal of Lexicography*, 9(4), pp. 321-357.
- Kruse, T. & Giacomini, L. (2019). Planning a domain-specific electronic dictionary for the mathematical field of graph theory: definitional patterns and term variation. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, Sintra, 1-3 October 2019. Brno: Lexical Computing CZ, s.r.o., pp. 676-693.
- Lakoff, G. & Núñez, R. (2000). *Where mathematics comes from: how the embodied mind brings mathematics into being*. New York: Basic Books.
- Lew, R. & Dziemianko, A. (2006). A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and its Usefulness for Conveying Syntactic Information. In *International Journal of Lexicography*, 19(3), pp. 225-242.
- Lew, R. & Dziemianko, A. (2012). Single-clause when-definitions: take three. In R. V. Fjeld, J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*, Oslo, 7-12 August 2012. Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 997-1002.
- Musen, M. A. The Protégé project: A look back and a look forward. *AI Matters*, 1(4), pp. 4-12.
- Pagel, R. & Schubotz, M. (2014). Mathematical Language Processing Project. In M. England, J. H. Davenport, A. Kohlhasse, M. Kohlhasse, P. Libbrecht, W. Neuper, P. Quaresma, A. P. Sexton, P. Sojka, J. Urban, S. M. Watt (eds.) *Joint Proceedings of the MathUI, OpenMath and ThEdu Workshops and Work in Progress track at CICM*, Coimbra, 7-11 June 2014. Aachen: CEUR Workshop Proceedings.
- Rösiger, I. & Bettinger, J. & Schäfer, J. & Dorna, M. & Heid, U. (2016). Acquisition of semantic relations between terms: how far can we get with standard NLP tools? In *Proceedings of the 5th International Workshop on Computational Terminology, Computerm 2016*, Osaka, 12 December 2016, pp. 41-51.
- Rundell, M. (2006). More than one Way to Skin a Cat: Why Full-Sentence Definitions Have not Been Universally Adopted. In E. Corino, C. Marelllo, C. Onesti (eds.) *Proceedings of the 12th EURALEX International Congress*, Torino, 6-9 September 2006. Torino: Edizioni dell'Orso, pp. 323-337.
- Schmidt-Thieme, B. & Weigand, H. (2015). Medien. In R. Bruder et al. (eds.) *Handbuch der Mathematikdidaktik*. Berlin/Heidelberg: Springer Spektrum, pp. 461-490.
- Stock, P. (1986). The structure and function of definitions. In M. Snell-Horny (ed.) *ZüriLEX '86 Proceedings, Papers read at the EURALEX International Congress*, University of Zürich, 9-14 September 1986. Tübingen: Francke, pp. 81-89.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Tarp, S. (2011). Pedagogical Lexicography: Towards a New and Strict Typology Corresponding to the Present State-of-the-Art. In *Lexikos*, 21(1), pp. 217-231.
- Vanetik, N. & Litvak, M. & Shevchuk, S. & Reznik, L. (2020). Automated Discovery of Mathematical Definitions in Text. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, S. Piperidis (eds.) *Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC 2020*, Marseille, 11-16 May 2020, Paris: ELRA, pp. 2086-2094.
- Wiegand, H. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlin/New York: de Gruyter.







# A Thematic Dictionary for Doctor–Patient Communication: The Principles and Process of Compilation

Kudashev I.S., Semenova O.V.

Tampere University, Finland

## Abstract

Increasing internationalisation has resulted in a constantly growing need for community interpreting worldwide. Healthcare is one of the most challenging domains for community interpreters, as misunderstandings, especially those caused by the use of incorrect terminology, may cost lives. In this paper, we describe the process of planning and compiling the *Finnish-Russian Thematic Dictionary for Doctor-Patient Communication* aimed at professional community interpreters and university-level students of community interpreting. We start by describing the theoretical background of dictionary planning and analysing the information needs of the target groups. We then describe and justify the selection of dictionary sources as well as the mega-, macro-, and microstructure of the dictionary. The dictionary has been compiled using a tailored version of the in-house dictionary writing system MyTerMS. We briefly report the details of the technical implementation of the project. Finally, we reflect on some challenges encountered in this project as well as its future prospects. The dictionary can be further developed by increasing the volume of its disease-specific part, adding verbs and usage examples, and customising the electronic version for various target groups and purposes.

**Keywords:** medical dictionary; medical glossary; medical terminology; doctor-patient communication; community interpreting; healthcare interpreting

## 1 Introduction

Increasing internationalisation has resulted in a constantly growing need for community interpreting worldwide. In Finland, like in many other European countries, representatives of language minorities in many cases have the right to communicate with officials and public service providers in their native language, which in practice means that a community interpreter is invited to the meeting.

The quality of community interpreting has lately become a hot topic in interpreting studies (e.g. Hale 2007; Valero-Garcés 2008; Flores et al. 2012; Maley 2018). Using the correct terminology is a key factor in quality interpreting. Healthcare is one of the most critical domains in this respect, as misunderstandings, especially those caused by the use of incorrect terminology, may cost lives.

In Finland, the training of community interpreters has been systematically developed (Mäntynen 2013). However, there is still a lot of variation in interpreters' competence levels (Ollila 2017: 28), which is also the case in many other countries (Roat & Crezee 2015). The project “Developing Healthcare Interpreting Training” (2019–2020) aimed at improving the university-level training of community interpreters working in the healthcare sector in Finland. One of the major goals of the project was the compilation of the *Finnish-Russian Thematic Dictionary for Doctor-Patient Communication*. Russian was chosen as the first target language because it is the most requested language in community interpreting in Finland (cf. Koskinen, Vuori & Leminen 2018: 9).

In this paper, we describe the process of planning and compiling the dictionary; present its mega-, macro-, and microstructure; report on the technical implementation; and reflect on some of the challenges encountered in this project and the dictionary's future prospects.

## 2 Factors Affecting Dictionary Planning

We begin by describing the theoretical background of dictionary planning. Factors affecting the planning of any dictionary can be divided into the factors deriving from the target group's needs and the factors which reflect the restrictions of the outside world. They can be called lexicographic factors proper and external lexicographic factors, respectively (Kudashev 2007: 66).

While performing some task (e.g. healthcare interpreting), dictionary users (e.g. community interpreters) have some information needs (e.g. they need information about the target language equivalents and their usage), as well as needs related to information retrieval and processing (e.g. they need to find and process this information very fast in the field). However, the lexicographer compiling the dictionary in most cases does not know precisely what the potential user's needs will be in a particular communicative situation. The general picture the lexicographer has is only an approximation, which can be improved with the help of surveys and interviews with target group representatives and by the careful selection of dictionary sources. In this way, the process of dictionary planning is highly affected by the methodology of gathering information about the users' needs, the volume and quality of this information, and the methodology of source selection. These factors are summarised in Figure 1 (cf. Kudashev 2007: 68).



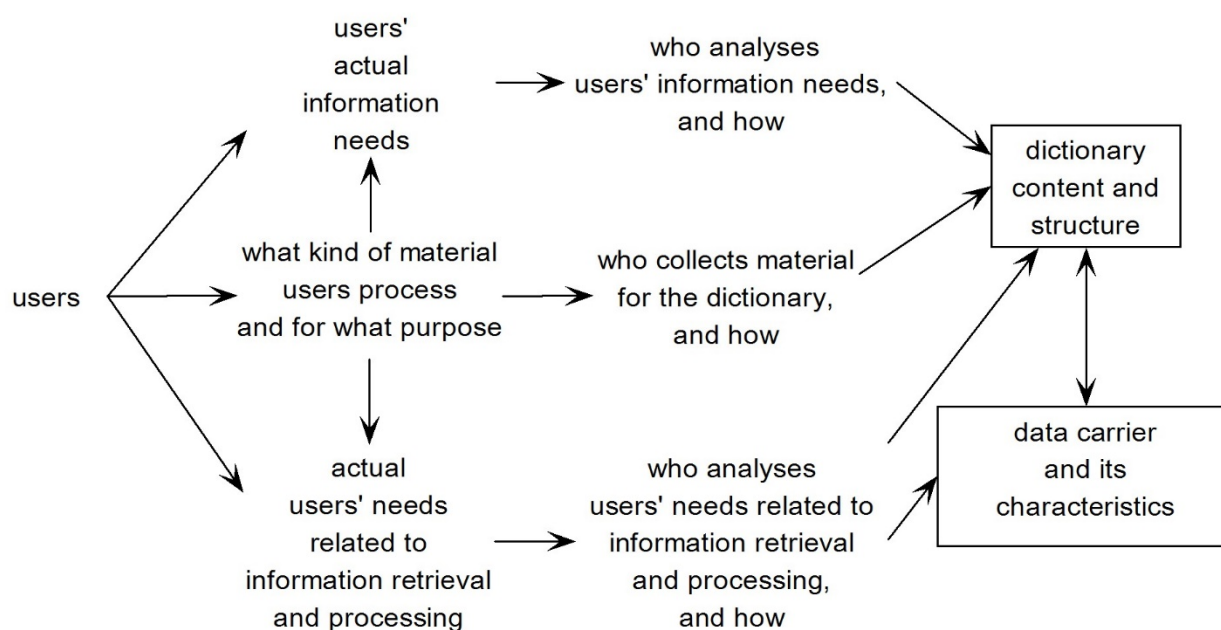


Figure 1: Factors Affecting Dictionary Planning.

External factors, in turn, are first of all related to the availability of resources, such as temporal, financial, human, and technical resources. Resource providers (e.g. sponsors, publishers) may set certain additional conditions of their own. Another external factor consists of the limitations imposed by the data carrier. The lexicographer's background, education, previous experience of dictionary-making, etc., may also affect the dictionary planning (Kudashev 2007: 71–73).

The target users' needs in our dictionary project, along with the methodology of gathering this information, are described in Section 3. The sources of the dictionary and the rationale behind their selection are described in Section 4. Among the external factors with a major impact on the dictionary project, we have to mention the tight schedule (one year) and limited budget, due to which we could only hire a part-time terminologist and only use volunteers from the healthcare sector as domain experts. However, we have managed to compile a dictionary with 4,000 entry words covering 30 common diseases.

### 3 Target Groups and Their Needs

Following the general theory of designing dictionaries of special languages (e.g. Bergenholtz & Tarp 1995; Fuertes-Olivera & Arribas-Baño 2008; Kudashev 2007), we started the project by analysing the information needs of the target groups. The dictionary's main target group comprises professional community interpreters and university-level students of community interpreting. Secondary target groups include patients, medical staff, and medical students.

Information about the target groups' needs was collected with the help of self-reflection and by studying professional literature on community interpreting in general and healthcare interpreting in particular. In addition, we interviewed Seija Koskinen, the Director of the Pirkanmaa Interpreter Centre (Tampere, Finland), and Girta Roots, a community interpreter working with the Finnish, Russian, and Estonian languages. Pirkanmaa Interpreter Centre provides community interpreting in the second most populated urban region in Finland. Almost 40% of the Centre's interpreting assignments are related to healthcare. Secondary target groups were represented by Galina Mäkäraäinen, a Russian-speaking medical doctor who participated in language courses for doctors with an immigrant background. Mäkäraäinen also consulted on various medical aspects and verified the equivalents. Students of the special course *Finnish-Russian Healthcare Interpreting* organised in the autumn semester of 2019 at Tampere University also gave their feedback on the first version of the dictionary.

Working community interpreters use medical dictionaries to revise medical terms while preparing for an assignment. They may also need the dictionary during the interpreting to check a term they do not know or do not remember. Students of interpreting use medical dictionaries for the same purposes, but the focus is on mastering the vocabulary and preparing for simulated interpreting sessions.

While preparing for a real or simulated assignment, both working interpreters and students familiarise themselves with the topic and study the terminology related to it. They cannot know beforehand what part of the material is going to be useful for interpreting the particular case (Veisbergs 2006: 1220). While students know the topic of the assignment beforehand, working interpreters do not always have this luxury due to the General Data Protection Regulation (GDPR) and similar regulations restricting access to confidential information, and they may therefore have very little time for preparation. To facilitate the quick learning of medical vocabulary by topic, the dictionary should be organised thematically and contain only the most relevant information (cf. Griniov-Grinevitš 2009: 68).

Students of interpreting have less background knowledge and lower linguistic competence than working interpreters. Consequently, they require more linguistic information about the terms and equivalents. For example, our experience has shown that Finnish-speaking students need information on the word stress for most Russian equivalents, while experienced



Finnish-speaking interpreters only want this information in difficult cases.

Interpreting assignments, for example doctor's appointments, are strictly limited in time, so a healthcare interpreter has very little time for checking terms. This implies that the dictionary must be well structured and have fast and convenient search tools.

The dictionary has two functions. The first one is pedagogical. Community interpreters and students must know some basic terminology by heart. As it is hardly possible to always remember more specific terminology related to each disease, they also need to revise disease-specific terminology while preparing themselves for a concrete assignment. The second function of the dictionary is referential, as interpreters may forget a particular term on the spot and want to check it. The dictionary must therefore be thematic and alphabetical in order to comply with both requirements.

#### 4 Dictionary Sources

The best source for a medical dictionary aimed at healthcare interpreters would probably have been recordings of authentic doctor's appointments. Unfortunately, this information is confidential, so getting access to it is problematic. The only possible alternative to authentic conversations is written sources, preferably aimed at doctor–patient communication. The following sources of Finnish terms were selected as the primary ones:

- evidence-based guidelines for patients (*Käypä hoito*) and descriptions of diseases published by the Finnish Medical Society, Duodecim (Terveyskirjasto 2020)
- international classifications and standards
  - International Classification of Diseases, 10<sup>th</sup> revision<sup>1</sup>
  - International Classification of Primary Care, 2<sup>nd</sup> edition<sup>2</sup>
  - Terminologia Anatomica (Kolesnikov 2003)
- Finnish classifications and ontologies
  - Finnish Classification of Surgical Procedures (Lehtonen, Lehtovirta, & Mäkelä-Bengs 2013)
  - Medicine Classification by the Finnish Medicines Agency FIMEA<sup>3</sup>
  - Finnish Ontology of Health and Welfare TERO<sup>4</sup>

The primary sources of Russian equivalents included:

- Evidence-based guidelines by the Ministry of Health of the Russian Federation<sup>5</sup>
- The Doctor's Handbook "2000 diseases from A to Z" (Denisov & Ševtšenko 2010)

As secondary sources of Finnish terms and Russian equivalents, we have also used course books and manuals for medical staff as well as mono- and bilingual dictionaries. As a rule, we have not used materials translated from other languages. However, at present, many classifications are international anyway. Besides, most materials on evidence-based medicine available in Russian are translations from English or other languages. For example, we have used a unique printed manual "Evidence-Based Medicine Guidelines" (Denisov, Ševtšenko, Kulakov, & Haitov 2002) translated from English. Translated texts, however, were used with care, and terms extracted from them were cross-checked in non-translated sources.

#### 5 Megastructure

The dictionary's front matter consists of the Preface and instructions on how to use the dictionary. The back matter contains the Russian-Finnish index and the list of sources.

The dictionary proper consists of two parts. Part 1 contains about 2,000 of the most common terms, which are supposed to be learned by heart. As terms are easier to learn thematically, they are grouped according to the domains listed below. Within each domain, the vocabulary is further divided into basic and advanced. For example, the basic level of anatomic terminology contains about 150 terms and the advanced level about 800. The domains in Part 1 are as follows:

- Healthcare system
- General medical vocabulary
- Diseases, symptoms, and complaints
- Anatomy and physiology
- Medical examinations
- Treatment and care, disease prevention, and rehabilitation
- Medication
- Medical equipment

The second part of the dictionary is disease-specific. Its function is twofold. First, it helps interpreters and students to prepare for the assignment related to a particular disease, even at short notice. The second function is referential. Even experienced interpreters – let alone students – cannot keep in mind all the terminology they may need. The second part of the dictionary provides them with disease-specific equivalents conveniently grouped into subdomains for quicker retrieval.

<sup>1</sup> <https://icd.who.int/browse10/2016/en> [09/05/2020]

<sup>2</sup> <https://www.who.int/classifications/icd/adaptations/icpc2/en/> [09/05/2020]

<sup>3</sup> <https://www.fimea.fi/> [09/05/2020]

<sup>4</sup> <https://finto.fi/tero/fi/> [09/05/2020]

<sup>5</sup> <http://cr.rosminzdrav.ru/#/> [09/05/2020]



Each disease-specific section in Part 2 is divided into the following thematic sections:

- Disease and its subclasses
- Associated diseases and diseases with similar symptoms
- Anatomy and physiology
- Symptoms and complaints
- Medical examinations
- Treatment and care, disease prevention, and rehabilitation
- Medication
- Medical equipment
- Miscellaneous

Terms from Part 2 may also be included in Part 1, which means that the two parts of the dictionary are not exclusive but partly overlap. In the current version of the dictionary, Part 2 covers about 30 of the most common diseases, which were selected on the basis of healthcare statistics by the Finnish Institute for Health and Welfare<sup>6</sup> and consultations with a domain expert. Domains in Parts 1 and 2 were selected by grouping terms extracted from medical texts into thematic classes and by performing a frame analysis of the communicative situation “doctor’s appointment” (cf. Madžajeva 2012; Gagarina 2012).

## 6 Macrostructure

Dictionary entries are organised into a table with two columns. Although the table format speeds up searches, the table borders should be light so as not to distract the user’s attention. The dictionary is divided into thematic sections as described above, but within all the sections of Part 1 and Part 2, the terms are arranged alphabetically. Thematic order supports the pedagogical function and alphabetical order the referential function.

Figure 2 demonstrates a portion of the dictionary from the disease-specific section *Reflux*. The first bolded caption is *Disease and its Subclasses*, and the second one is *Associated Diseases and Diseases with Similar Symptoms*.

<b>Osa 2 -- Refluksi -- Sairaus ja sen alatyypit</b>	
refluksi; refluksitauti	рефлюкс
refluksitauti → refluksi	рефлюкс
<b>Osa 2 -- Refluksi -- Liitännäissairaudet ja oireiltaan samantapaiset sairaudet</b>	
adenokarsinooma	аденокарцинома
ahtauma I; striktuura	сужение канала; стриктура
anemia	анемия; малокровие (разг.)
aspiraatio	аспирация
Barrettin epiteeli → Barrettin ruokatorvi	пищевод Барретта
Barrettin ruokatorvi; Barrettin epiteeli	пищевод Барретта
ylävatsavaivat mon.; dyspepsia; ruoansulatushäiriö; ruoansulatusvaivat mon. (ark.)	расстройство пищеварения; диспепсия
esofagiitti → ruokatorven tulehdus	эзофагит; воспаление слизистой оболочки пищевода
gastriitti → mahatulehdus	гастрит

Figure 2: Example of a Disease-specific Section.

By default, the synonyms of Finnish terms are put both under the main entry word and as cross-references in their alphabetical place. Cross-references are vital when the user checks an unfamiliar term. For the user’s convenience, cross-references are also provided with Russian equivalents. However, as repetitions may bother users who are learning terms, it is possible to switch cross-references off in the electronic version of the dictionary.

Homonyms, i.e. words or word combinations referring to two or more concepts, are provided with Roman indexes and disambiguation notes. In the example below (see Figure 3), *kuivuminen I* refers to *dryness* (as in *skin dryness*) and *kuivuminen II* to *dehydration* (as in *dehydration of the body*).

<b>Osa 1 -- Sairaudet, oireet ja vaivat -- Vaikeusaste A</b>	
kuivuminen I	сухость (ж.) (кожи, слизистых и т.п.)
kuivuminen II; dehydraatio; nestehukka	обезвоживание (организма); дегидратация; дегидратация

Figure 3: Treatment of Homonyms.

<sup>6</sup> [https://sampo.thl.fi/pivot/prod/fi/avo/perus06/summary\\_icd1001](https://sampo.thl.fi/pivot/prod/fi/avo/perus06/summary_icd1001) [09/05/2020]



The concept-oriented approach prevents one of the “deadly sins” of bilingual lexicography (cf. Kromann, Riiber, & Rosbach 1991: 2724), when multiple meanings are presented in the same entry without proper disambiguation. This results in a long line of translation equivalents, some of which are not interchangeable as they refer to different concepts. Besides, the lack of disambiguation notes may slow down the process of choosing the correct equivalent and result in translation mistakes. The primary need of any translator or interpreter using a dictionary is to locate a correct translation equivalent as quickly as possible and to be sure that this is the right choice (Varantola 1998: 181; Griniov-Grinevitš 2009: 68; Nkwenti-Azeh 2001: 604–606).

## 7 Microstructure

Entries in the dictionary consist of the following data fields: main Finnish term, its possible synonyms, the grammar and usage labels related to them, Russian equivalent, its possible synonyms, and the grammar and usage labels related to them. The order of synonyms within the entry was determined with the help of statistical analysis and domain experts.

Assuming that most community interpreters have a good command of the foreign language, we have provided only a limited amount of phonetic and grammatical information. However, in some cases we also had to take into consideration the needs of the second target group, students.

All Russian equivalents are provided with stresses marked in bold. Terms in the plural form and abbreviations are equipped with the corresponding labels. Informal, colloquial forms are marked with a usage label (see Figure 4). For example, *HDL cholesterol* is often informally called *good cholesterol*. Both terms are included in the dictionary, but the latter one is marked as colloquial.

HDL-kolesteroli; hyvä kolesteroli ( <i>ark.</i> )	липопротеиды высокой плотности <u>мн.ч.</u> ; ЛПВП <u>сокр.</u> ; хороший холестерин ( <u>разг.</u> )
---	---

Figure 4: Examples of Grammatical and Usage Labels.

Irregular plural forms are given in italics in round parentheses to warn users about the unusual inflection. Terms which are typically used in the plural are provided with a note (see Figure 5).

pohje ( <i>mon. pohkeet</i> )	икра ( <i>чаще мн.ч. - икры</i> )
-------------------------------	-----------------------------------

Figure 5: Examples of Irregular Grammatical Forms and a Note on Predominant Usage in the Plural.

Russian equivalents ending in the soft sign (‘ь’) are provided with gender labels, as their gender is not obvious (see Figure 6).

käheys; äänen käheys	осиплость ( <i>ж.</i> ); хрипота
köhä → yskä	кашель ( <i>м.</i> )

Figure 6: Examples of the Gender Label.

Partial equivalents are marked with the ≈ sign (see Figure 7). For example, the Finnish concept *luontaistuote* (≈ natural product/health food) differs from the Russian concept *биодобавка* (≈ dietary supplement), although in many cases they can be used as contextual equivalents.

luontaistuote	≈ биодобавка; БАД <u>сокр.</u> ; биологически активная добавка
---------------	--

Figure 7: Treatment of Partial Equivalents.

Ideally, the differences between the concepts should also be commented on, but such comments would at the same time prevent the dictionary from being a compact reference work. This is an example of the inevitable inner contradictions of a dictionary that has multiple functions and/or is aimed at multiple target groups. In such cases, dictionary compilers must either prioritise or make compromises.

## 8 Challenges Encountered in the Project

Among the main challenges of any bilingual dictionary project are of course culture-specific terms, as national concept systems and terminology differ even in such an international domain as medicine. In particular, terms related to the organisation of healthcare in different countries may cause problems. However, to understand the nuances of translating culture-specific terms, one must have a good command of the languages used in the dictionary, in our case Finnish and Russian. As there is no point in discussing such issues in an article written in English, we will skip this topic and focus on two less language-specific challenges: the opposition between scientific and informal terms, and the treatment of synonyms and abbreviations.

### 8.1 Scientific vs Informal Terms

Medical discourse, like many other languages for special purposes, is multi-dimensional. In particular, it can be categorised according to the participants. One can distinguish communication between doctors, doctors and nurses, doctors and



patients, etc. (cf. Bergenholtz & Tarp 1995: 19; Alexeeva & Mishlanova 2002: 104–105; Gotti 2018: 13–14). Doctor–patient communication differs from other communicative situations in many respects. Typically, patients (and community interpreters) lack a medical background, which means that they tend to communicate using everyday vocabulary rather than scientific medical terminology. This presents a major challenge for dictionary compilers, who have to balance between the general and professional dimensions.

For example, the Russian word *гипертония* (arterial hypertension) used to be an official medical term. It is generally understood and very frequent in everyday discourse. However, it has become obsolete in scientific parlance and has been substituted by the term *артериальная гипертензия*, which is much less familiar to patients. A similar case is the pair of terms *аденома простаты* and *гиперплазия предстательной железы* (prostatic hyperplasia). As our dictionary describes communication between doctors and patients, we have decided to include informal yet widely used terms alongside official ones. However, official terms precede the informal ones. It is not obvious how such informal terms should be labelled. From the doctor’s point of view, they are obsolete, but from the patient’s perspective they are not. We have decided to mark them as colloquial, as they are a part of the patients’ vocabulary which has been “determinologosated”.

## 8.2 Treatment of Synonyms and Abbreviations

An abundance of synonyms and abbreviations is typical of medical language (e.g. Kuryshko 2001: 23–24, 102). This presents a number of challenges related to the selection of synonyms and their placement in a particular order in the entry. We have used the following criteria when solving these issues:

- understandability (patients and interpreters should understand as many terms as possible)
- frequency (frequently used terms are typically more comprehensible and easier to remember)
- simplicity and shortness (simple and short terms are easier to remember)
- diversity (to serve the reference function, the dictionary should also contain terms lying outside the interpreter’s active word stock).

However, these criteria may contradict each other. For example, abbreviated forms are shorter and in principle should be easier to remember. In some cases, this works. For instance, the abbreviation *УЗИ* (medical ultrasound) is much more frequent in Russian than its full form, *ультразвуковое исследование*. In addition, everybody knows this abbreviation. Placing the abbreviation before the full form is therefore quite justified. However, some abbreviations are less well known to patients. For example, the abbreviation *НПВС* (nonsteroidal anti-inflammatory drugs, NSAIDs) is quite rare, despite the fact that NSAIDs themselves are probably familiar to everyone. In this case, the full form should be placed first. These examples demonstrate that prioritisation of the general guiding principles is different in each individual case. To verify their decisions, dictionary makers should consult domain experts and representatives of the target groups.

## 9 Technical Implementation

The dictionary will be published both as a database accessible over the web and as a printable dictionary in pdf format. The electronic version enables efficient searches as well as some useful dynamic features. For example, it is possible to switch some data (e.g. cross-references, administrative data, and domain labels) on and off depending on the intended use. The printable version, in turn, can be used even in situations where the use of electronic devices is not allowed or the Internet connection is poor.

The dictionary has been compiled using a tailored version of the in-house dictionary writing system MyTerMS (see Kudashev & Kudasheva 2006). MyTerMS is a web interface to the underlying lexicographic database. MyTerMS has been used in several dictionary projects and serves as a terminology management system for terminological projects at Tampere University and the University of Helsinki, Finland.

MyTerMS performs all basic operations that can be expected from dictionary management software, such as adding, editing, searching, browsing, printing, and deleting entries. It also automates many operations. For example, while adding a number of entries belonging to the same domain, it is possible to prefill the domain field instead of selecting it manually every time. MyTerMS also helps ensure the integrity of the data, for example, by preventing duplicate entries and automatically managing cross-references. In addition, it ensures the correctness of the input by performing a compliance check before saving the data. MyTerMS also automatically generates the Russian-Finnish index.

The entries are segmented into data fields, and the articles are formed “on the fly” with the help of scripts and cascading style sheets. The layout of the entries is as close to the final as possible except for the presence of some administrative data, which is visually separated from the final data with colour. Administrative data as well as cross-references can be switched on and off by ticking the corresponding checkboxes.

One of the strengths of MyTerMS is its advanced search. Users can perform even very complex searches by using wildcards and regular expressions and combining multiple search conditions. This helps extract data almost by any criteria or their combination. Figure 8 demonstrates the main window of the programme. The green frame is the control panel with various control buttons and options. The two lists on the left are the termlist and the hitlist. The largest frame is the entry frame. Letter rangers in the top left corner allow the length of the termlist to be limited for faster refreshing. In the same corner, there is also a ‘quick search’ pane facilitating navigation through the termlist and the hitlist. The user only needs to type the initial letters of the term for which they are searching.



Figure 8: The Main Window of the MyTerMS Dictionary Writing System.

Entries are added and edited with the help of an HTML form (see Figure 9). The form allows up to seven synonyms to be added for each language. Additional sections for the synonyms open on demand. The programme allows adding inline formatting (e.g. bolded font, italics, upper and lower indexes). However, plain text copies of the corresponding fields are also saved to ensure correct and fast searches. Most labels related to grammar and usage are predefined, but users can also provide additional free-form notes related to these categories.

Figure 9: HTML Form for Adding and Editing Entries.



## 10 Conclusion

The compilation of a thematic dictionary for healthcare interpreters turned out to be a very interesting yet challenging task. Some information needs and those related to data retrieval differ even among the main target groups of the dictionary and in different situations when using the dictionary. The combination of the pedagogical and the reference functions creates some tension, too. Additional target groups (such as patients, medical doctors with immigrant background, etc.) would have aggravated the situation further. In the current project, lexicographic contradictions were also complicated by external factors, such as the tight timetable and limited budget. However, under the circumstances, we are satisfied with the first version of the dictionary. At the same time, we plan to develop the dictionary further in several directions.

While the number of terms that should be learned by heart (about 2,000) approaches the optimal level, the disease-specific part of the dictionary is undeniably modest. Our next goal is to cover about 100 of the most common diseases, which corresponds to approximately 10,000–12,000 terms. The diseases will be selected on the basis of official healthcare statistics. According to the Pareto principle (also known as the 80/20 rule), we expect to reach a reasonable saturation point when the dictionary covers the top 80% of reasons for visiting a doctor. A potential problem here is that the classes represented in the statistics are often “too big” and include multiple undifferentiated diseases.

During the healthcare interpretation course organised at Tampere University, we noticed that the students’ knowledge of medical terms, most of which are nouns, is insufficient. To use the terms correctly in the context, students also need to master verbs and collocations. We plan to enrich our dictionary with usage examples. In the electronic version, these could be switched on and off depending on the intended use of the dictionary.

Our dictionary is already dynamic and customisable to some extent. However, the degree of customisation can be further increased. Our ultimate goal is a multipurpose medical dictionary aimed at multiple target groups, in which the contents could be customised according to the target group and intended usage. However, this requires a more profound study of the users’ perspectives and a great deal of lexicographic and software engineering.

## 11 References

- Alexeeva, L.M. & Mišlanova, S.L. (2002). *Medicinskij diskurs: teoretičeskie osnovy i principy analiza*. Perm’: Izdatel’stvo Permskogo universiteta. [= Medical Discourse: Theoretical Basis and Principles of Analysis].
- Bergenholtz, H. & Tarp, S. (1995). *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: Benjamins.
- Denisov, I. N. & Ševtšenko, Ju. L. (eds.) (2010). *Spravočnik-putevoditel praktikujuščego vratša. 2000 boleznei ot A do Ja*. 2-je izd. Moscow: GEOTAR-Media. [= The Doctor’s Handbook: “2000 Diseases from A to Z”].
- Denisov, I.N., Ševtšenko, Ju.L., Kulakov, V.I. & Haitov, R.M. (eds.) (2002). *Kliničeskie rekomendatsii dlja praktikujuščih vratšei, osnovannyje na dokazatelnoi meditsine*. Moscow: GEOTAR-MED. [= Evidence-Based Medicine Guidelines].
- Flores, G., Abreu, M., Barone, C.P., Bachur, R. & Lin, H. (2012). Errors of Medical Interpretation and their Potential Clinical Consequences: A Comparison of Professional versus Ad Hoc versus no Interpreters. In *Annals of Emergency Medicine*, 60(5), pp. 545–553.
- Fuertes-Olivera, P.A. & Arribas-Baño, A. (2008). *Pedagogical Specialised Lexicography: The Representation of Meaning in English and Spanish Business Dictionaries*. Amsterdam/Philadelphia: Benjamins.
- Gagarina, Je.Ju. (2015). Freim “on-line-konsultacija” v meditsinskih Internet-forumah. In *Vestnik MGOU. Serija: Lingvistika*. № 4, pp. 39–43. [= “Online Consultation” Frame on the Medical Internet Forums].
- Gotti, M. (2018). LSP as Specialised Genres. In J. Humbley, G. Budin & Ch. Laurén (eds.) *Language for Special Purposes: An International Handbook*. Berlin: De Gruyter, pp. 3–25.
- Grin’ov-Grinevič, S.V. (2009). *Vvedenije v terminografiju: Kak prosto i legko sostavit slovar. Utšebnoje posobie*. Izd. 3-je, dop. Moscow: LIBROKOM. [= Introduction to Terminography].
- Hale, S.B. (2007). *Community Interpreting*. Basingstoke: Palgrave Macmillan.
- Käypä hoito -suositukset potilaalle. Accessed at: <https://www.kaypahoito.fi/potilaalle> [09/05/2020]. [= Evidence-based Guidelines for Patients].
- Kolesnikov, L.L. (ed.) (2003). *Terminologia Anatomica* = Meždunarodnaja anatomičeskaja terminologija (S ofits. spiskom rus. ekvivalentov). RANK Ros. anat. nomenklatur. komis. Minzdrava RF, Moscow: Meditsina.
- Koskinen, K., Vuori, J. & Leminen, A.-K. (2018). Johdanto. In K. Koskinen, J. Vuori & A.-K. Leminen (eds.) *Asioimistulkkaus. Monikielisen yhteiskunnan arkea*. Tampere: Vastapaino, pp. 7–28. [= Community Interpreting].
- Kromann, H.-P., Riiber, T. & Rosbach, P. (1991). Principles of Bilingual Lexicography. In F.J. Hausmann et al. (eds.) *Wörterbücher: ein internationales Handbuch zur Lexikographie: an international encyclopedia of lexicography: encyclopédie internationale de lexicographie*. Teilbd 3. Berlin: de Gruyter, pp. 2711–2728.
- Kudashev, I. & Kudasheva, I. (2006) Software Demo: The Terminographic Processor MyTerMS. In G.-M. de Schryver (ed.) DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems, Turin, Italy, 5 September 2006 (Pre-EURALEX 2006). Pretoria: (SF)<sup>2</sup> Press, pp. 35–40.
- Kudashev, I. (2007). *Proektirovanie perevodčeskikh slovaroj special’noj leksiki*. Helsinki: Yliopistopaino. [= Designing LSP Dictionaries for Translators].
- Kuryshko, G.F. (2001). *Javlenije sinonimii v professionalnoi leksike (na materiale nemetskoi i russkoi meditsinskoi terminologii)*. Moscow: Narodnyi utšitel. [= Synonymy in Professional Vocabularies (Case: Russian and German Medical Terminology)].
- Lehtonen, J., Lehtovirta, J. & Mäkelä-Bengs, P. (2013). *THL-toimenpideluokitukset*. Terveystieteiden ja hyvinvoinnin laitos. [= Finnish Classification of Surgical Procedures].



- Madžajeva, S.I. (2012). Freimovyi podhod k sistematizatsii terminologitšeskih znani. In *Utšenyje zapiski: elektronnyi nautšnyi žurnal Kurskogo gosudarstvennogo universiteta*, 2(22), pp. 130-137. [= Using Frame Analysis for Systematisation of Terminological Data].
- Maley, J.H. (2018). Hemoptysis or Hematemesis? – The Importance of Professional Medical Interpretation: A Teachable Moment. In *JAMA Internal Medicine*, 178(6), pp. 841-842.
- Mäntynen, A. (2013). Asioimistulkkaus maahanmuuttajien terveystalveluyksikössä Tampereella. In K. Koskinen (ed.) *Tulkattu Tampere*. Tampere: Tampere University Press, pp. 106-125. [= Community Interpreting at the Healthcare Centre for Immigrants in Tampere].
- Nkwenti-Azeh, B. (2001). User-Specific Terminological Data Retrieval. In S.E. Wright & G. Budin (eds.) *Handbook of Terminology Management*. Vol. 2. Application-Oriented Terminology Management. Amsterdam/Philadelphia: Benjamins, pp.600-612.
- Ollila, S. (ed.) (2017). *Tulkkaus terveydenhuollossa: Lähtökohtana asiakkaan ymmärrys*. Vaasa: University of Vaasa. [= Interpreting in Healthcare Services].
- Roat, C. E. & Crezee, I. (2015). Healthcare interpreting. In H. Mikkelsen & R. Jourdenais (eds.) *The Routledge Handbook of Interpreting*. London: Routledge, pp. 236-253.
- Terveyskirjasto*. Accessed at: <https://www.terveyskirjasto.fi/terveyskirjasto/tk.koti> [09/05/2020]. [= Descriptions of Diseases].
- Valero-Garcés, C. (2008). Hospital Interpreting Practice in the Classroom and the Workplace. In C. Valero-Garcés & A. Martin (eds) *Crossing Borders in Community Interpreting: Definitions and Dilemmas*. Amsterdam/Philadelphia: Benjamins, pp. 165-185.
- Varantola, K. (1998). Translators and their Use of Dictionaries. In B.T.S. Atkins (ed.) *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Niemeyer, pp. 179-192.
- Veisbergs, A. (2006). Dictionaries and Interpreters. In E. Corino, C. Marello & C. Onesti (eds.) *Proceedings of the XII EURALEX International Congress 2006*, Turin, 6–9 September 2006. Turin: Turin University, pp. 1219-1224.

### Acknowledgements

We wish to express our gratitude to the Finnish *Cultura* Foundation (<https://culturas.fi>) for funding the project “Developing Healthcare Interpreting Training”. We also gratefully acknowledge the contribution of the Director of the Pirkanmaa Interpreter Centre (Tampere, Finland) Seija Koskinen, community interpreter Girta Roots, medical doctor Galina Mäkäraäinen, and University Instructor Miia Santalahti (Tampere University).









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicography and Semantic Theory**







# Τοπωνύμια τηςελληνικής και η σχέση τους με τη νεοελληνική γλωσσική εικόνα του κόσμου

O.B. Bobrova

*Lomonosov Moscow State University, Moscow, Russia*

## Abstract

Από ψυχολinguιστικής άποψης ο γλωσσικός πολιτισμός κάθε λαού αποτελεί ένα σύνολο γνώσεων, ιδεών, συνειρμών (associations) και υποδηλώσεων (connotations), αντιλήψεων και ομαδικών αναμνήσεων που βρίσκουν την έκφρασή τους σε αντίστοιχα γλωσσικά μέσα. Σπουδαίο μέρος της γλωσσικής εικόνας του κόσμου αποτελούν σίγουρα τα χωρικά στοιχεία της (περιοχές, χώρες, πόλεις κ.λπ.) που διαμορφώνουν την “πολιτιστική γεωγραφία” κάθε γλώσσας και έχουν γίνει ήδη αντικείμενο ψυχολinguιστικών μελετών.

Το παρόν εξετάζει τα χωρικά στοιχεία του νεοελληνικού γλωσσικού πολιτισμού τα οποία είναι συνδεδεμένα με όλων των ειδών τόπους και τοποθεσίες: πόλεις (Σπάρτη, Μέκκα), χώρες (Αμερική, Βαβυλωνία), όρη (Γολγοθάς, Όλυμπος), ποταμούς (Μαίανδρος, Αήθη), λίμνες (Πρέσπα) κ.ά. (Θερμοπύλες, Βασίλη). Κατά τη σημασιολογική ανάλυση των τοπωνυμίων αναλύονται οι ιδιαιτερότητες της μεταφορικής χρήσης τους, καθώς και ο ρόλος τους στο νοητικό λεξικό του φυσικών ομιλητών της νέας ελληνικής. Τα αποτελέσματα της ανάλυσης μπορούν να χρησιμεύσουν στην επικαιροποίηση και εμπλουτισμό των υπαρχόντων λεξικών της νέας ελληνικής, καθώς και στη συλλογή και περιγραφή στοιχείων που αποτελούν μέρος του “πολιτιστικού λεξικού” της.

**Keywords:** τοπωνύμια, γλωσσική εικόνα του κόσμου, μετωνυμία, στερεότυπη μεταφορά, ψυχολinguιστολογία

## 1 Εισαγωγή

Είναι ευρέως γνωστό πως οι γεωγραφικές γνώσεις αποτελούν αναπόσπαστο και σημαντικό μέρος των επιστημών που έχουν ως αντικείμενο τους τον άνθρωπο και τις ανθρώπινες κοινωνίες. Η γεωγραφία αναγνωρίζεται ως κλειδί όχι μόνο στην ιστορία μιας χώρας ή ενός λαού, αλλά και του κόσμου που ως ολότητα σε πολλούς πολιτισμούς εκλαμβάνεται ως “το σπίτι της ανθρωπότητας” (Boas 1940: 642). Το ενδιαφέρον προς “το σπίτι”, δηλ. στον τόπο διαμονής με την ευρεία έννοια του όρου, ήταν πάντα έκδηλο στον άνθρωπο και εκφραζόταν συνήθως στην επιδίωξη του να γνωρίσει όσο το δυνατόν καλύτερα τη γεωγραφία του τόπου, της χώρας και του πλανήτη του, καθώς και να περιγράψει και να “οργανώσει” με κατάλληλο τρόπο τον κόσμο γύρω του δίνοντας ονόματα σε τόπους και τοποθεσίες.

Για τους λαούς της Ευρώπης ιδιαίτερος ρόλος που ανατίθετο στα γεωγραφικά ονόματα, διαφάνηκε το Β' ήμισυ του XIX αιώνα. Κατά την περίοδο αυτή τα γεωγραφικά ονόματα, όπως και η γλώσσα του έθνους αυτή καθ' εαυτήν, χρησίμευαν ως μέσο αυτοπροσδιορισμού, αναγνώρισης “φίλου” και “εχθρού” (Woodman 2012: 274).

Ως εκ τούτου, τα τοπωνύμια κάθε εθνικής γλώσσας προκαλούν ενδιαφέρον από μόνα τους, αφού αντικατοπτρίζουν την ιστορική ανάπτυξη του λαού που τη μιλά, τις μετακινήσεις του στο βάθος της ιστορίας, τις αλλαγές στην κοινωνική οργάνωσή του κ.ά.: “Κάθε όνομα που δίνεται σε ένα τόπο, έχει μια ιστορία πίσω του. Το όνομα δόθηκε από κάποιον, δόθηκε κάποτε και για μια ορισμένη αιτία” (Tent 2015: 67). Αυτή η υπόσταση των τοπωνυμίων είναι αντικείμενο της ονοματολογίας (για αρχαία και νεοελληνικά τοπωνύμια βλ. ενδεικτικά Matthews 2007; Fraser, Matthews 2010; Συμειωνίδης 2010 κ.ά.).

Ταυτόχρονα ο ρόλος που παίζουν τα τοπωνύμια στον αυτοπροσδιορισμό και την εξέλιξη ενός λαού, επιτρέπει να καταλήξει κανείς στο συμπέρασμα πως τα ονόματα τόπων μπορούν να προσκομίσουν ενδιαφέρον και πλούσιο υλικό όσον αφορά την “πολιτιστική γεωγραφία” μιας φυλής ή εθνικής οντότητας. Αυτό οφείλεται στο γεγονός ότι το νοητικό λεξικό κάθε γλώσσας συμπεριλαμβάνει όχι μόνο τις έννοιες που σχετίζονται με τόπους εν γένει (Zinken 2008: 59), αλλά και ιδέες, υποδηλώσεις και ομαδικές αναμνήσεις που προήλθαν από αυτούς (Krasnykh 2003: 172). Όλα αυτά τα στοιχεία αποτελούν “μακρινή περιφέρεια” της σημασίας αρκετών γεωγραφικών ονομάτων. Τα στοιχεία αυτά διατηρούνται στο λεξιλόγιο και το “γλωσσικό πολιτισμό” του λαού και μεταδίδονται μέσω της γλώσσας στις επόμενες γενιές ομιλητών της, καθορίζοντας ως ένα βαθμό τη συμπεριφορά και την κοσμοαντίληψή τους (Boas 1940: 259).

Εκτός των πραγματικών τόπων και τοποθεσιών ιδιαίτερο ρόλο στον πολιτισμό μπορούν να αποκτήσουν και τόποι φανταστικοί, μυθολογικοί ή παραμυθένιοι. Ενδεικτικό παράδειγμα είναι τα Τάρταρα και τα Ηλύσια πεδία της αρχαίας ελληνικής μυθολογίας ή οι λαϊκές αντιλήψεις για μακρινές ή/και παραμυθένιες χώρες στην ευρωπαϊκή μεσαιωνική παράδοση (*Shlaraffenland/The Land of Cockaigne* ‘γη της αμέριμνης ζωής’, ‘γη όπου ρέει μέλι και γάλα’). Το ίδιο σημαντικά είναι και πραγματικά μέρη και τοποθεσίες που χάρη στο ρόλο τους στην ιστορία και στη σημασία τους για την πολιτιστική παράδοση του συγκεκριμένου λαού έχουν αποκτήσει ένα σύνολο συγκεκριμένων και σταθερών υποδηλώσεων και συνειρμών. Για παράδειγμα, η Βαβυλωνία, ως γνωστόν, ήταν ένα μεγάλο και πλούσιο κράτος του αρχαίου κόσμου με πρωτεύουσα τη Βαβυλώνα. Στη νέα ελληνική όμως το όνομά της χρησιμοποιείται μεταφορικά ως συνώνυμο σύγχυσης και ασυνεννοησίας. Αυτή η μεταφορική χρήση της λέξης ανάγεται στην βιβλική ιστορία για την κατασκευή του Πύργου της Βαβέλ, αλλά, όπως φαίνεται, έλαβε νέα ώθηση χάρη στην περίφημη κωμωδία του Δ. Βυζαντίου.

Η λεπτομερής και πλήρης ανάλυση τοπωνυμίων που έχουν αποκτήσει ιδιαίτερη “πολιτιστική σημασία” στη νέα ελληνική θα μπορούσε, πρώτον, να συμβάλει σημαντικά στην επικαιροποίηση και εμπλουτισμό των υπαρχόντων λεξικών της.



Αυτή η ανάλυση, δεύτερον, θα μπορούσε να σταθεί ένα σημαντικό βήμα στη μελέτη και περιγραφή του “γλωσσικού πολιτισμού” της ελληνικής γλώσσας. Τρίτον, η ανάλυση αυτή είναι δυνατόν να θεωρηθεί προκαταρκτική φάση της σύνταξης ενός “πολιτιστικού λεξικού” της νέας ελληνικής. Μέχρι στιγμής έχουν γίνει πολλές και επιτυχημένες προσπάθειες εντοπισμού και περιγραφής στοιχείων του “πολιτιστικού λεξικού” μερικών ευρωπαϊκών γλωσσών (βλ. ενδεικτικά Wierzbicka 1997; Krasnykh 2016). Η σύνταξη ενός παρόμοιου λεξικού με νεοελληνικά δεδομένα θα αποτελούσε ένα “γνώθι σαυτόν” για τους ομιλητές της νέας ελληνικής, καθώς και ένα χρήσιμο εγχειρίδιο για αυτούς που τη μαθαίνουν ως δεύτερη/ξένη.

## 2 Υλικό και Στόχοι της Ανάλυσης

Είναι ευρέως αποδεκτή η άποψη πως το λεξιλόγιο της γλώσσας αντικατοπτρίζει πλήρως το περιβάλλον και τις κοινωνικές σχέσεις των φυσικών ομιλητών της, καθώς και αποτελεί μέσω έκφρασης και μετάδοσης ιδεών, ενδιαφερόντων και ασχολιών τους (Sapir 1973: 90-91). Οι αλλαγές στη σημασία των λέξεων, η εξαφάνιση παλιών λέξεων και εμφάνιση καινούριων σχετίζονται αδιάρρηκτα με τις αλλαγές που γίνονται στον πολιτισμό και τη συνείδηση του λαού (Ibid.: 27). Το παρόν αναλύει το ρόλο που παίζουν στη νέα ελληνική τα τοπωνύμια που, χάρη σε πρόσθετες μεταφορικές σημασίες που έχουν αποκτήσει με την πάροδο του χρόνου, μπορούν να θεωρούνται λέξεις “με ειδική πολιτιστική σημασία” (Wierzbicka 1997: 5). Εξετάζεται μια ομάδα λεξημάτων αποτελούμενη από 47 τοπωνύμια (ονόματα χωρών, περιοχών και πόλεων, συμπεριλαμβανομένων και φανταστικών, ποταμών, ορών, λιμνών κ.ά.). Η πλήρης λίστα παρουσιάζεται στον πίνακα 1 (βλ. παρακάτω).

Αναφορά ( <i>reference</i> ) του λεξήματος	Λέξη/μα	Αριθμός (ποσοστό %)
Πόλη, οικισμός	Αθήνα      Άουσβιτς Βαβυλώνα      Βατερλώ Γόμορρα      Γοργοπόταμος Ιερουσαλήμ      Καλαμάτα Καρδαμύλη      Κορώνη Μαραθώνας      Μάτι Μέκκα      Μυτιλήνη Νυρεμβέργη      Πόλη Πομπηία      Σόδομα      Σπάρτη Σύβαρη	20 (42,55)
Φυσικά γεωγραφικά φαινόμενα		14 (29,79)
Όρος	Γολγοθάς      Έβερεστ Όλυμπος      Παρνασσός Άγιο(ν) Όρος*	5 (10,6)
Νησί, ήπειρος	Ατλαντίδα      Ίμια      Λέσβος	3 (6,38)
Ποταμός	Λήθη, Μαιάνδρος	2 (4,25)
Γεωγραφική περιοχή	Λακωνία, Μάνη	2 (4,25)
Πέρασμα	Θερμοπύλες	1 (2,13)
Κόλπος	Ναβαρίνο	1 (2,13)
Λίμνη	Πρέσπες	1 (2,13)
Κράτος, χώρα	Αγγλία/Μεγάλη Βρετανία      Αίγυπτος Αμερική      Βαβυλωνία Γερμανία      Ελβετία Ελντοράντο	7 (14,9)
Φρούριο, μοναστήρι	Αγία Λαύρα      Βαστίλη	2 (4,25)
Τόπος, περιοχή με άγνωστη τοποθεσία	Γέεννα      Παράδεισος	2 (4,25)
Άλλα: ξενοδοχείο	Γουότεργκειτ	1 (2,13)
	Σύνολο:	47 (100)

Πίνακας 1: Λίστα των προς ανάλυση τοπωνυμίων και οι αναφορές (*references*) τους.

\*Το όνομα Άγιον Όρος αναφέρεται κυρίως στο βουνό αλλά και στη χερσόνησο όπου βρίσκεται.

Τα προς ανάλυση λεξήματα και τα περικείμενά τους προέρχονται από εγκυκλοπαιδικά και άλλα λεξικά της κοινής νεοελληνικής γλώσσας (Μπαμπινιώτης 2002, Πάπυρος Larousse 2003, Συμμεωνίδης 2010, Χρηστικό λεξικό της νεοελληνικής γλώσσας 2014), διαθέσιμα νεοελληνικά σώματα κειμένων (Σώμα Ελληνικών Κειμένων<sup>1</sup>, Corpus of Modern Greek<sup>2</sup>, eITenTen<sup>3</sup>, ΕΘΕΓ<sup>4</sup>), ηλεκτρονικές εκδόσεις μεγάλων ελληνικών εφημερίδων (“Καθημερινή”,

<sup>1</sup> <http://www.sek.edu.gr/search.php>

<sup>2</sup> <http://web-corpora.net/GreekCorpus/search/>



“Το Βήμα”, “Εφημερίδα των Συντακτών” κ.ά.), καθώς και ελληνικές ιστοσελίδες στο Διαδίκτυο το οποίο αξιοποιείται ως πηγή παραδειγμάτων χρήσης σύγχρονης ζωντανής γλώσσας (Gatto 2014: 1).

Πρέπει να σημειωθεί πως είναι αρκετές οι περιπτώσεις χρήσης όχι των ίδιων των τοπωνυμίων αλλά των παράγωγών τους. Έτσι, εκτός από τοπωνύμια, τα χωρικά στοιχεία του νεοελληνικού γλωσσικού πολιτισμού εκφράζονται με:

- επίθετα παράγωγα από τοπωνύμιο τα οποία συνδέονται ελεύθερα με ουσιαστικά (σπαρτιάτικη αγωγή, σπαρτιάτικο γεύμα, σπαρτιάτικο φαγητό)·
- πολυλεκτικά λεξήματα τύπου “επίθετο παράγωγο από τοπωνύμιο + ουσιαστικό” (μαραθώνιος δρόμος)·
- ουσιαστικοποιημένα επίθετα παράγωγα από τοπωνύμιο (μέγκλα, καλαματιανός).

Το συμπέρασμα ότι το τοπωνύμιο έχει “ιδιαίτερη πολιτιστική σημασία” βασίζεται στην δυνατότητα μεταφορικής ή μετωνυμικής χρήσης του. Εκτός αυτού, ένδειξη της “ιδιαίτερης πολιτιστικής σημασίας” αποτελεί η χρήση του εν λόγω τοπωνυμίου στα ρητά, γνωμικά και παροιμίες της νέας ελληνικής.

Με κριτήριο τη σημασία τους τα παραπάνω τοπωνύμια χωρίζονται στις εξής κατηγορίες:

- τοπωνύμια που χρησιμοποιούνται μετωνυμικά στο λόγο (δηλ. είναι μετωνυμίες τύπου “τόπος αντί γεγονότος”): εκλογικό Βατερλώ, ένα νέο Ναβαρίνο (20 λεξήματα).
- τοπωνύμια που χρησιμοποιούνται μεταφορικά στο λόγο (η Ευρώπη είναι παράδεισος, να γίνει η Ελλάδα Μέκκα καινοτομίας) (14 λεξήματα).<sup>5</sup>
- τοπωνύμια των οποίων τα παράγωγα χρησιμοποιούνται στο λόγο μεταφορικά ή μετωνυμικά (σπαρτιάτικο γεύμα, μαραθώνιες συνομιλίες, συβαρίτης, λακωνίζω) (9 λεξήματα).

Με σκοπό να αποδείξουμε πως ο χαρακτηρισμός “λεξήματα με ιδιαίτερη πολιτιστική σημασία” χρησιμοποιείται πολύ βάσιμα αναφορικά με τα παραπάνω τοπωνύμια, θεωρούμε σκόπιμο να τα αναλύσουμε παρακάτω πιο λεπτομερώς ανά κατηγορίες.

### 3 Τοπωνύμια “με Ιδιαίτερη Πολιτιστική Σημασία” και η Μέθοδος Ανάλυσής τους

Όπως αναφέραμε πιο πάνω, τα τοπωνύμια που αποτελούν το αντικείμενο του παρόντος, έχουν αποκτήσει ένα σύνολο πρόσθετων σημασιολογικών στοιχείων. Τα στοιχεία αυτά προέρχονται κυρίως από γενικές εγκυκλοπαιδικές γνώσεις. Η σωστή μεταφορική χρήση του λεξήματος Βαβυλωνία, επί παραδείγματι, βασίζεται άμεσα στις γνώσεις του ομιλητή για την ιστορία της κατασκευής του Πύργου της Βαβέλ. Ως εκ τούτου, η ανάλυση αυτού και παρόμοιων λεξημάτων με παραδοσιακά σημασιολογικά μέσα θα αποδειχτεί αναποτελεσματική, επειδή τα “πολιτιστικά φορτισμένα” σημασιολογικά στοιχεία τους, οι σημασιολογικές τους σχέσεις (associations) και συνειρμοί (connotations) βρίσκονται στην μακρινή περιφέρεια της σημασίας τους. Η ψυχολογολόγος V. Krasnykh με σκοπό τον εντοπισμό και την περιγραφή αυτών των λεπτομερειών και λεπτών αποχρώσεων της σημασίας προτείνει έναν αλγόριθμο αποτελούμενο από τα εξής βήματα:

- 1) προσδιορισμός της προέλευσης του λεξήματος (παγκόσμια ιστορία, μυθιστορήματα, Ιερά Ιστορία κ.ά.)·
- 2) περιγραφή της κυριολεκτικής σημασίας του λεξήματος, της αναφοράς (reference) του και σχετικές εγκυκλοπαιδικές πληροφορίες·
- 3) περιγραφή μεταφορικής/μετωνυμικής χρήσης του λεξήματος στο λόγο·
- 4) παράθεση παραδειγμάτων χρήσης του λεξήματος και εκφράσεων που το περιλαμβάνουν (Krasnykh 2016: 418).

#### 3.1 Τοπωνύμια-μετωνυμίες “Τόπος αντί Γεγονότος”

Σύμφωνα με τον G. Lakoff, η μετωνυμία αποτελεί αναπόσπαστο μέρος της ανθρώπινης σκέψης. Η μεγάλη διάδοση των μετωνυμιών στη γλώσσα οφείλεται στο γεγονός ότι διευκολύνουν σημαντικά την επικοινωνία, επιτρέποντας στον ομιλητή να συγκεντρώσει την προσοχή του σε μία και μόνο όψη μιας σύνθετης κατάστασης ή μιας αφηρημένης έννοιας κάνοντάς την πιο απλή και προσιτή (well-understood, easy-to-perceive) (Lakoff 1987: 77).

Όπως αναφέραμε παραπάνω, τα περισσότερα τοπωνύμια που εξετάσαμε χρησιμοποιούνται στο λόγο μετωνυμικά δηλώνοντας ένα γεγονός αντί του τόπου στον οποίο συνέβη. Ακολουθεί η ανάλυσή μερικών από αυτά με βάση τον αλγόριθμο που μόλις προαναφέραμε.

##### Θερμοπύλες

- 1) Το λέξημα δηλώνει ένα παραθαλάσσιο πέραςμα στην Ελλάδα. Η σφαίρα προέλευσης του λεξήματος είναι παγκόσμια ιστορία και, συγκεκριμένα, η ιστορία της Αρχαίας Ελλάδας.
- 2) Τα Στενά των Θερμοπύλων είναι γνωστά στην ιστορία από τη μάχη με τους Πέρσες που έγινε εκεί το 480 π.Χ. Εκεί έπεσαν μαχόμενοι ο Λεωνίδας και οι 300 Σπαρτιάτες του. Η θυσία τους έγινε σύμβολο αυταπάρνησης και πατριωτισμού (ΠΛ 2003: 682).
- 3) Χρησιμοποιείται για να δηλώσει αιματηρή, δύσκολη και ηρωική αντίσταση σε κάτι κακό.
- 4) Παραδείγματα χρήσης:

<sup>3</sup> <https://www.sketchengine.eu/eltenten-greek-corpus/>

<sup>4</sup> <http://hnc.ilsp.gr/index.php>

<sup>5</sup> Ως γνωστόν, μέχρι τώρα δεν έχουν τεθεί σαφή κριτήρια διαφοροποίησης μεταφοράς και μετωνυμίας (Bartsch 2003: 49) που παρουσιάζουν πολλές ομοιότητες ως προς τη γνωσιακή υπόσταση και τη συμβολική χρήση τους (Lakoff, Johnson 1980: 37). Στο παρόν οι χρήσεις των λεξημάτων αναγνωρίζονται ως μεταφορικές ή μετωνυμικές με βάση τη σημασιολογία: η μεταφορά έχει να κάνει με μια κατάσταση ως ολότητα (μεταφορά μερικών σημασιολογικών στοιχείων από μια οντότητα σε μια άλλη) ενώ στην περίπτωση της μετωνυμίας πρόκειται για μεταφορά ενός σημασιολογικού στοιχείου.



- (1) Σε ιδιώτες δίνονται τα διόδια και προβλέπεται αύξηση 125%. Τέτοιο τίμημα δεν πλήρωσαν ούτε οι Πέρσες στις *Θερμοπύλες*<sup>6</sup>  
 (2) τα Ίμια είναι οι σύγχρονες *Θερμοπύλες*.<sup>7</sup>

Εκφράσεις: *φυλάω Θερμοπύλες* ‘προστατεύω και διαφυλάσσω κάτι, κυρ. θεσμό ή αξία, που θεωρείται σημαντικό και πολύτιμο’ (XP 2014: 69).

Παραδείγματα χρήσης:

- (3) Πάντως, η ίδια η Μέρκελ σε ομιλία της στο Ανόβερο τάχθηκε κατά της πολιτικής αύξησης των ελλειμμάτων για την τόνωση της ανάπτυξης, διαμηνύοντας *urbi et orbi* ότι εξακολουθεί και ότι θα εξακολουθήσει μέχρι τέλους να *φυλά τις Θερμοπύλες* της δημοσιονομικής ορθοδοξίας.<sup>8</sup>  
 (4) Ο Έβρος και η Θράκη εδώ και χρόνια *φυλούν Θερμοπύλες*.<sup>9</sup>

#### Πομπηία

- 1) Είναι όνομα ιταλικής πόλης. Η σφαίρα προέλευσης του λεξήματος είναι παγκόσμια ιστορία.  
 2) Η πόλη της Πομπηίας θάφτηκε από στρώμα τέφρας το 79 μ.Χ. κατά την έκρηξη του Βεζουβίου (ΠΛ 2008: 1443).  
 3) Χρησιμοποιείται στο λόγο για να περιγράψει την κατάσταση αναμονής μιας καταστροφής ή (συχνότερα) την κατάσταση πανωλεθρίας και μαζικών θανάτων.  
 4) Παραδείγματα χρήσης:  
 (5) όλη η Γης σήμερα είναι μια *Πομπηία*, λίγη ώρα πριν από την έκρηξη (Ν. Καζαντζάκης, Αδελφοφάδες)<sup>10</sup>  
 (6) σύγχρονη *Πομπηία*! Ιστορίες που συγκλονίζουν [...] Οι λέξεις είναι φτωχές για να περιγράψουν την τραγωδία από την καταστροφική φωτιά στο Μάτι και στο Νέο Βουτζά της Αττικής<sup>11</sup>  
 (7) Σύγχρονη *Πομπηία* η Ανατολική Αττική.<sup>12</sup>

Εκφράσεις: δεν υπάρχουν.

Η μετωνυμία “τόπος αντί γεγονότος” είναι πολύ διαδεδομένη στη γλώσσα (Kövesces 2002: 144-145). Πιθανότατα σε αυτό οφείλεται και το ότι η ελληνική γλώσσα όχι μόνο έχει μια αξιοσημείωτη λίστα μετωνυμιών “με πολιτιστική σημασία” (βλ. πίνακα 1) αλλά και εμπλουτίζεται συνεχώς με τέτοιου είδους μετωνυμίες. Για παράδειγμα, κατά τη διάρκεια της ανάλυσής μας, εκτός από πολύ διαδεδομένες και καταχωρισμένες σε λεξικά μετωνυμίες, εντοπίσαμε λεξήματα των οποίων η μετωνυμική χρήση σημειώνεται σχετικά πρόσφατα:

#### Μάτι

- 1) είναι οικισμός του ν. Αττικής.  
 2) Το καλοκαίρι του 2018 εκεί σημειώθηκαν φονικές πυρκαγιές που στοίχισαν τη ζωή πολλών ανθρώπων. Είναι η φονικότερη πυρκαγιά στην ιστορία του νεοελληνικού κράτους και μια από τις φονικότερες στη σύγχρονη ιστορία.  
 3) Χρησιμοποιείται στο λόγο ως χαρακτηρισμός καταστροφικής πυρκαγιάς με πολλά θύματα.  
 4) Παραδείγματα χρήσης:  
 (8) Ηθελαν «Νέο Μάτι;» - Σχέδιο εμπρησμού διερευνά η Πυροσβεστική.<sup>13</sup>  
 (9) Φόβοι για νέο Μάτι: Αυτές είναι οι εννέα περιοχές σε κίνδυνο.<sup>14</sup>

Εκφράσεις: δεν υπάρχουν.

<sup>6</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=2000&start\\_sent\\_no=&end\\_sent\\_no=&found1=556.20&query=page=2&sid=93520&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=2000&start_sent_no=&end_sent_no=&found1=556.20&query=page=2&sid=93520&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>7</sup> <https://www.enikos.gr/international/554173/ta-imia-einai-oi-synchrones-thermopyles-ti-leei-o-fantaros-pou-ep> (accessed 20.04.2020).

<sup>8</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=8234&start\\_sent\\_no=&end\\_sent\\_no=&found1=198.38&query=page=2&sid=93520&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=8234&start_sent_no=&end_sent_no=&found1=198.38&query=page=2&sid=93520&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>9</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=2523&start\\_sent\\_no=&end\\_sent\\_no=&found1=303.10&query=page=2&sid=93520&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=2523&start_sent_no=&end_sent_no=&found1=303.10&query=page=2&sid=93520&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>10</sup> [http://web-corpora.net/GreekCorpus/search/?interface\\_language=ru](http://web-corpora.net/GreekCorpus/search/?interface_language=ru) (accessed 20.04.2020).

<sup>11</sup> <http://patrastimes.gr/home-page-slider/%CF%83%CF%8D%CE%B3%CF%87%CF%81%CE%BF%CE%BD%CE%B7-%CF%80%CE%BF%CE%BC%CF%80%CE%B7%CE%AF%CE%B1-%CE%B9%CF%83%CF%84%CE%BF%CF%81%CE%AF%CE%B5%CF%82-%CF%80%CE%BF%CF%85-%CF%83%CF%85%CE%B3%CE%BA%CE%B%CE%BF/> (accessed 20.04.2020).

<sup>12</sup> <https://empros.gr/2018/07/syngchroni-pompiia-i-anatoliki-attiki/> ().

<sup>13</sup> <https://www.tanea.gr/2019/07/23/greece/ithelan-neo-mati-sxedio-emprismou-diereyna-i-pyrosvestiki/> (accessed 20.04.2020).

<sup>14</sup> <https://www.in.gr/2019/03/09/greece/fovoi-gia-neo-mati-aytes-einai-oi-ennea-perioxes-pou-kindyneyouun/> (accessed 20.04.2020).



*Πρέσπες*

- 1) είναι γνωστή λιμναία περιοχή στα σύνορα Ελλάδας, Βόρειας Μακεδονίας και Αλβανίας.
- 2) Στη Λίμνη της Μεγάλης Πρέσπας στις 17 Ιουνίου 2018 υπεγράφη η διακρατική συμφωνία Ελλάδας-ΠΓΔΜ για τη μετονομασία της τελευταίας σε Δημοκρατία της Βόρειας Μακεδονίας.
- 3) Χρησιμοποιείται στο λόγο ως συνώνυμο αμφίβολης συμφωνίας με την υπογραφή της οποίας η θέση της Ελλάδας στη διεθνή πολιτική σκηνή υποβαθμίζεται σημαντικά.
- 4) Παραδείγματα χρήσης:

(10) Προς νέες *Πρέσπες* φαίνεται πως οδηγείται η Ελλάδα με την Τουρκία, με κρυφές συμφωνίες για Αιγαίο και Ν. Μεσόγειο για τα κοιτάσματα αλλά και την ΑΟΖ!<sup>15</sup>

Εκφράσεις: δεν υπάρχουν.

*Νυρεμβέργη*

- 1) είναι πόλη της Γερμανίας.
- 2) Εκεί έλαβε χώρα η δίκη των μελών του Εθνικοσοσιαλιστικού κόμματος και άλλων οργανώσεων της ναζιστικής Γερμανίας για διάπραξη εγκλημάτων κατά της ανθρωπότητας (ΠΛ 2003: 1255).
- 3) Χρησιμοποιείται ως συνώνυμο μιας δίκης που οριστικά βάζει τέλος σε παράνομες πράξεις και εγκλήματα.
- 4) Παραδείγματα χρήσης:

(11) Πότε θα στηθεί μια νέα *Νυρεμβέργη*, όπου ο κάθε Ερντογάν και κάθε ιμπεριαλιστής ηγέτης θα λογοδοτήσουν για τόσα εγκλήματα πολέμου και τόσες παραβιάσεις του διεθνούς δικαίου και των δικαιωμάτων του ανθρώπου;<sup>16</sup>

Εκφράσεις: δεν υπάρχουν.

*Ιμια*

- 1) είναι βραχονησίδες στο ΝΑ Αιγαίο
- 2) Εκεί το 1996 σημειώθηκε μια κρίση με αφορμή την προσάραξη τουρκικού πλοίου. Ακολούθησε η αμφισβήτηση της ελληνικής θαλάσσιας κυριαρχίας εκ μέρους της Τουρκίας. Η κατάσταση παραλίγο να εξελιχτεί σε μια γενικευμένη σύγκρουση αλλά με την παρέμβαση του ΝΑΤΟ οι δύο χώρες απέσυραν τις ναυτικές δυνάμεις τους.
- 3) Χρησιμοποιείται ως συνώνυμο μιας ένοπλης σύγκρουσης (κυρίως με την Τουρκία) με αφορμή την αμφισβήτηση της κυριαρχίας της Ελλάδας.
- 4) Παραδείγματα χρήσης:

(12) Οδεύουμε προς νέα... *Ιμια*;<sup>17</sup>

Εκφράσεις: *κρίση των Ιμίων*.

**3.2 Τοπωνύμια-μεταφορές**

Τα λεξήματα αυτού του τύπου είναι λιγότερα σε σχέση με τα τοπωνύμια-μετωνυμίες και, σε αντίθεση με αυτά, δεν περιγράφουν ένα γεγονός αλλά παρουσιάζουν έναν γενικό χαρακτηρισμό μιας κατάστασης με βάση την υπόρρητη σύγκρισή της με μια άλλη.

*Μέκκα*

- 1) Είναι όνομα μιας πόλης της Σαουδικής Αραβίας.
- 2) Ιερή πόλη των μουσουλμάνων (ΠΛ 2003: 1069). Η μεταφορική χρήση του λεξήματος έχει σχέση με το ισλάμ και τα ιερά του κείμενα στα οποία η Μέκκα αναφέρεται ως γενέτειρα του προφήτη Μωάμεθ.
- 3) Χρησιμοποιείται στο λόγο ως συνώνυμο για τόπο ιερό και λατρεμένο ή για πόλο έλξης ανθρώπων που ασκούν ένα συγκεκριμένο επάγγελμα.
- 4) Παραδείγματα χρήσης:

(13) “*Μέκκα* του Καπιταλισμού” Wall Street<sup>18</sup>

(14) η δράση μεταφέρεται από τη Βενετία στη *Μέκκα* του τζόγου, στο Λας Βέγκας<sup>19</sup>

(15) *Μέκκα* της μόδας.<sup>20</sup>

<sup>15</sup> [https://viralgreece.eu/katrovgkalos-i-toyrkia-echei-dikaiomata-sta-koitasmata-stin-n-mesogeio?fbclid=IwAR1YncOwHK9IJBVGzQIi9vmw-X0ECXfk9qosnwkQOEajf3vl8\\_0rUrB7Q0](https://viralgreece.eu/katrovgkalos-i-toyrkia-echei-dikaiomata-sta-koitasmata-stin-n-mesogeio?fbclid=IwAR1YncOwHK9IJBVGzQIi9vmw-X0ECXfk9qosnwkQOEajf3vl8_0rUrB7Q0) (accessed 16.04.2020).

<sup>16</sup> <https://simerini.sigmalive.com/article/2019/10/20/mia-nea-nuremberge/> (accessed 20.04.2020).

<sup>17</sup> <https://www.philenews.com/koinonia/epistoles/article/848033/odevome-pros-nea-imia> (accessed 16.04.2020).

<sup>18</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=8835&start\\_sent\\_no=&end\\_sent\\_no=&found1=103.7&query=sid=11410&page=1&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=8835&start_sent_no=&end_sent_no=&found1=103.7&query=sid=11410&page=1&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>19</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=445&start\\_sent\\_no=&end\\_sent\\_no=&found1=7.22&query=sid=11410&page=1&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=445&start_sent_no=&end_sent_no=&found1=7.22&query=sid=11410&page=1&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>20</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=8585&start\\_sent\\_no=&end\\_sent\\_no=&found1=119.7&query=page=2&sid=11410&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=8585&start_sent_no=&end_sent_no=&found1=119.7&query=page=2&sid=11410&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).



Εκφράσεις: δεν υπάρχουν.

#### *Εβερρεστ*

- 1) είναι βουνό των Ιμαλαίων.
- 2) Είναι το ψηλότερο βουνό του κόσμου.
- 3) Χρησιμοποιείται μεταφορικά ως χαρακτηρισμός για κάτι τεράστιο, υπέρογο ή κάτι πολύ σπουδαίο και περίφημο.
- 4) Παραδείγματα χρήσης:

(16) Δεν έχει προφανώς αντίρρηση με τη... σπουδή του πολιτικού συστήματος να κατακτήσει το *Εβερρεστ* της διαφάνειας.<sup>21</sup>

(17) «Ο Νονός»: Το *Εβερρεστ* του αμερικανικού σινεμά.<sup>22</sup>

Εκφράσεις: δεν υπάρχουν.

### 3.3 Τοπωνύμια των οποίων τα Παράγωγα Χρησιμοποιούνται στο λόγο Μεταφορικά ή Μετωνυμικά

Σε αρκετές περιπτώσεις το τοπωνύμιο δεν χρησιμοποιείται ως μεταφορά ή μετωνυμία. Μεταφορική χρήση όμως έχουν τα παράγωγά του που συνήθως χρησιμοποιούνται είτε μόνα τους είτε αποτελούν μέρος λεξιλογικών συνάψεων (lexical collocations).

#### *Μαραθώνας*

- 1) είναι πόλη και δήμος στην Αττική, Ελλάδα.
- 2) Είναι γνωστός για τη μάχη που έγινε το 490 π.Χ. μεταξύ των Αθηναίων και των Περσών. Η είδηση για τη νίκη των Αθηναίων υπό τον Μιλτιάδη μαθεύτηκε στην Αθήνα χάρη στον δρομέα Φειδιππίδη που έπεσε νεκρός από την εξάντληση μόλις ανήγγειλε τη νίκη (ΠΑ 2003: 1043).
- 3) Μεταφορική χρήση στο λόγο έχει το παράγωγο λέξημα *μαραθώνιος* (ως επίθετο και ως ουσιαστικό) που δηλώνει επίπονη προσπάθεια ή κάτι που έχει μεγάλη διάρκεια (Ibid.).
- 4) Παραδείγματα χρήσης:

(18) εξεταστικός *μαραθώνιος* των υποψηφίων<sup>23</sup>

(19) *μαραθώνιες* συνομιλίες συνολικής διάρκειας έξι ωρών.<sup>24</sup>

Εκφράσεις: δεν υπάρχουν.

#### *Λακωνία*

- 1) είναι ιστορική περιοχή της Σπάρτης.
- 2) Οι Σπαρτιάτες είναι γνωστοί στην ιστορία ως υπέροχοι πολεμιστές, φημίζονταν επίσης για λιτό και αυστηρό τρόπο ζωής τους.
- 3) Μεταφορική χρήση στο λόγο έχουν τα παράγωγα λεξήματα *λακωνίζω*, *λακωνικός* -ή -ό, *λακωνίζειν* που δηλώνουν πολύ λιτό, περιεκτικό και εύστοχο τρόπο έκφρασης.
- 4) Παραδείγματα χρήσης:

(20) επισημαίνει *λακωνικά* η «Zeit»<sup>25</sup>

(21) Περιορίστηκε απλώς σε μια *λακωνική* παρατήρηση<sup>26</sup>

Εκφράσεις: *λακωνίζειν* εστί φιλοσοφείν 'το να λες λίγα είναι φιλοσοφία':

(22) Το αρχαίο ρητό "Το *λακωνίζειν* εστί φιλοσοφείν" είναι ο χρυσός κανόνας για την ειδησεογραφία.<sup>27</sup>

### 3.4 Τοπωνύμια ως Μέρος του Νοητικού Λεξικού και του "Πολιτιστικού Λεξικού" της Νέας Ελληνικής

Όπως προκύπτει από την ανάλυση, το να αποκτά ένα τοπωνύμιο κάποιες πρόσθετες υποδηλώσεις και μεταφορικές σημασίες είναι πολύ διαδεδομένο στη κοινή νεοελληνική. Μπορούμε όμως με βάση τα αποτελέσματα της παραπάνω

<sup>21</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=8707&start\\_sent\\_no=&end\\_sent\\_no=&found1=454.14&query=page=3&sid=72821&search\\_language=greek&interface\\_language=en&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=8707&start_sent_no=&end_sent_no=&found1=454.14&query=page=3&sid=72821&search_language=greek&interface_language=en&contexts_output_language=greek) (accessed 20.04.2020).

<sup>22</sup> [http://www.cinemamagazine.gr/themata/arthro/godfather\\_bday-130998731/](http://www.cinemamagazine.gr/themata/arthro/godfather_bday-130998731/) (accessed 20.04.2020).

<sup>23</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=535&start\\_sent\\_no=&end\\_sent\\_no=&found1=78.7&query=sid=43110&page=1&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=535&start_sent_no=&end_sent_no=&found1=78.7&query=sid=43110&page=1&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>24</sup> [http://web-corpora.net/GreekCorpus/search/context.php?doc\\_id=9208&start\\_sent\\_no=&end\\_sent\\_no=&found1=759.13&query=sid=55558&page=1&search\\_language=greek&interface\\_language=ru&contexts\\_output\\_language=greek](http://web-corpora.net/GreekCorpus/search/context.php?doc_id=9208&start_sent_no=&end_sent_no=&found1=759.13&query=sid=55558&page=1&search_language=greek&interface_language=ru&contexts_output_language=greek) (accessed 20.04.2020).

<sup>25</sup> <http://hnc.ilsp.gr/index.php#> (accessed 20.04.2020).

<sup>26</sup> Ibid.

<sup>27</sup> Ibid.

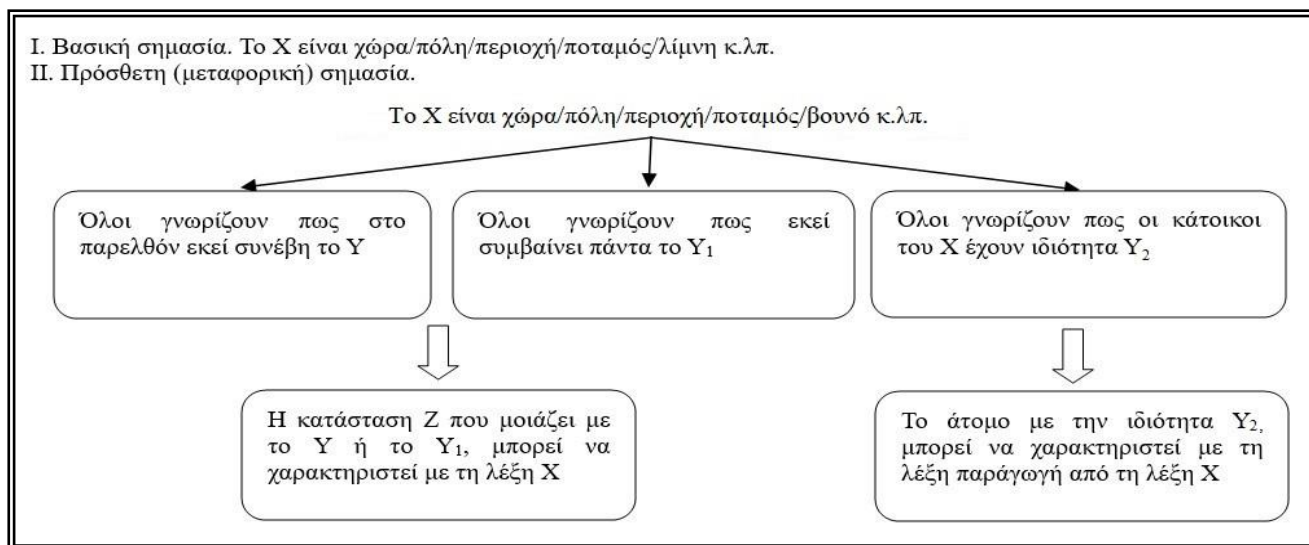


ανάλυσής να βγάλουμε κάποια συμπεράσματα σχετικά με το πως παρουσιάζονται οι σημασίες των τοπωνυμίων στο νοητικό λεξικό των ομιλητών της νέας ελληνικής;

Όπως φαίνεται, η δυνατότητα της μεταφορικής χρήσης των εν λόγω τοπωνυμίων στηρίζεται σχεδόν αποκλειστικά σε εγκυκλοπαιδικές γνώσεις του ομιλητή. Κατά την ανάλυση της σημασίας συγκεκριμένων λεξημάτων, για παράδειγμα, είναι αδύνατον να παραλειφθούν οι αναφορές σε πραγματικά γεγονότα ή/και άλλες γενικές πληροφορίες που εξηγούν την μεταφορική χρήση του λεξήματος. Επομένως, πρόκειται για λεξήματα για τα οποία είναι σχεδόν αδύνατος ο “διαχωρισμός της σημασίας και εγκυκλοπαιδικής προέκτασής της” [Atchinson 2008: 50].

Εκτός αυτού, πρέπει να σημειωθεί πως η μεταφορική χρήση των τοπωνυμίων αποτελεί έναν τρόπο έκφρασης συναισθημάτων στο λόγο, πράγμα που κάνει τα εν λόγω τοπωνύμια “συναισθηματικά φορτισμένα”. Με σκοπό να διευκολυνθεί η σημασιολογική ανάλυσή τους, θεωρήθηκε σκόπιμο να χρησιμοποιηθεί η μέθοδος παραφράσεων που προτάθηκε από την A. Wierzbicka και αποδείχθηκε αποτελεσματική στη σημασιολογική ανάλυση σύνθετων αφηρημένων εννοιών (βλ. Wierzbicka 1997).

Με βάση τα παραπάνω μπορούμε να παρουσιάσουμε τη σημασία των τοπωνυμίων ως ακολούθως (βλ. εικόνα 1):



Εικόνα 1: Σημασία των τοπωνυμίων όπως παρουσιάζονται στο νοητικό λεξικό της νέας ελληνικής.

#### 4 Συμπεράσματα

Η ανάλυση έδειξε πως τα τοπωνύμια της νέας ελληνικής αρκετά συχνά αποκτούν πρόσθετες μετωνυμικές ή μεταφορικές σημασίες. Ο σημασιολογικός τύπος “τόπος αντί γεγονότος” έχει την πιο ευρεία χρήση.

Συνήθως μεταφορικά ή μετωνυμικά χρησιμοποιούνται τα τοπωνύμια που σχετίζονται με γεγονότα της ελληνικής ή παγκόσμιας ιστορίας (το ποσοστό τους ανέρχεται σε περίπου 66%), μυθολογία ( $\approx 15\%$ ), θρησκεία ( $\approx 15\%$ ), γεωγραφία ( $\approx 4\%$ ).

Η μεταφορική χρήση των τοπωνυμίων είναι δυνατή μόνο σε περίπτωση που ο ομιλητής διαθέτει αρκετές γενικές εγκυκλοπαιδικές γνώσεις που του επιτρέπουν να πραγματώσει τη “συμβολική αξία” του λεξήματος και να το χρησιμοποιήσει σωστά. Ως εκ τούτου, η σημασιολογική ανάλυση των τοπωνυμίων και παρόμοιων λεξημάτων “με ιδιαίτερη πολιτιστική σημασία” πρέπει να γίνεται όχι αποκλειστικά με σημασιολογικές μεθόδους, αλλά και με βάση τα εξωγλωσσικά στοιχεία που αποτελούν “κοινές γνώσεις” των ομιλητών της νέας ελληνικής.

Κατά τη διάρκεια της ανάλυσης εντοπίστηκαν επίσης μερικά τοπωνύμια η μεταφορική ή μετωνυμική χρήση των οποίων είναι συχνή στο προφορικό και γραπτό λόγο αλλά δεν αναφέρεται στα λεξικά (*Ιμια, Νυρεμβέργη, Πρέσπες*). Τα παραπάνω δεδομένα είναι δυνατόν να ληφθούν υπόψη κατά τον εμπλουτισμό των υπαρχόντων λεξικών της νεοελληνικής, καθώς και κατά την περιγραφή του “πολιτιστικού λεξικού” της νέας ελληνικής.

#### 5 References

- Atchinson, J. (2008). *Words in the mind : An introduction to the mental lexicon*. Oxford; Cambridge, Blackwell Publishing.
- Bartsch, R. (2003) Generating polysemy: Metaphor and metonymy. In R. Dirven, R. Pörrings (eds.) *Metaphor and Metonymy in Comparison And Contrast*. Berlin, New York, Mouton de Gruyter, pp. 49-74.
- Boas, F. (1940) *Race, language and culture*. New York: The Macmillan Company.
- Fraser, P., Matthews, E. (eds.) (2010) *Lexicon of Greek personal names*. Oxford: Clarendon Press.
- Gatto, M. (2014) *The Web as a Corpus : theory and practice*. London, New York, Bloomsbury.
- Goddard, C., Wierzbicka, A. (eds.) (2002) *Meaning and Universal Grammar. Theory and empirical findings*. V. 1. Amsterdam, Philadelphia, John Benjamins Publishing Company.
- Kövecses, Z. (2002) *Metaphor : A Practical Introduction*. Oxford : Oxford University Press.



- Krasnykh, V. V. (2003) "Svoy" sredi "chuzikh": mif ili realnost? [At home or among strangers : myth or reality?] Moscow, Gnozis.
- Krasnykh, V. V. (2016) *Slovar i grammatika lingvokultury : Osnovy psikholingvokulturologii* [Lexicon and Grammar of Linguistic Culture : Prolegomena to psycholinguoculturology]. Moscow, Gnozis.
- Lakoff, G., Johnson M. (1980) *Metaphors we Live By*. Chicago: The University of Chicago Press.
- Lakoff, G. (1987) *Women, Fire, and Dangerous Things*. Chicago, London: The University of Chicago Press.
- Matthews, E. (ed.) (2007) *Old and New Worlds in Greek Onomastics* (Proceedings of the British Academy). Oxford: Oxford University Press.
- Sapir, E. (1973) Language and Environment. In Mandelbaum D. (ed.). *Selected writings of Edward Sapir in language, culture and personality*. Berkeley: University of California Press, pp. 89-103.
- Tent, J. (2015) Approaches to research in Toponymy. In *Names: A Journal of Onomastics*, 2 (63), pp. 65-74.
- Wierzbicka, A. (1997). *Understanding cultures through their key words*. Oxford: Oxford University Press.
- Woodman, P. (2012) Toponymy in a landscape of aggression: Geographical names in National Socialist Germany. In P. Woodman (ed.) *The great toponymic divide: Reflections on the definition and usage of endonyms and exonyms*. Warszawa, Head Office of Geodesy and Cartography, pp. 271-302.
- Zinken, J. (2008) Linguistic pictures of the world or language in the world? Metaphors and methods in ethnolinguistic research. In *Etnolingwistyka*, 20, pp. 51-62.
- ΠΑ - Εικονογραφημένο εγκυκλοπαιδικό λεξικό και πλήρες λεξικό της νέας ελληνικής γλώσσας. 2003. Αθήνα, ΠΑΠΥΡΟΣ.
- Πολίτης, Ν. (1899) Μελέται περί του βίου και της γλώσσης του ελληνικού λαού. Παροιμιαί. Τ. Α'. Αθήνα, Σακελλαρίου.
- Συμειωνίδης, Χ. Π. (2010) *Ετυμολογικό λεξικό των νεοελληνικών οικωνυμίων. Ττ. 1-2*. Λευκωσία, Κέντρο Μελετών της Ιεράς Μονής Κύκκου.
- ΧΡ - Χρηστικό λεξικό της νεοελληνικής γλώσσας. 2014. Αθήνα, Ακαδημία Αθηνών.



# Definitions of the Oxford English Dictionary and Explanatory Combinatorial Dictionary of I. Mel'čuk

Margalitadze T.

Ivane Javakhishvili Tbilisi State University, Georgia, e-mail: tinatin.margalitadze@tsu.ge

**Abstract:** The Oxford English Dictionary (OED) was an innovative dictionary from many points of view. The paper focuses on one of such innovative features of the OED, namely the method of description of word meaning. One of the ambitions of the OED team was 'to show more clearly and fully than has hitherto been done, or even attempted, the development of the sense or various senses of each word from its etymology and from each other'. For this purpose, the OED editors described semantic structures of English words, mechanisms of development of transferred senses from different semantic components of word meaning. This approach transformed the OED definitions into a very valuable source for the study and investigation of semantic structures of English words.

I. Mel'čuk's theory has a considerable impact on the development of methodology of semantic description of different languages. This theory, from my point of view, is also interesting as it has returned to the lexicographic practice and further elaborated long forgotten great ideas of the OED editors, and particularly James Murray. The paper discusses some parallels between the OED semantic theory and the Explanatory Combinatorial Dictionary (ECD) of I. Mel'čuk.

**Key words:** the OED; semantic theory; I. Mel'čuk; the ECD; polysemous models

## 1. Introduction

On April 19, 1928 the last, 125<sup>th</sup> fascicle of *The New English Dictionary on Historical Principles* was printed and it was immediately followed by the publication of the full Dictionary in ten bound volumes. The appearance of the second edition of the Dictionary in twenty volumes in 1989, under the new title - The Oxford English Dictionary (OED), had enormous response from the public and the media: '... the great publishing event of the century', '... a scholarly Everest', '... one of the wonders of the world' – these are a few excerpts from newspaper articles of the time<sup>1</sup>.

This enthusiasm and positive reaction to the publication of the OED is in no way an exaggeration. The dictionary, conceived and created in the second half of the XIX century fascinates not only by the scope of work implemented in it, but also by innovative approach to all aspects of dictionary-production. 'I feel that in many respects I and my assistants are simply pioneers, pushing our way experimentally through an untrodden forest, where no ... man's axe has been before us', - wrote J. Murray, one of the longest serving editors of the OED (Mugglestone 2000: 1). The OED team collected the biggest corpus for the project, over 10 mln illustrative sentences for sense discrimination and description. This material was collected from 5,000 sources, covering seven centuries of the development of the English language. 249,300 etymologies, included in the Dictionary were based on the scholarly approach, introduced by the comparative-historical method; all 2,412,400 usage quotations of the OED had references to sources and dates, and many others.

These principles were developed long before the electronic age. But it was exactly this innovative approach, which enabled software developers of the end of the XX century to create a CD-ROM version of the OED second edition (1992) with numerous search functionalities. E.g. it is possible to search English words by date of their appearance in English; It is possible to search for cognates of English words in Sanscrit, Latin, Greek, Avestan, Old Irish, Lithuanian, Old Church Slavonic or any other old/dead or/and new Indo-European languages, because this information is provided in etymologies of words; It is possible to search for Shakespearean quotations in the Dictionary, which helps to understand the Shakespearean meaning of a word, often different from its Modern English meaning. Likewise, it is possible to look up biblical quotations, or quotations from any other authors, included in the OED, etc.

## 2. Semantic Theory of the OED

The present paper focuses on one of such innovative features of the OED, namely the method of description of meaning of English words.

One of the achievements of the OED was the separation of etymology from semantics, rigorous scholarly research of the empirical material and fine sense-division. Besides, the Philological Society had an important principle, well formulated in the quotation given below from *Proposal for the Publication of a New English Dictionary by the Philological Society* (1859): 'In the treatment of individual words the historical principle will be uniformly adopted; - that is to say, we shall endeavor to show more clearly and fully than has hitherto been done, or even attempted, the development of the sense or various senses of each word from its etymology and from each other' (Silva 2000: 78). This

<sup>1</sup> [https://en.wikipedia.org/wiki/Oxford\\_English\\_Dictionary](https://en.wikipedia.org/wiki/Oxford_English_Dictionary) [accessed 24.01.2020]



principle aimed to reveal the mechanisms of appearance of transferred senses either from the primary meaning of a word, or from its other polysemous meanings.

The methodology of defining meaning in the XIX century dictionaries followed the tradition of Aristotelian and Thomistic philosophy, which is known as a definition ‘per genus proximum et differentias specificas’, i.e. ‘by stating the superordinate class to which something belongs, together with the specific characteristics that differentiate it from the other members of the class’ (Geeraerts 2010: 76). But the OED editors went even further and, in their definitions, brought forward and described supplementary features of meaning, different so-called potential components of meaning. As N. Hultin writes in his ‘The Web of Significance: Sir James Murray’s Theory of Word-Development’: the dictionary reflects ‘an implicit theory of language in which reason acts as the guide for the development of word signification’ (Silva 2000: 79). This principle was implemented in dictionary definitions of the OED transforming it into a very valuable source for the study and investigation of semantic structure of English words.

The adoption of this principle, revealing semantic transfer processes, implied description and explication of the basis of this change in a dictionary definition, i.e. a concrete semantic component of the semantic structure of a word, which served as the basis of metaphor, metonymy and other mechanisms of semantic change and determined the appearance of this or that new sense.

Moreover, the OED entries have several levels of numbering, using Roman numerals, Arabic numerals and letters of the alphabet in order to show the **sense-structure** of a polysemous word.

As an example, below will be given the analysis of dictionary definitions of some polysemous meanings of the adjective **thick**: 1) dense, crammed (*thick forest, thick hair ...*); 2) numerous, occurring in large numbers in a limited area (*a thick crowd ...*); 3) viscous in consistency (*thick coffee, thick soup, thick greese*); 4) having the component particles densely aggregated so as to hinder vision (*thick fog, thick smoke, thick mist ...*).

The definitions from the OED, namely definitions 4.a., 5.a., 6.a., 7.a – reveal the semantic component - **consisting of closely occupied, filled or set individual components** - which is the basis for the above-cited meanings of **thick** (see Picture 1).

Picture 1.

4.a. Closely occupied, filled, or set with objects or individuals; composed of numerous individuals or parts densely arranged; dense, crowded. Of hair: Bushy, luxuriant.

5.a. Of the individual things collectively: Existing or occurring in large numbers in a relatively small space, or at short intervals; densely arranged, crowded; hence, numerous, abundant, plentiful.

6.a. Having great or considerable density, either from natural consistence or from containing much solid matter; dense, viscid; stiff. (Said of liquids, semi-liquids, etc).

7.a. Of mist, fog, smoke, etc.: Having the component particles densely aggregated, so as to intercept or hinder vision. Hence of the weather, etc.: Characterized by mist or haze; foggy, misty.

**Individual components**, as is clear from the OED definitions, may be trees (as in a forest), or human beings (as in a crowd), or solid matter (as in liquids), or hair, or particles of mist, fog, smoke and so on.

Thus, from the analysis of the OED definitions it is possible to reconstruct the development of transferred meanings of **thick**:

- 1) dense, crammed  
*thick forest, thick hair ...*  
(**consisting of closely occupied, filled or set hair, trees**)
- 2) numerous, occurring in large numbers in a limited area  
*a thick crowd, thick throngs of young people*  
(**consisting of closely occupied, filled or set individuals**)
- 3) viscous in consistency  
*thick coffee, thick soup, thick greese*  
(**consisting of closely occupied, filled or set components of solid matter**)
- 4) having the component particles densely aggregated so as to hinder vision  
*thick fog, thick smoke, thick mist, the air was thick*  
(**consisting of closely occupied, filled or set component particles of mist, fog, smoke, etc**).

These meanings of **thick** have several level of numbering. Apart from using Arabic numerals and letters, Roman numeral II is also used to show that meanings 1) – 4) develop on the basis of one semantic component. Thus, definitions reveal not only semantic component and mechanism of semantic change on its basis, but an entry also shows the **sense-structure** of a polysemous word. The next 3 meanings of **thick** have Roman numeral III, indicating that they are



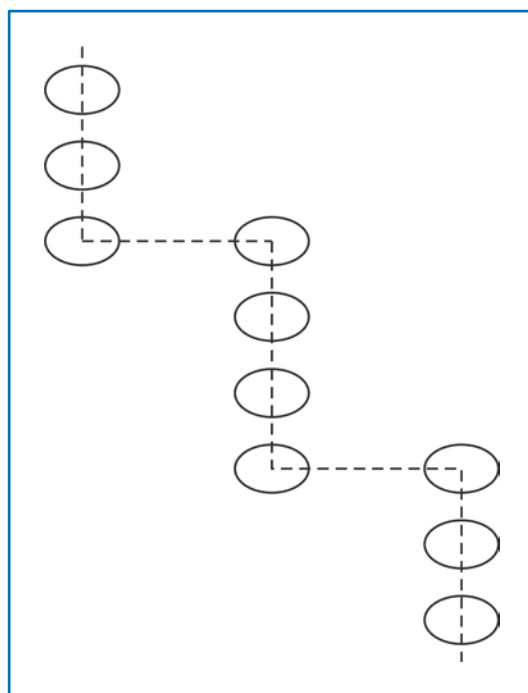
developed on the basis of the semantic component *being indistinct*. This component itself is generated from the meaning of **thick** - having the component particles densely aggregated so as **to hinder vision** (definition 7.a.).

- 1) not clear, foggy, misty (of weather)  
*thick day, thick morning, thick weather*  
(*being indistinct in vision [because of haze, fog, etc]*)
- 2) imperfectly articulated, muffled, guttural  
*thick words*  
(*being indistinct in sounding*)
- 3) (of hearing) dull of perception  
*thick of hearing*  
(*being indistinct in hearing*)
- 4) obtuse, stupid  
*thick-headed*  
(*being indistinct in mind, understanding*).

As a result of this analysis, the polysemous model of the adjective **thick** can be drawn (see Picture 2). I termed this model a **tiered model** (Margalitadze 1982). A dotted line indicates the semantic component, the basis for the development of transferred senses, an ellipse represents a polysemous meaning, developed on the basis of this semantic component. Thus, the sense structure shows three strings of senses: The first one, the primary meaning of **thick** has the semantic component *having relatively great extension between the opposite surfaces*, which also generates some other meanings:

- 1) having relatively great extension between the opposite surfaces  
*thick book, thick wall, thick glass*  
(*having relatively great extension between the surfaces*)
- 2) deep  
(*having relatively great extension from top to bottom*)
- 3) heavily built, burly, muscular  
*thick shoulders, thick figure, thick arms*  
(*having relatively great bulk "from one surface to its opposite"*).

The second string of senses is generated on the basis of the semantic component: *consisting of closely occupied, filled or*



*set individual components*. The third string has the semantic component: *being indistinct* (see the discussion above).

Picture 2.



A different method is adopted for the description of nouns and their meanings in the OED. Dictionary definitions of the noun **heart** (*The hollow muscular or otherwise contractile organ which, by its dilatation and contraction, keeps up the circulation of the blood in the vascular system of an animal*) reveal numerous semantic components **heart** is associated with in English: *the seat of life; the seat of one's inmost thoughts and secret feelings; the seat of the emotions generally; the seat of love or affection; the seat of courage; the seat of the mental or intellectual faculties; the innermost or central part of anything; the vital, essential, or efficacious part, etc.*

Below are given definitions from the entry **heart** which contain the above-cited semantic components:

Definition 2. Considered as the centre of vital functions: **the seat of life**; the vital part or principle; hence in some phrases = life. Obs. or arch.

Definition 6.a. **The seat of one's inmost thoughts and secret feelings**; one's inmost being; the depths of the soul; the soul, the spirit.

Definition 9.a. **The seat of the emotions generally**; the emotional nature, as distinguished from the intellectual nature placed in the head.

Definition 10.a. More particularly, **The seat of love or affection**, as in many fig. phrases: to give, lose one's heart (to), to have, obtain, gain a person's heart. Hence = Affection, love, devotion.

Definition 11.a. **The seat of courage**; hence, Courage, spirit. Especially in to pluck up heart, gather heart, keep (up) heart, lose heart.

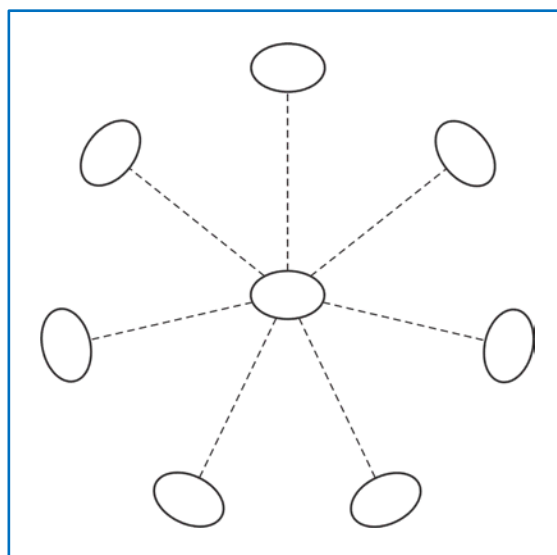
Definition 12. **The seat of the mental or intellectual faculties.** Often = understanding, intellect, mind, and (less commonly) memory. arch. exc. in phrase by heart.

Definition 17.a. **The innermost or central part of anything**; the centre, middle.

Definition 20. **The vital, essential, or efficacious part** (essence)

And so on.

Dictionary definitions clearly point to the fact, that above-cited semantic components are features, associated with **heart**, therefore they are supplementary components of the main meaning of the word. The transferred meanings of **heart** are based on these semantic components, as is seen from the definitions above. They surround the main meaning of **heart**



like rays of the sun (see Picture 3). Hence the name, I gave to this model – a **solar model** (Margalitadze 2006).

Picture 3.

Still another method of description is adopted for adjectives, denoting very general features. The entry for the adjective **straight**, for example, brings forth the general semantic component – *free from curvature, bending or angularity* (definition 2.a.) which is present in every lexical unit of the adjective **straight**, but in each sense it is concretized by different features: e.g. human body, deportment, course, flight, conduct, socially acceptable behavior and so on. This type of description enabled me to identify one more model of polysemous adjectives (and also verbs), which I call **one-dimensional model** (Margalitadze 1982; 2014).

Each entry of the OED is a real treasure for semantic analysis of a word and is a proof that the OED editors remained true to their intention “... to show more clearly and fully than has hitherto been done, or even attempted, the development of the sense or various senses of each word from its etymology and from each other” (Silva 2000: 78).



### 3. Method of Analysis of Dictionary Definitions

The OED was a well-known dictionary in Russian, and generally, Soviet lexicology and lexicography. *A New English-Russian Dictionary*, edited by an outstanding Russian linguist and lexicographer Ilya Galperin (NERD), one of the best bilingual dictionaries published in Russia, was largely based on the OED. This fact, from my point of view, influenced the popularity of the Distinctive-Feature Semantics in Soviet linguistics which viewed word meaning as a structure consisting of semantic components arranged in a hierarchical order (Margalitadze 2018: 250–253).

The knowledge of the OED, the knowledge of the method of defining meaning in the OED in an analytical way, by splitting it up into more basic semantic components, gave rise to the development of one of the methods of componential analysis of meaning – the so-called definitional method of analysis (see e.g. Arnold, 1966), or method of analysis of dictionary definitions, Дефиниционный Метод Анализа. The method is based on the comparison and analysis of definitions of comprehensive explanatory dictionaries, primarily definitions of the OED. Definitions of an entry are not always explicit about semantic structure of a lexical unit. Therefore, a variety of this method was developed later, called the Method of Transformation of Definitions (Arnold, 1979). According to this method, when a dictionary definition is not sufficient and does not fully reveal semantic structure of a lexical unit, in this case, some words from this definition are replaced by their definitions from the same dictionary and this process continues until the whole semantic structure and components of meaning of a word are revealed.

This method was actively used in the Soviet linguistics as one of the methods of study word meaning. The editorial team of the Comprehensive English-Georgian Dictionary (CEGD) also largely relied on the OED and its definitions for the analysis of English words included in the CEGD. I still use this method as one of the methods of semantic research and, alongside with other methods, teach it to my students of MA program in lexicography (Margalitadze 2006).

### 4. Explanatory Combinatorial Dictionary of I. Mel'čuk

Explanatory Combinatorial Dictionary (ECT) of I. Mel'čuk, which is based on his Meaning-Text Theory (MTT) is a complex theory which cannot be dealt in all details in the present paper. The aim of this chapter is to discuss some important aspects of semantic description of word meaning in the ECD, which, from my point of view, are reminiscent of the great ideas of the OED editors and particularly James Murray.

I. Mel'čuk, an outstanding Russian and Canadian linguist, lived in Russia till 1977, before emigrating to Canada. He undoubtedly had deep knowledge of semantic theories based on the OED, wide-spread and popular in Soviet linguistics. This knowledge is perceptible in some parts of his theory.

I. Mel'čuk introduces three linguistic conditions that the definition of a lexical unit (LU<sup>2</sup>) must conform to in the ECD: denotational potential of a LU, showing its links with the extralinguistic world; paradigmatic potential, explicating semantic links of a LU with related LUs in the lexicon; and syntagmatic potential showing syntagmatic links of a LU with other LUs in the sentence (Mel'čuk, 2013: 282). Thus, description of a paradigmatic potential of a LU aims at revealing its semantic links or 'semantic bridges' with other related LUs of a lexicon. Description of a paradigmatic potential of a LU comprises polysemy, derivation and phraseology (Mel'čuk, 2013: 298). A 'semantic bridge' is a very important concept in the theory of I. Mel'čuk. In case of polysemy, 'semantic bridges' link LUs of a vocable<sup>3</sup> by revealing hidden semantic links between them. An example of a 'semantic bridge' is discussed for LUs of the vocable CLOUD on the page 298 (Mel'čuk, 2013).

LU = CLOUD<sup>I</sup> is defined as 'accumulation of grayish white substance...that partially hides the sky'. 'Semantic bridge' in this definition is '...that partially hides the sky'. This 'semantic bridge' appears in the definition of the other LU of the vocable CLOUD, CLOUD<sup>III</sup>

(σ)<sup>4</sup> = (... that partially hides the sky);

LU = CLOUD<sup>III</sup> 'fact X ... that (partially) spoils the positive character of the fact Y [as if X were a cloud that partially hides the sky]' (as in *This sad news was the only cloud on the otherwise excellent vacation*).

The existence of CLOUD<sup>III</sup> shows the linguistic relevance of the component (σ) in the definition of CLOUD<sup>I</sup>. The semantic link between CLOUD<sup>III</sup> and CLOUD<sup>I</sup> is obvious to an English speaker – it is a comparison with CLOUD<sup>I</sup> [a live, even if conventional, metaphor]; it has to be shown in the definitions of both lexemes. On the other hand, according to the concept of vocable, two lexemes of the same vocable should explicitly manifest their semantic bridge. As a result, we have to include the component (σ) in the definition of CLOUD<sup>I</sup>, which allows us to have the component (as if X were a cloud that (partially) hides the sky...) in the definition of CLOUD<sup>III</sup>, and the semantic link – a semantic bridge – is ensured (Mel'čuk 2013: 299).

'Semantic bridges' reveal hidden semantic transfer mechanisms functioning between different polysemous meanings of a word, linking them to one another. LUs of a vocable maybe linked directly by a 'semantic bridge' or indirectly by a chain of 'semantic bridges' via other LUs of a vocable. Different configurations of 'semantic bridges' result in radial polysemy or chain polysemy (Mel'čuk 2013: 325).

It is not difficult to see similarities between 'semantic bridges' and the definitions of the OED, discussed in chapter 2 of the present paper. The OED is a historical dictionary and the aim of its editors was to show the development of senses from the primary meaning of a word or from each other. In order to reveal this mechanism of sense development, editors

<sup>2</sup> LU corresponds to a polysemous meaning of a word.

<sup>3</sup> Vocable corresponds to a polysemous word.

<sup>4</sup> A symbol of a 'semantic bridge'.



had to describe and explicate in dictionary definitions those semantic components, semantic links, 'semantic bridges' in the terminology of I. Mel'čuk, which served as the basis for the appearance of transferred senses of polysemous words. Unlike the OED, the ECD is not a historical dictionary, therefore it aims to show how LUs are linked. For this purpose it is necessary to describe the linking semantic components, 'semantic bridges' in dictionary definitions. It is not also difficult to see parallels between radial and chain sense structures and sense structures suggested by the definitions of the OED, discussed above.

Further I. Mel'čuk expounds on the linguistic relevance of a semantic component in a dictionary definition for derivation and phraseology. The similar approach to this issue can be also amply illustrated by the OED definitions.

Like the OED, I. Mel'čuk also introduces several levels of sense numbering with Roman numerals, Arabic numerals and letters. Roman numerals express larger semantic distances between LUs, Arabic numerals – smaller ones, and letters – the smallest distances between LUs.

## 5. Conclusion

Great ideas that lexicographers developed in the XIX century, theories of description of word meaning proposed by them, were well forgotten in the first half of the XX century. After being the philologists' prime object of investigation in the XIX century, the lexicon had been neglected in favour of syntax and phonology, as it was more difficult to describe and encapsulate it in rules (Bejoint 2010: 264).

As you surely know, one of the many surprising facts about the discipline of linguistics in the 20th century was that the study of lexis and meaning was largely neglected in America, Britain, and their spheres of influence. Honourable exceptions were in the European Saussurean tradition — notably German semantic field theorists such as Trier, Porzig, and Weisgerber and the Romanian Eugene Coseriu; British Firthians such as Halliday and Sinclair, Russians such as Mel'čuk and Apresjan, and others. But these past researchers were hampered by, among other things, lack of evidence and the political crises of their time", wrote Patrick Hanks in the new proposal of the University of Wolverhampton "Studying meaning in the 21st century" (Margalitadze 2018).

It is very disappointing to realize that this theory of description of lexicon of a language, and generally, the study of semantics was neglected for decades after the tremendous achievement of the OED team (Margalitadze 2018: 250). I. Mel'čuk's innovative theory has considerable impact on the development of methodology of semantic description of different languages. This theory, from my point of view, is also interesting, as it has returned to the lexicographic practice and further elaborated long forgotten great ideas of the OED editors, and particularly James Murray.

Why is the OED semantic theory important for the modern theory of lexicography, as well as dictionary-making practice? The OED theory of description of meaning provides excellent scientific study of each word, its meaning, its semantic structure, semantic transfer mechanisms underlying its polysemous meanings, sense structure of polysemous meanings of a word. The study of each word is based on rigorous scholarly research of the vast empirical material. Such description of word meanings turns the OED into a reliable source for the study of polysemy or meaning in general, for the comparative study of word meanings in different languages, for the study of complex semantic processes which are at play in a language. The OED is one of the reliable sources in bilingual projects, as its entries help bilingual lexicographers understand every shade of meaning of a word in order to provide it with adequate equivalents in another language. The dictionary helps bilingual lexicographers solve many problems, including the problem of equivalence between languages. Our editorial team has worked with the OED for 35 years, while editing entries of the CEGD and we know the efficiency and reliability of this great lexicographic endeavor from our personal experience.

I strongly believe that implementation of the OED approach is important in monolingual, explanatory dictionaries and I. Mel'čuk's theory is the proof that I am not alone in this belief.

"The structure now reared will have to be added to, continued, and extended with time, but it will remain, it is believed, the great body of fact on which all future work will be built", – James Murray (Silva 2000: 94).

## 6. References

- Arnold, I.V. (1966). *Semantic Structure of Word in Modern English and Methods of its Study*. Leningrad: Prosveshchenie (in Russian).
- Arnold, I.V. (1979). Potential and Hidden Semantic Components and their Actualization in English Literary Texts. In: *Inostrannye Yazyki v Shkole*, № 5, pp. 10–14 (in Russian).
- Béjoint, H. (2010). *The Lexicography of English*. Oxford: Oxford University Press.
- CEGD = *Comprehensive English–Georgian Dictionary* (eds. Margalitadze, Chanturia et al), vol. I–XIV (1995–2012). The Online Version: [www.dict.ge](http://www.dict.ge). Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.
- Geeraerts, D. (2010). *Theories of Lexical Semantics*. New York: Oxford University Press.
- Mel'čuk, Igor (2013). *Semantics. From Meaning to Text*, vol. 2. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- NERD = *A New English-Russian Dictionary* (ed. Galperin) in 2 volumes (1987–1988). Moscow: Russky Yazyk Publishers.
- Margalitadze, T. (1982). The Main Models of the Semantic Structure of Adjectives in Modern English. In *Bulletin of the Academy of Sciences of the Georgian SSR*, 105(3), pp. 181–184.
- Margalitadze, T. (2006). *Meaning of a Word and Methods of its Research*. Tbilisi: Ivane Javakhishvili Tbilisi State



- University, Lexicographic Centre. [www.margaliti.com](http://www.margaliti.com).
- Margalitadze, T. (2014). Polysemous Models of Words and Their Representation in a Dictionary Entry. In Abel, A., C. Vettori and N. Ralli. (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 1025–1037. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism. <https://euralex.org/category/publications/euralex-2014> [28/05/2020]
- Margalitadze, T. (2018). Once Again Why Lexicography Is Science. In *Lexikos*, 28. <https://lexikos.journals.ac.za/pub/issue/view/86> [28/05/2020]
- Mugglestone, L. (2000). 'Pioneers in the Untrodden Forest': The New English Dictionary. In *Lexicography and the OED. Pioneers in the Untrodden Forest* (ed. Mugglestone L.). Oxford: Oxford University Press, pp. 1–22.
- OED = *Oxford English Dictionary on Historical Principles* (Vol. I–XIII; suppl. I–IV). Oxford: Oxford University Press.
- Silva, P. (2000). Time and meaning: Sense and Definition in the OED. In *Lexicography and the OED. Pioneers in the Untrodden Forest* (ed. Mugglestone L.). Oxford: Oxford University Press, pp. 77–96.







# Frame Semantics in the Specialized Domain of Finance: Building a Termbase to Aid Translation

Pilitsidou V.<sup>1</sup>, Giouli V.<sup>2</sup>

<sup>1</sup> National and Kapodistrian University of Athens, Greece

<sup>2</sup> Institute for Language and Speech Processing, ATHENA Research Centre, Athens

## Abstract

Frame semantics (Fillmore 1977, 1982, 1985) is one of the most important developments for lexicography in the 20th century. The semantic frames approach to lexicon building and semantic representation of meaning at word and phrase level – or even beyond – has been the focus of research in computational linguistics and in Natural Language Processing. The present paper is aimed at describing completed work for the creation of a domain-specific frame-semantic lexicon in Greek (EL) and its alignment to the English (EN) FrameNet. Building on Fillmore's Frame Semantics (Fillmore 1977, 1982, 1985) and on the example set by the FrameNet project (Baker et al. 1998), we developed a bilingual EL-EN lexical resource in the financial domain based on corpus evidence. Our motivation was two-fold: (a) to better account for the semantics of the specialized lexicon – especially the verbs and predicative nouns of the financial domain, and (b) to make cross-lingual alignments at the word level in a way that is meaningful for the translation process.

**Keywords:** frame semantics; FrameNet; frame; financial domain; translation; terminology; terminological resource

## 1 Introduction

The development of FrameNet (Baker et al. 1998) and FrameNet-compatible language resources as both human- and machine-readable lexica is based on annotating examples of how words are used in actual texts. Frame Semantics has many possible applications, terminography and domain-specific translation being one of them. In this paper we describe completed work for the creation of a terminological resource for the specialised domain of finance. The resource will be based on the principles of Frame Semantics (Fillmore 1977, 1982, 1985) and the example of FrameNet (Baker et al. 1998). The paper is organised as follows. In section 2 we briefly present the theoretical framework and the previous work on related projects, especially regarding the Greek language. The research scope and goals are set out in section 3. Section 4 and its subsections outline the methodology adopted towards developing the bilingual resource, whereas section 5 provides a detailed description of the resource and its components. In section 6 we discuss different aspects of the procedures and the results of our work and, finally, in section 7 we provide our conclusions and prospects for future research.

## 2 Theoretical Framework and Related Work

The theory of Frame Semantics by Charles J. Fillmore (Fillmore 1977, 1982, 1985) focuses on the continuity that exists between language and experience (Petruck 1996). According to this theory, words gain their meaning in a semantic frame which can be an event or a relation. In this context, the term “semantic frame” or “frame” refers to any system of meanings which is connected in a way that to understand any one of these meanings we must be able to understand the whole structure to which it belongs; when one of the elements of such a structure is used in a text or a discussion, then all the other elements automatically become available (Fillmore 1982: 111). Fillmore calls these elements “Frame Elements” (FEs). The words that evoke the semantic frames are called “Lexical Units” of the frame (LUs) and they are predicates which are mainly verbs, other parts of speech (names, adjectives, adverbs) as well as multi-word expressions (Tantos et al. 2015: 167).

Frame Semantics is the theoretical framework on which FrameNet (Baker et al. 1998), a lexical resource for the English language, is based. This semantic representation includes Frames and their LUs and allows the connection of all the grammatical categories (noun, adjective, verb, adverb) with a Frame.

Consequently, the theory of the semantic Frames has been further utilised for the formulation of the Frame-based Terminology (FBT) theory and for the concomitant creation of terminological bases. According to this approach, the way that the senses which belong to a thematic field are realised and connected with each other depends on the events of the field (task-oriented). FBT is, according to Faber (2011, 2012, in Faber 2014:14), a cognitive approach to terminology which directly connects the specialised knowledge with Cognitive Linguistics and Semantics. It uses a modified version of Fillmore's Frames (Fillmore 1982, 1985) along with the premises of Cognitive Linguistics (Faber 2011).

In this context, Faber (2011) describes the specialised language as dynamic and supports that its representation should be dynamic as well, although this approach is not being used adequately in the terminological resources. She further supports that the way that items are represented in our brain means that current methods and ways of creating representations of specialised knowledge should be modified in order to take this information into account. Specialized language concepts cannot be activated in isolation unless they are part of a larger structure or event. Our knowledge about a concept initially gives us the context or the event in which the concept has a meaning for us. Consequently, concept representations should, instead of being presented out of context and being static, be presented inside their context and be dynamic (Faber 2011).



Over the years, a number of frame-based language resources have been developed for other languages for general purposes (FrameNet Brazil (Salomão 2009), Spanish FrameNet (Subirats 2009) and Japanese FrameNet (Ohara 2009), and the Swedish FrameNet++ (Ahlberg et al. 2014), inter alia).

As far as the Greek language is concerned, there has been previous work in language for general purposes, but no work has been reported for language for specific purposes. In fact, an initial attempt to build a frame semantics lexical resource for Greek is reported in Gotsoulia et al. (2007); however, this work was conceived of as the preliminary phase of a pilot project for the development of the basic infrastructure and design of the actual resource. Later, a frame-driven approach was followed by Dalpanagioti (2012) to the bilingual lexicographic process for creating a bilingual lexical database of motion verbs for EL and EN. Yet, these studies are fragmented and represent only a part of the general language. Finally, Giouli et al. (2020) report on work towards the development of the Greek (EL) counterpart of the Global FrameNet in the context of the Shared Annotation Task, where the annotation methodology employed, the current status and progress made so far, as well as the problems raised during annotation of the EL corpus are put into focus. In this paper, we have tried to compile a frame-based lexical resource for the domain of finance.

### 3 Research Scope and Goals

An important – yet still under investigation – characteristic of Frame Semantics and FrameNet is the universal nature of frames. In theory, frames are universal, as are the different concepts across languages. In practice, however, this has yet to be proven. One of the purposes of this work was to examine whether a new approach to terminology, the FBT, can have this characteristic. We tried to see if the frames that include terminology of a specific domain can also be used in another language. To this end, a parallel knowledge base of finance terms has been created for Greek and English.

The present work has been conducted from a translation perspective, as well. One of the aims is to use the resulting resource for facilitating the translation procedure. This is feasible, since FrameNet is structured in a way which makes it readable by people as well as computers. Therefore, it would be possible to utilise the term base in tools that assist translation (machine translation tools).

### 4 Methodology

For the creation of the term base, the methodology was viewed as a four-step approach: as a first step, the corpus was compiled, consisting of two sub-corpora – one for each language. The second step was the extraction, selection and grouping of terms; in particular, the terms were classified into groups according to their underlying meaning in order to be assigned a frame. The third step was the frame-creation step and other procedures entailed by it; after the creation of the frames that was led by our terms, definitions of FEs were provided and the terms were assigned a frame in both languages. At fourth step, the frames' sentences-examples were annotated in various layers. These steps are described in detail in the following subsections.

#### 4.1 Corpus Creation and Term Selection

For the purposes of this work, a special-purpose comparable corpus was created comprised by two sub-corpora – the EL sub-corpus and the EN sub-corpus. The EL sub-corpus is comprised by 73,069 tokens και 8,146 types, while the EN sub-corpus is comprised by 92,105 tokens and 7,129 types, size which was considered adequate for this work. As Pearson (1998: 56) mentions, when the corpus is designed for special purposes, then a smaller corpus which is derived from a given thematic field is more appropriate than a wider corpus.

The texts selected for this corpus belong to the same genre or thematic field and they were chosen according to their terminology content and coherence.<sup>1</sup> The sources are divided into two groups: (a) journalistic/news texts with finance content and (b) banks' financial results reports. The first group includes Greek and English articles from online newspapers with financial content as well as some with more general content; the second group, on the other hand, contains only financial texts or documents, which are very rich in terminology. This combination allowed us to find very specialised and less specialised terms. Both groups' articles cover topics such as economy, business, markets, stock market, bonds and banks - in other words, the corpus is compiled with texts from specialised fields, with high term ratio, which are suitable for terminological research.

The table below (Table 1) provides details on quantitative data of the two sub-corpora.

	Group (a) - EL	Group (b) - EL	Total EL	Group (a) - EN	Group (b) - EN	Total EN
<b>Texts</b>	121	10	131	92	5	97
<b>Tokens</b>	48,344	24,725	73,069	54,670	37,435	92,105
<b>Types</b>	7,333	2,118	9,451	6,579	1,845	8,424

Table 1: Quantitative data of the two sub-corpora.

<sup>1</sup> For the EL sub-corpus, the sources are: Capital.gr (<https://www.capital.gr/>), naftemporiki (<https://www.naftemporiki.gr/>), ΚΕΡΔΟΣ (<http://www.kerdos.gr/>), Η ΚΑΘΗΜΕΡΙΝΗ (<https://www.kathimerini.gr/>), ΕΘΝΙΚΗ ΤΡΑΠΕΖΑ (<https://www.nbg.gr/>) and ΤΡΑΠΕΖΑ ΠΕΙΡΑΙΩΣ (<https://www.piraeusbank.gr/el/idiwtes>).

For the EN sub-corpus, the sources are: CITY A.M. (<https://www.cityam.com/>), Bloomberg (<https://www.bloomberg.com/europe>), REUTERS (<https://www.reuters.com/>), The Guardian (<https://www.theguardian.com/international>), Barclays (<https://www.barclays.co.uk/>) and LLOYDS BANKING GROUP (<https://www.lloydsbankinggroup.com/>).



On step two, the terms of both sub-corpora were extracted semi-automatically using the software AntConc (Laurence 2016), and in particular its word list, keyword list, concordance and clusters functions for frequency analyses. To be more precise, initially the keywords were extracted and then they were examined in order to locate candidate terms, which later were individually processed through concordance and clusters in order to find candidate multi-word terms. Said procedures were followed in the same order for both sub-corpora.

At this initial stage, a total of 561 candidate terms were identified and selected that pertain to the Noun, Adjective and Verb grammatical categories; being a terminological work, we could locate a substantial number of multi-word nouns (also referred to as “multi-word terms”) as well. The distribution of the so-identified lexical items per grammatical category or part-of-speech (POS) is depicted in Table 2.

POS	EL	EN
Nouns (single)	100	100
Nouns (multi-word)	110	184
Adjectives	24	13
Verbs	14	16
Total	248	313

Table 2: POS of candidate terms.

## 4.2 Creation of Frames and Annotation

The third step, creation of the semantic frames, was the most important and the most laborious one. Our initial attempt was to use frames which have already been created for the lexical resource FrameNet (Ruppenhofer et al. 2016). To this end, the EL terms were grouped according to their meaning and the scene or frame that they could evoke, and then effort was made to assign FrameNet’s frames to them. For some of the terms this procedure was quite straightforward, such as those which evoke the frames of lending or borrowing, or those which belong to the scene of commerce. However, for terms of the language for special purposes – which formed the majority of our terms – the creation of new frames was necessary. This was accomplished with deep search in the FrameNet resource and extensive lexicographical research.

As a result of this process, the EL terms were grouped and divided into 9 scenes and 39 frames. At the next stage, we examined whether the frames which had been created according to the EL data could be used for the EN data as well, in order to find out if the already created frames can be used in a language other than the one they have been created. The decision to start with the creation of the EL frames was based on two considerations: firstly, in this way we could ensure that the terminology has been fully understood and organised into frames correctly, and secondly, that the conceptual structures created (the frames) adequately represent the respective terminology.

Since the lexical resource to be created was a terminological resource, we had to annotate (at this stage, manual annotation was performed to assign Core and Non-core FEs to the frames) beginning from nouns instead of verbs, so we deviated from FrameNet’s method, as the majority of terms are nouns (single- and multi-word) and only a small proportion of them are verbs. Therefore, the frames were defined according to the terms and the states or events that they represent. The method for the EL frames creation was corpus-driven, whereas that for the EN frames creation was corpus-based, because the already created frames had to be used. This is why the web as a secondary source was used for the EN frames in addition to the primary source for collecting examples in English in cases where the EN sub-corpus was not enough (the frames for which such examples were used are: Withdrawal, Bank\_account\_management, Stock\_exchange\_transactions, Owing and Social\_contributions).

Regarding the LUs that result from the frames, there are some which have not been derived from the corpus but needed to be added in order to complete the frame. These LUs were found either through the respective frame of the other language or from our general knowledge on a topic. For example, only few of the LUs of the frame Commerce existed in the corpus; however, having the LUs αγοράζω.ν (to buy) and αγοραστής.ν (buyer) without their opposites πουλάω.ν (to sell) and πωλητής.ν (seller) would make no sense. Additionally, some LUs have been assigned more than one frames indicating cases of polysemy or not very specialised terms – for example, the Greek noun επενδυτής.ν (investor) could be used in the frames Obtaining\_a\_loan, Bond\_issuing and Commerce.

Another important element that had to be added were the definitions of the FEs in the form of glosses; they were created according to each frame’s needs in a way that they represent the concepts involved in a frame and, consequently, the terminology itself as concretely as possible. Moreover, an attempt was made to create these definitions as language independent as possible. The example of FN and Frame Semantics was once again followed for this process: FEs’ definitions connect the concepts of each frame. To better account for the creation of the appropriate glosses (definitions), we consulted various lexica and reference works, such as dictionaries, term bases and language portals, namely: the Greek database of financial terms<sup>2</sup>, various dictionaries that are available on the Greek Language Portal<sup>3</sup>, EcoLexicon (Faber

<sup>2</sup> Available online at <https://www.euretario.com/>

<sup>3</sup> Available online at <http://www.greek-language.gr/greekLang/index.html>



2011)<sup>4</sup>, the English FrameNet database<sup>5</sup>, Glosbe<sup>6</sup> parallel dictionary, the EU's IATE (Interactive Terminology for Europe) terminology database<sup>7</sup>, Investopedia<sup>8</sup>, the EL and EN branches of Wikipedia, Linguee EL-EN database<sup>9</sup>, Merriam-Webster online dictionary<sup>10</sup>, Oxford Dictionaries<sup>11</sup>, and WordReference<sup>12</sup> online dictionary. These were used not only for assisting us to better understand the terminology, but also to properly create the FEs' definitions. An example of a frame (namely, the Obtaining\_a\_loan one) and the encoding performed in terms of annotated FEs and the definitions provided is depicted in Table 4.

The fourth and last step was the annotation of a second corpus including the sentences which were used as examples of the frames (also referred to as "sentences-examples"). This corpus is comprised of 255 sentences (EL: 130, EN: 125), or 6,923 words (EL: 3,832 words, EN: 3,091 words). In order to make the most out of this corpus, automatic pre-processing was in order. For this reason, the web tool UDPipe (Straka & Starková 2017) was used for processing the corpus comprising the sentences-examples of both languages, and tokenization, POS tagging and dependency parsing were performed. This pre-processing is essential because it provides a set of characteristics which are necessary for the following semantic analysis. Finally, annotation on lexical level was performed manually using the web annotation tool WebAnno (Yimam et al. 2013), which allows users to annotate texts in any level of lexical analysis by defining their own annotation scheme. Our annotation scheme were the FEs of our semantic frames, which appeared in the tool as tags. The number of tags we defined are 253, 147 of which are Core FEs and 106 Non-Core FEs. The tags are of course common for both languages, as the EL semantic frames were also used for EN.

## 5 Lexical Resource Description

The result of the previously described methodology is a bilingual terminological resource consisting of three components: (a) the EL and EN frames including examples, (b) the lexicon comprised of the LUs of the frames, and (c) the annotated corpus with the sentences-examples of the frames. All three components of this work are available and are described in detail below.

### 5.1 Frames and Scenes

The core of this work are the semantic frames. Together with the examples and the LUs that have been assigned to them, they are a way of presenting and understanding the terminology in both languages. Nine scenes have been created, which are divided into 39 frames. The scenes are the context in which a frame belongs. The following table (Table 3) illustrates the scenes and the frames.

<sup>4</sup> Available online at <http://ecolexicon.ugr.es/en/index.htm>

<sup>5</sup> Available online at <https://framenet.icsi.berkeley.edu/fndrupal/>

<sup>6</sup> Available online at <https://el.glosbe.com/>

<sup>7</sup> Available online at <https://iate.europa.eu/home>

<sup>8</sup> Available online at <https://www.investopedia.com/>

<sup>9</sup> Available online at <https://www.linguee.com/english-greek>

<sup>10</sup> Available online at <https://www.merriam-webster.com/>

<sup>11</sup> Available online at <https://www.oxforddictionaries.com/>

<sup>12</sup> Available online at <http://www.wordreference.com/>



SCENES	FRAMES
Scene 1: Lending	1. Obtaining a loan 2. Loan reimbursement 3. Increase/reduction of interest rate 4. Interest
Scene 2: Bank transactions	5. Deposit 6. Withdrawal 7. Bank account management 8. Way if transaction
Scene 3: Bonds	9. Bond issuing 10. Bond yield
Scene 4: Stock market	11. Stock 12. Stock market results 13. Stock indices 14. Stock exchange transactions 15. Distribution of earnings to shareholders
Scene 5: Financial results	16. Earnings 17. Profit 18. Expenditures 19. Ratio 20. Assets 21. Balance sheet data 22. Final results 23. Risks
Scene 6: Domestic economy	24. Lending a state 25. National budget 26. Facilitation for indebted state 27. Tax payment 28. Return to the financial markets 29. Owing 30. Economic performance index 31. Social contributions
Scene 7: Consumption	32. Commerce 33. Consumer spending 34. Price trend
Scene 8: Economy	35. Financing 36. Positive/negative economic activity 37. Change in price level 38. Financial crisis
Scene 9: Change in economy	39. Change position in a scale

Table 3: The scenes and the frames for the field of finance.

The frames are common for both languages and they are constructed as follows: the name of the frame is followed by its FEs. After that, the LUs are noted, which are the words or terms that evoke said frame, and at the end the frame's examples are listed, which are sentences from the corpus (or the web) that comprise the terminology of the frame. All frames include Core FEs and most of them Non-core FEs as well; the former are necessary elements for the conceptual structure that they describe, and the latter are non-obligatory elements.

As an example, the frame *Obtaining\_a\_loan* is shown in Table 4, along with the LUs and some of the examples in both languages. As one can see, the ID number of the frame (in our example: 1.), the name of the frame as well as all the Core and Non-core FEs with their definitions in the form of glosses are common in both languages and they are both written in English for reasons of consistency and better visualisation. The effort to create language independent FEs and definitions seems to have been successful, as for all our frames the same FEs were used for both languages, just like the example of the *Obtaining\_a\_loan* frame. This means that the general concepts exist in both EL and EN, given that the scene is common and the concepts, the states and the events that it expresses do not differ between the languages. The differentiation begins at the level of the LUs. In other words, the difference between the two languages lies on the lexicalisation of the concepts, rather than the concepts themselves. More details about the lexical units are provided at the next section.



1. Obtaining a loan - EL	1. Obtaining a loan - EL
<p>Core FEs:  Borrower: The person or institution who receives the Theme from the Lender for a Duration.  Lender: The person or institution who gives the Theme to the Borrower for a Duration.  Theme: The object that is transferred from the Lender to the Borrower for a Duration.  Non-Core FEs:  Duration: The amount of time in which the Borrower has possession of the Theme.  Manner: The way in which the Lender lends the Theme.  Place: The location in which the Lender lends the Theme to the Borrower.  Purpose: The aim of the Lender which they believe will be accomplished by lending the Theme to the Borrower.  Time: The time when the lending event occurs.  Amount: The amount of money of the Theme.</p>	<p>Core FEs:  Borrower: The person or institution who receives the Theme from the Lender for a Duration.  Lender: The person or institution who gives the Theme to the Borrower for a Duration.  Theme: The object that is transferred from the Lender to the Borrower for a Duration.  Non-Core FEs:  Duration: The amount of time in which the Borrower has possession of the Theme.  Manner: The way in which the Lender lends the Theme.  Place: The location in which the Lender lends the Theme to the Borrower.  Purpose: The aim of the Lender which they believe will be accomplished by lending the Theme to the Borrower.  Time: The time when the lending event occurs.  Amount: The amount of money of the Theme.</p>
LUs	LUs
<p>δανείζω.ν, δανείζομαι.ν, δάνειο.ν, παίρνω δάνειο.ν, λαμβάνω δάνειο.ν, δίνω δάνειο.ν, δανεισμός.ν, τραπεζικός δανεισμός.ν, δανειστής.ν, δανειστικός.α, δανειακός.α, δανειοδοτώ.ν, δανειοδότηση.ν, κόκκινο δάνειο.ν, τράπεζα.ν, εμπορική τράπεζα.ν, κεντρική τράπεζα.ν, συστημική τράπεζα.ν, αγροτική τράπεζα.ν, επενδυτής.ν, κεφάλαια ρευστότητας.ν, χρήματα.ν, χρηματοδότηση.ν, χρηματοδοτικός.α, πιστωτής.ν, πίστωση.ν, πιστωτικός.α, πιστώνω.ν, επαγγελματικό δάνειο.ν, επιχειρηματικό δάνειο.ν, στεγαστικό δάνειο.ν, ομολογιακό δάνειο.ν, εποχικό δάνειο.ν, άτοκο δάνειο.ν</p>	<p>borrow.v, lend.v, take out a loan.v, borrowing.n, lending.n, lender.n, creditor.n, issue.v, loan.n, home loan.n, mortgage.n, unsecured loan.n, consumer credit.n, defaulted loan.n, bond issue.n, bank.n, Bank of England.n, central bank.n, investment bank.n, investor.n, funding.n, secured lending.n, corporate lending.n, bank lending.n, corporate debt.n</p>
Examples	Examples
<p>Οριστική λύση για τη διαχείριση [theme κόκκινων δανείων] συνολικού ύψους [amount 1,75 δις. Ευρώ] που EIXAN ΛΑΒΕΙ [borrower 82.000 αγρότες και κτηνοτρόφοι] από την πρώην [lender Αγροτική Τράπεζα] δίνει το υπουργείο Αγροτικής Ανάπτυξης.</p> <p>Η [borrower κεντρική τράπεζα] μείωσε επίσης το αντίστροφο επιτόκιο επαναγοράς –το επιτόκιο με το οποίο ΔΑΝΕΙΖΕΤΑΙ [theme χρήματα] από τις [lender εμπορικές τράπεζες] - κατά 0,25% στο 5,75%.</p>	<p>The Commission hopes that by improving the securitisation process, where assets such as mortgages and consumer credit are bundled together and sold on to investors as bonds, it will unlock [amount up to €150bn] [theme of funding] for [lender banks] to LEND to [borrower consumers] and [borrower growing businesses].</p> <p>Today we get a new healthcheck on Britain's economy, with new figures showing how much new credit was lapped up by consumer last month, and how many [borrower people] TOOK OUT [theme mortgages].</p>

Table 4: The frame Obtaining\_a\_loan (EL, EN).

A fundamental principle for the creation of the frames was to use FrameNet's frames to the greatest extent possible. However, the FrameNet's frames that could be adopted without any modification were not a lot due to the different nature of the two resources. Particularly, this work's frames Stock (FN: Capital\_stock) and Change\_position\_on\_a\_scale (FrameNet: Change\_position\_on\_a\_scale) are the only ones which are exactly the same as the corresponding ones in FrameNet. To the frames Obtaining\_a\_loan (FrameNet: Lending) and Lending\_a\_state (FN: Lending) the FE *amount* was added, as our data required. Similarly, the frame Commerce (FN: Commerce\_sell) is the same as its corresponding one in FrameNet, with the addition of the FE *payment*. For the frame Loan\_reimbursement some FEs have been used (BORROWER, THEME, LENDER, TIME, AMOUNT, from Lending), while the rest were newly created. In the frame Earnings one FE from FrameNet's frame Earnings\_and\_losses was used (TIME, which was renamed into *time period*) as well as the names of two FEs (EARNINGS and EARNER), but not their definitions because they did not completely fit to the specific terminology. Moreover, for the definition of *profit.n* in the frame Profit we used FrameNet's *profit.n* as lexical entry. Similarly, for the definition of *asset.n* in the frame Assets FrameNet's definition of *asset.n* as lexical entry was used. Finally, in a number of frames (for example, Change\_in\_price\_level, Stock\_market\_results, Stock\_exchange\_transactions, Earnings, Profit, Expenditures, Ratio) the FEs FINAL VALUE and FINAL STATE of FrameNet's frame Change\_position\_on\_a\_scale have been used.

It should be mentioned that some FEs can be found in more than one frames; an example are the FEs Borrower and Lender of the frames Obtaining\_a\_loan, Loan\_reimbursement, Increase/reduction\_of\_interest\_rate and Interest of the Lending



scene as well as the frame Lending\_a\_state of the Domestic\_economy scene. All the frames except for one belong to the same scene, the scene that is about lending, which shows that scenes are the general context into which the different frames belong (in this case, the context of lending or borrowing).

## 5.2 Lexical Units

The resulting terminological resource consists of 374 LUs for EL and 368 LUs for EN. The LUs of the frames do not only include the terms that have been extracted from the corpus, but other LUs that evoke the frames as well. Therefore, we can say that the plurality of the LUs are the terms of this field, even though some of them are also used in general language. For example, the LUs *borrow.v*, *lend.v*, *take out a loan.v* of the frame Obtaining\_a\_loan are also used in language for general purposes.

Following again the example of FrameNet, each frame is accompanied by the LUs that evoke it. A bilingual lexicon has thus been created including the information shown on Table 5 which comprises some of the EN LUs of the frames Deposit and Withdrawal.

Entry ID	Lexical unit	Part of speech	Frame	Full form	Abbreviation	Alternative form	Synonym	Antonym	Hypernym	Example
EN_049	saving	noun	Deposit							yes
EN_050	transaction	noun	Deposit							no
EN_051	withdrawal	verb	Withdrawal					deposit		yes
EN_052	withdrawal	noun	Withdrawal					deposit	transaction	no
EN_053	money	noun	Withdrawal							yes
EN_054	bank	noun	Withdrawal						credit institution	yes

Table 5: Example of LUs of the lexicon.

This type of representation allows us to explore distinct meanings of the terms, particularly through the lexical relations (synonymy, antonymy, hypernymy/hyponymy) and the definitions of the FEs, and it provides a helpful tool for cases of polysemy. The lexical relations do not only connect the terms, but some frames as well, in a way that they make it easier for someone to understand the concepts that they represent. Here, it should be mentioned that the LUs are listed per frame and there are cases where an LU appears in more than one frames; this is one more way to study instances of polysemy.

It must be clear that there is no full form, abbreviation, alternative form, synonym, antonym and hypernym/hyponym for every LU. Also, even though we tried to provide examples for the majority of the terms, there are some terms or other LUs that evoke certain frames that are not in the sentences-examples of the frames. These are evoked through the lexical relations and we have listed them in the LUs' tables. With a possible future expansion of our resource, more terms and examples can be added.

Additionally, effort was made to align the EL and the EN terms. The unique EL terms are 190 and the unique EN terms 172, out of which there are 137 aligned sets of terms. This number shows us that starting with parallel corpora and following the above-mentioned procedures for two languages, we can end up with a bilingual term base which is useful in many ways, and especially for the translation process. With a future extension of the corpus and subsequent extraction of more terms, the aligned terms will increase, providing us with a term base with valuable semantic information.

## 5.3 Annotated Sentences-Examples

The annotated sentences are the third component of our resource. In total, the annotated corpus consists of 255 sentences (130 EL and 125 EN), which correspond to 6,923 words (3,832 EL and 3,091 EN). These sentences, which are the frames' examples bare two layers of linguistic annotation via UDPipe and WebAnno tools, and are available for use and further extension.

An example of annotated sentences with WebAnno is shown in Figure 1, where EN sentences of the frame Expenditures are annotated. Above each annotated word or phrase the frame to which it has been assigned and the FE appear. For example, the LU "operating expenses" of sentence 16 evokes the frame Expenditures and has been annotated with the FE "expenditure". The tags in colour make it easier to see the different FEs and may also be useful for discovering repeated patterns in a particular frame. Here, for example, in the frame Expenditures there seems to be a tendency of mentioning first the theme (expenditure), then the predicate and after the FEs *final state* and *final value*. The FE *cause* seems to be used at the end of the sentence.



Essentially, we have provided a semantically annotated corpus which is a useful resource for seeing the relations between the elements that comprise each sentence.

	expenditures.expenditure	expenditures.predicate	expenditures.final-state	expenditures.final-value	expenditures.cause	
16	Total operating expenses	reduced	6%	to £12,019m	reflecting savings from strategic cost programmes as well as lower litigation and conduct charges.	
	expenditures.expenditure	expenditures.predicate	expenditures.final-state	expenditures.final-value	expenditures.cause	
17	Operating costs	were	1 per cent lower than in the first quarter of 2016 at	£1,968 million	reflecting tight cost control and further benefits from the Simplification programme.	
	expenditures.expenditure	expenditures.predicate	expenditures.final-state	expenditures.final-value	expenditures.cause	
18	Restructuring costs	were	£157 million	(2016: £161 million) and comprised severance costs relating to the Simplification programme, the announced rationalisation of the non-branch property portfolio and the work on implementing the ring-fencing requirements.		
	expenditures.expenditure	expenditures.predicate	expenditures.final-value			
19	Analysis of loans and advances to customers at	amortised cost				
	expenditures.expenditure					
20	Credit impairment charges	improved	90%	to £3m	driven by lower impairment charges in Europe reflecting business exits.	
	expenditures.expenditure	expenditures.predicate	expenditures.final-state	expenditures.final-value	expenditures.cause	

Figure 1: The frame Expenditures (EN) in WebAnno (Yimam et al. 2013).

## 6 Discussion

The outcome of this work is a bilingual term base which can be used either in its present form or for the creation of an electronic data base. The term base can also be utilised for the assistance of the translation procedure, for example in software designed for this purpose. In fact, the possibility to use an annotated corpus like the one described in this paper breaks new ground for terminography for assisting translation.

One of the biggest challenges that we had to face during the alignment of the EL and EN frames was that of equivalence. The major problem of language and the primary concern of linguistics is “equivalence in difference” (Jacobson 1959, in Munday 2002: 71), meaning that, regardless of differences someone may deal with while translating, the total equivalence between source and target language must be ensured; and since translation from one language into another includes two equivalent messages in two different codes, equivalence in difference is the only acceptable form of equivalence (Kentrotis 1996: 283). An approach to terminography like the one of FBT aims at exactly this kind of equivalence.

During this effort, we needed to make alignments at two levels: first, at the frame level, and then at the LU level. Frame alignment was not problematic; however, there were some problems when we tried to align the EL and EN terms. Firstly, there is a number of pragmatic elements that is hard to translate, even if the concept exists in both languages. One example is the frame *Social contributions*, as it contains terms from two different financial systems. Additionally, the expressive meaning (Baker 1992: 23) of the terms might differ, like the EL term *κεφαλαιακοί περιορισμοί.n* (*capital controls.n*), which is more emotionally loaded than the English equivalent due to the Greek financial crisis.

If we look again at the LUs of the frame *Obtaining a loan* that is available in Table 4, we can see that most terms can be aligned, which proves the universal nature of specialised terminology and of the semantic frames. There are also cases where a translational equivalent exists, but it has not been found in our corpus – probably in a future expansion of the resource these terms could also be added.

A notable example of difference in the lexicalisation of concepts is the concept of *δανείζω.v* (give a loan) and *δανείζομαι.v* (take a loan). In Greek, the same verb is used in active and passive voice for expressing two different equivalent verbs in English: *lend.v* and *borrow.v*, respectively. In FrameNet these two concepts form two distinct frames which are called *Lending* and *Borrowing*. In our resource, however, we decided to combine them under one common frame (*Obtaining a loan*), because our aim was to gather all the terms that pertain to bank lending in one frame. This difference in lexicalisation can also lead to other issues, for example difficulty in annotation due to discrepancies among languages in the lexicalisation of concepts. The majority of the discrepancies, however, tend to be with verbs denoting financial-related events, rather than nouns, the latter being more technical and specific.

The present resource is of course smaller in size than the one of FrameNet, and as result the frames are not so extended, in a sense that they include only the FEs that are essential for covering the terms. In a possible future expansion of the resource, more frames and more FEs to the existing frames can be added, so that the resource can cover the biggest possible part of financial terminology.

It should be taken into account that the development methodology of our resource differs in a few fundamental aspects than the one of FrameNet. The most important one is that the frames were created based on the terminology that had to be accounted for; to put it another way, the frames, the FEs and their definitions were all created in order to express the terms in the most precise way possible. They were also viewed as a way to examine whether frames can be used in both languages. Also, it is worth mentioning that the aim of this work was not to compile a complete terminological resource, but rather to explore the methodology and the process for developing a bilingual frame-based resource of a specialized domain.

## 7 Conclusion

We have presented work which includes the development of a bilingual resource for the domain of finance which is based on Frame Semantics (Fillmore 1977, 1982, 1985). The outcome of our work is: (a) a bilingual lexical resource in electronic format which contains LUs in Greek and English (c. 560 terms, 740 LUs in total). The LUs are described both in terms of their semantic frames and through the listing of the lexical relations with which they are linked to each other; (b) a number of scenes and semantic frames for the semantic field of finance; in particular, the specialised vocabulary is organised around



9 scenes and 39 frames which are common for both languages; and (c) a fully annotated corpus composed of the sentences-examples in which the LUs are attested.

Future work has already been planned towards enriching the Greek component of FrameNet, as well as making comparisons between the Greek and English language. In particular, we participate in the Global FrameNet project, which is a joint effort to bring together FrameNets in different languages. Another possible future prospect is the expansion of our resource, which can be done in a number of ways. More data can be added to the corpus in order to extract more terms and find more examples for our frames. In this way, more FEs can also be added to the frames. From another perspective, the resource can be extended to other domains of language for specific purposes, as a way to examine whether the same principles would apply when following the above-described methods. The resource has been made freely available for research purposes via CLARIN-EL repository.

## 8 References

- Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L.J., Uppström, J. (2014). Swedish FrameNet++ The Beginning of the End and the End of the Beginning. In *Proceedings of the Fifth Swedish Language Technology Conference*, Uppsala, 13-14 November 2014.
- Baker, M. (1992). *In other words. A coursebook on translation*. New York: Routledge.
- Baker C. F., Fillmore C. J., Lowe J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics – Volume 1 (COLING '98)*, Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA: 86-90. Accessed at <https://dl.acm.org/citation.cfm?doid=980451.980860> [29/05/2020]
- Dalpanagioti, T. (2012). *A Frame-driven Approach to Bilingual Lexicography: Lexical Meaning and Usage Patterns in Greek and English Motion Verbs*. PhD Thesis. Εθνικό αρχείο διδακτορικών διατριβών. Accessed at <https://www.didaktorika.gr/eadd/handle/10442/27162> [29/05/2020]
- Faber, P. (2011). The dynamics of specialized knowledge representation: Simulational reconstruction or the perception–action interface. *Terminology* 17, no 1 (January 1): 9-29. Accessed at <http://dx.doi.org/10.1075/term.17.1.02fab> [29/05/2020]
- Faber, P. (2014). Frames as a Framework for Terminology. In book: *Handbook of Terminology*, Vol. 1, edited by H. J. Kockaert and F. Steurs. Amsterdam/Philadelphia: John Benjamins: 14-33.
- Fillmore, C. J. (1977). Scenes-and-frames Semantics. In *Linguistic Structures Processing. Fundamental Studies in Computer Science*, vol. 59. North Holland Publishing: 55-81.
- Fillmore, C. J. (1982). Frame Semantics. *Linguistics in the Morning Calm*, ed. The Linguistic Society of Korea. Selected papers from SICOL-1981: 111-137.
- Fillmore, C. J. (1885). Frames and the Semantics of Understanding. *Quaderni di semantica*. Vol. 6, no 2: 222-254.
- Giouli, V., Pilitsidou, V., Christopoulos, H. (2020). Greek within the Global FrameNet Initiative: Challenges and Conclusions so far. In *Proceedings of the International FrameNet Workshop 2020, Towards a Global, Multilingual FrameNet. LREC 2020 Workshop, Language Resources and Evaluation Conference*, 11–16 May 2020, Marseille, France. 48-55. Accessed at <https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/frameNet2020book.pdf> [31/07/2020]
- Gotsoulia, V., Desipri, E., Koutsombogera, M., Prokopidis, P., Papageorgiou, H., Markopoulos, G. (2007). Towards a Frame Semantics Lexical Resource for Greek. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*.
- Kentrotis, G. (1996). *Theory and Practice of Translation* (in Greek). Athens: Diavlos.
- Laurence, A. (2016). AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Accessed at <http://www.laurenceanthony.net/> [29/05/2020]
- Munday, J. (2002). *Introducing Translation Studies: Theories and Applications*. (in Greek). Athens: Metaixmio.
- Ohara, K. H. (2009). *Frame-based Contrastive Lexical Semantics in Japanese FrameNet: The Case of Risk and Kakeru*: 163–182. Berlin/New York: Mouton de Gruyter.
- Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins.
- Petruck, M. R. L. (1996). *Handbook of Pragmatics*. Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen (eds.). Philadelphia: John Benjamins.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. Berkeley, California: International Computer Science Institute.
- Salomão, M. M. M. (2009). Framenet brasil: um trabalho em progresso. *Caleidoscópio* 7(3): 171–182.
- Straka, M., Starková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics: 8899. Vancouver, Canada. Accessed at <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf> [29/05/2020]
- Subirats, C. (2009). *Spanish FrameNet: A Frame-semantic Analysis of the Spanish Lexicon*. Berlin/New York: Mouton de Gruyter.
- Tantos, A., Markantonatou, S., Anastasiadi-Symeonidi, A., Kyriakopoulou, T. (2015). *Computational Linguistics* (in Greek). [E-book] Athens: Hellenic Academic Libraries Link. Accessed at <https://repository.kallipos.gr/bitstream/11419/2205/13/nlp2ndEdition.pdf> [29/05/2020]
- Yimam, S. M., Iryna, G., de Castilho, R. E., Biemann, C. (2013). *WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations*. ACL.



**Acknowledgements**

The authors would like to thank the anonymous reviewers for their valuable comments to the manuscript that contributed to improving the final version of the paper. The research leading to the results presented here was conducted in the context of the Translation and Interpreting MA programme of the Faculty of Turkish and Modern Asian Studies of the National and Kapodistrian University of Athens. Also, they would like to specially thank Prof. E. Sella and Prof. I.E. Saridakis, for their support throughout this effort and their comments.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Phraseology and Collocation**







# Le Traitement des Proverbes dans les Dictionnaires Explicatifs Roumains du XIX<sup>e</sup> Siècle

Aldea M.

Université Babeş-Bolyai de Cluj-Napoca, Roumanie

## Résumé

Cette étude s'intéresse au traitement des proverbes dans deux dictionnaires explicatifs roumains du XIX<sup>e</sup> siècle, *Lexiconul de la Buda* [Lexicon de Buda] (LB<sup>e</sup>) et *Vocabularu romano-francesu* [Vocabulaire roumain-français] (VRF). En nous appuyant sur plusieurs exemples de proverbes puisés dans ces deux ouvrages lexicographiques, nous examinons la manière dont ces proverbes sont enregistrés dans la structure même des articles, la place occupée par le mot vedette dans la structure du proverbe, les variations lexicales et les éventuels changements de sens, de même que la circulation des proverbes au cours de presque un demi-siècle, leur typologie et les proverbes correspondants dans les autres langues enregistrées dans notre corpus, telles que le latin, le hongrois, l'allemand ou le français, et, en même temps, dans d'autres langues romanes. Au terme de cette analyse, nous pouvons constater la dynamique de la langue roumaine, sa tendance à employer des formules figées, des mots appartenant au lexique fondamental et aussi bien que des mots récents, des mots empruntés aux langues néolatines pour rendre ces proverbes.

**Mots-clés** : proverbe ; le Lexicon de Buda ; Ion Costinescu ; Vocabulaire roumain-français ; dictionnaires explicatifs roumains ; traitement lexicographique

## 1 Introduction

Ces dernières décennies, on observe un déplacement de l'attention des linguistes vers le domaine de la phraséologie en général et, au sein de celle-ci, sur l'analyse des unités phraséologiques qui constituent son objet – syntagmes, expressions, collocations, etc. (voir Cowie 1998 ; Moon 1998 ; Granger et Meunier 2008 ; Munteanu 2013, etc.). Parmi les éléments à structure fixe, les proverbes occupent une place privilégiée. L'abondance d'études consacrées aux proverbes envisagés à partir de différentes perspectives (folklorique, grammaticale, lexicale, pragmatique, traductologique et lexicographique) met en évidence leur importance (voir Tabarcea 1982 ; Negreanu 1983 ; Gheorghe 1986 ; Kleiber 1999 ; Tamba 2000 ; Visetti et Cadiot 2006 ; Kleiber 2010 ; Milică 2013 : 141-189 ; Bogdan Oprea 2019 ; Kleiber 2019 ; Čermák 2019 ; Vírva 2019 ; Pelea 2020 : 307-314 et *passim*, etc.). De toutes ces perspectives possibles, nous nous limitons ici à l'analyse de leur présence dans des lexicographiques. Différentes études plus ou moins étendues (voir Zgusta 1971 : 138-163 ; Svensén 1993 : 110-111, 207-209 ; Atkins et Rundell 2008 : 166-176 ; Marelllo 2010 : 1347-1354 ; Kosch 2016 : 145-161 ; Čermák 2019 ; Ştefănescu 2019 : 121-234, etc.) montrent la difficulté d'établir des frontières très précises concernant le traitement de ces unités phraséologiques au niveau propositionnel ou phrastique dans les dictionnaires. Considérés comme « des expressions sui-generis », « des ensembles de mots qu'ils doivent être entendu comme des ensembles [...] » (Zgusta 1971 : 153 ; nous traduisons), les proverbes sont des unités sémantiques qui ne peuvent pas être expliquées ou remplacées par un synonyme constitué d'un seul mot ; on peut le faire plutôt par une paraphrase ou par un ample commentaire explicatif : « leur principale caractéristique est de transmettre une certaine expérience accumulée, une remarque généralisée, ou de conseiller ; ils transmettent cette signification parfois d'une manière métaphorique » (*ibidem*, 151-152 ; n.t.) ; « [l']aspect le plus général et évident, ainsi que la particularité des proverbes, est leur base lexicale » (Čermák 2019 : 10 ; n.t.).

En nous appuyant sur ces observations, nous voudrions nous pencher dans la présente recherche sur le traitement des proverbes dans deux dictionnaires roumains du XIX<sup>e</sup> siècle, le *Lexicon de Buda* (Buda, 1825) et le *Vocabulaire roumain-français* (Bucarest, 1870) (désormais abrégés LB<sup>e</sup>, respectivement VRF). Malgré leur format plurilingue (voir le LB<sup>e</sup>), respectivement bilingue (voir le VRF), les deux ouvrages ont l'avantage d'être également des dictionnaires explicatifs. Parus au XIX<sup>e</sup> siècle, les deux dictionnaires marquent de véritables tournants dans la lexicographie roumaine (cf. Aldea 2016 ; Aldea 2018) : le LB<sup>e</sup> marque le début de la lexicographie roumaine moderne, tandis que le VRF est « le [p]remier et seul dictionnaire explicatif et général de la langue roumaine qui soit imprimé » (Seche 1966 : 73 ; n.t.) jusqu'à 1870, et dont la publication suit d'ailleurs de près la parution, en 1825, du LB<sup>e</sup>.

Avant de passer à l'analyse proprement-dite de notre corpus, il convient de faire quelques remarques concernant le mot « proverb » [proverbe].<sup>1</sup> Examinant les deux dictionnaires de notre corpus, nous avons constaté que les rédacteurs du LB<sup>e</sup> ne consignent pas le mot « proverb » [proverbe] dans une entrée à part, mais ils accordent un article distinct au mot « adagiu » [adage] défini par la série synonymique « zisă, pildă, parimie » [dicton, exemple, parémie], chacun de ces termes bénéficiant d'un traitement indépendant. Néanmoins, le rédacteur du VRF enregistre le mot « proverb » [proverbe] (< lat. *proverbium*, fr. *proverbe*) comme mot vedette et l'explique par le paradigme des formes sentencieuses considérées synonymes : « sentință, maximă, parimie, pildă comună și vulgară în puține cuvinte » [sentence, maxime, parémie,

<sup>1</sup> Toutes les traductions entre crochets sont de nous.



conseil commun et populaire en peu de mots] (cf. VRF, s.v.), chacune d'entre elles bénéficiant aussi d'une entrée distincte.

## 2 Analyse du Corpus

L'inventaire de ces unités figées enregistrées tant par le LB<sup>e</sup> que par le VRF a été établi selon le critère de la présence de la marque « proverbe ». Dans le cas du LB<sup>e</sup>, la sélection a été faite automatiquement grâce à l'édition électronique, tandis que, dans le cas du VRF, le dépouillement a été réalisé manuellement.

Le LB<sup>e</sup> enregistre 15 proverbes répartis entre 14 articles (voir les exemples de 1 à 14<sup>2</sup>) sur un total de plus de 12.000 articles (soit 771 pages), tandis que le VRF recense 129 proverbes répartis entre 113 articles (pour une sélection de ces proverbes – voir les exemples de 15 à 65) sur un total d'environ 28.000 articles (soit 1.332 pages).

Dans le LB<sup>e</sup>, la marque du domaine « proverb » [proverbe] est désignée par plusieurs abréviations : *prov.* (exemples de 1 à 7 et 13), *proverb.* (exemples 8, de 10 à 12 et 14) ou *provb.* (exemple 9), et elle introduit soit le proverbe roumain (exemples 13 et 14), soit le proverbe latin (exemples de 1 à 7 et de 9 à 12) ; parfois, elle est précisée après les équivalents dans les trois langues du dictionnaire (exemple 8).

Dans le VRF, la marque du domaine apparaît toujours avant proverbe et se présente aussi bien par le biais de structures du type « se zice proverbial / proverbialmente » [on dit proverbiallement] (exemples 29 et 36), « în proverbial » [proverbiallement] (exemple 37), « locuțiune proverbială » [locution proverbiale] (exemple 59), « în stil [...] proverbial » [en style [...] proverbial] (exemple 16), « expresiuni proverbiale » [expressions proverbiales] (exemple 61) que sous une forme abrégée comme : *pro.* (exemple 51), *proverb.* (exemple 63), *proverb.* (exemples 62 et 65), *prover.* (exemples 31, 49 et 60), *prv.* (exemples de 22 et 38) et *prov.* (exemples 18, 21, de 24 à 26, 30, de 32 à 35, 39, 41, 42, de 44 à 46, 48, 52 et de 55 à 58). Parfois, la marque « proverbe » est doublée par d'autres marques, telles que *fig.* « figurat » [figuré] (exemples 15, 28, 29, 40, 43, 50, 54 et 60), *iro.* « ironic » [ironique] (exemple 23), *pop.* « popular » [populaire] (exemple 53), *fam.* « familiar » [familier] (exemple 20) et un double marquage *fig. fam.* « figurat, familial » [figuré, familier] (exemple 27). Il convient de mentionner que, dans le LB<sup>e</sup>, nous avons pu identifier un seul article qui contient deux proverbes (exemple 9), tandis que, dans le VRF, il y a des nombreux articles qui contiennent deux ou plusieurs unités phraséologiques marquées comme proverbes (exemples 27, 30, 41, 54, etc.).

À l'exception du proverbe de l'exemple 8 enregistré comme deuxième sens de la seconde valeur du mot vedette (celle réfléchie), toutes les autres structures puisées dans le LB<sup>e</sup> (voir les exemples de 1 à 7 et de 9 à 14) sont placées dans la proximité du sens de base et mettent en évidence le sens fondamental du lemme dans le but de le clarifier, de le rendre plus compréhensible. Cependant, dans le VRF (voir les exemples de 15 à 65), les structures bénéficiant de la marque « proverbe » se trouvent soit au milieu de l'article, soit à sa fin et illustrent souvent la valeur figée ou métaphorique.

La lecture attentive de ces structures enregistrées par les deux ouvrages en tant que proverbes nous a permis de constater qu'un grand nombre de ces unités phraséologiques sont perçues de nos jours comme des expressions verbales – un seul article dans le LB<sup>e</sup> (exemple 4), plus de 50 articles dans le VRF (exemples 15, 16, 20, 28, 32, 41, 50, 52, 53, 61, 63 et 64) –, des syntagmes – 2 articles dans le LB<sup>e</sup> (exemples 2 et 3), 6 articles dans le VRF (exemples 43, 44, 62, etc.) et des locutions – 4 articles en VRF (exemples 22, 36, 55, etc.).

En nous appuyant sur ces nouvelles données, nous procédons ci-dessous à l'analyse proprement dite des proverbes répertoriés dans notre corpus en tenant compte de plusieurs aspects : la classe lexico-grammaticale et l'origine du mot vedette, ainsi que son emplacement dans la structure du proverbe, le changement de sens, les variations lexicales et sémantiques à l'intérieur d'un même proverbe, la typologie des proverbes, et leur correspondant dans d'autres langues.<sup>3</sup> En ce qui concerne la classe lexico-grammaticale du mot vedette, nous observons que la grande majorité appartiennent à la catégorie grammaticale du nom – 8 articles dans le LB<sup>e</sup> (exemples 1, 6, 7 et de 9 à 12) et 43 articles dans le VRF (exemples 18, 19, 25, de 30 à 32, 34, 35, 38, 42, 43, de 45 à 49, 53, 54, 56, 57, 60 et 65, etc.), à la catégorie de l'adjectif – un seul article dans le LB<sup>e</sup> (exemple 13) et 7 articles dans le VRF (exemples 17, 26, 33, 51, etc.), et à celle du verbe – 2 articles dans le LB<sup>e</sup> (exemples 5 et 8) – ou du verbe substantivé – 11 articles en VRF (exemples 21, 24, 27, 37, 39, 58, etc.).

Dans la structure du proverbe, nous observons que la base lexicale, le mot vedette, est placée soit au début (exemples 9, 11, 12, 14 dans le LB<sup>e</sup>, respectivement les exemples 27, 31, 34, 39, 42, 45, 48, 54, 65, etc. dans le VRF), soit au milieu du proverbe (exemples 1, 7, 13 et 8 dans le LB<sup>e</sup>, respectivement les exemples 16, 19, 21, 27, 30, 32, 33, 35, 37, 43, 54, 56, 57, 58, etc. dans le VRF), soit à la fin (exemples 5 et 6 dans le LB<sup>e</sup>, respectivement les exemples 17, 18, 24, 25, 26, 38, 42, 46, 47, 51, 60, etc. dans le VRF).

Les différents aspects qui ont attiré notre attention concernent le changement de sens et les variations lexicales et sémantiques à l'intérieur d'un même proverbe. L'examen du corpus nous a permis d'observer que, dans de nombreux cas, le sens des proverbes est expliqué. Dans le LB<sup>e</sup>, nous avons identifié un seul exemple de ce type : « una (o) rândunea nu face primăvară [...] însemnează că o întâmplare nu face regulă » « une hirondelle ne fait pas le printemps »<sup>4</sup> [...] [signifie qu'un événement ne fait pas la règle] (exemple 14), où le commentaire est introduit par le verbe métalinguistique « însemnează » [signifie]. Pour justifier l'absence d'une glose dans les autres proverbes du LB<sup>e</sup>, nous pouvons avancer comme possible explication le fait que la signification est probablement connue en général par les lecteurs / locuteurs,

<sup>2</sup> Tous les exemples auxquels nous renvoyons se trouvent dans l'Annexe.

<sup>3</sup> L'objectif de ce travail est d'offrir une perspective globale sur la présence et le traitement des proverbes dans deux dictionnaires explicatifs roumains du XIX<sup>e</sup> siècle. Dans un prochain ouvrage, partant du modèle exposé dans les études de Conenna (1988 ; 2000) et Tabarcea (1982), nous nous proposons de nous pencher sur l'analyse syntaxique des proverbes.

<sup>4</sup> Toutes les traductions placées entre guillemets anglais sont de celles données aussi bien par les rédacteurs dans le LB<sup>e</sup> que par Ion Costinescu dans le VRF.



donc l'explication n'est plus nécessaire.<sup>5</sup> Pourtant, dans le VRF, à l'exception de quelques articles dans lesquels le proverbe n'est pas expliqué (exemples 21, de 24 à 26, etc.), le sens global, général, du proverbe est défini.

En outre, nous avons répertorié le même proverbe roumain dans deux ou plusieurs articles.

Ainsi, dans le LB<sup>e</sup>, le proverbe « o oaie răioasă împle toată turma » [une brebis galeuse gâte tout le troupeau] apparaît aussi bien sous l'entrée « Ôe » [Brebis] (exemple 11), que sous l'entrée « Răiosu » [Galeux] (exemple 13), mais avec des différences d'ordre graphique (ôe / ôue [brebis] ; răiosă / rōiosă [galeuse]), un ajout lexical, le numéral « una » [une] et la présence de la forme étymologique du verbe « a umple » [gâter]. De plus, si, dans l'exemple 13, le proverbe roumain trouve ses équivalents dans les trois autres langues du LB<sup>e</sup>, dans l'exemple 11, le correspondant hongrois du proverbe n'est pas donné.

Par contre, dans le VRF, les exemples sont plus nombreux et nous allons les présenter ci-dessous.

« Nu face haina pe călugăr » [l'habit ne fait pas le moine] : ce même proverbe est répertorié dans trois articles : « Călugăru » [Moine] (exemple 18), « Facere » [Faire] (exemple 27), « Haină » [Habit] (exemple 35) ; la seule différence consiste dans la manière de présenter la description sémantique, la définition. Bien que la signification détachée soit la même, le commentaire du sens varie par rapport aux moyens lexicaux employés : « nu trebuie să judeci pe cineva după cele din afară » [il ne faut pas juger quelqu'un selon son aspect] (exemple 18), « să nu judecăm pe oameni după aparință și nu este sânt cel ce se arată a fi » [qu'on ne juge pas les gens selon leur aspect et il n'est pas saint celui qui fait semblant de l'être] (exemple 27) et « să nu considerăm persoana după aparinție, după cele de afară » [ne jugeons pas une personne selon son aspect, d'après ce qui se voit à l'extérieur] (exemple 35).

La structure grammaticale « V + S + COD » présente dans l'exemple ci-dessus se retrouve dans le proverbe « nu face haina pe medic » « la robe ne fait pas le médecin » enregistré dans l'entrée « Medic » [Médecin] (exemple 42), mais le choix lexical porte sur l'expression du COD, le mot « călugăr » [moine] étant substitué par le mot « medic » [médecin], ce qui donne lieu à un autre sens, complètement différent : « nu este titlu care face pe om învățat » [il n'y a pas de qualification qui rende l'homme érudit].

Comparant les deux proverbes ci-dessous, « apa curge, pietrele rămân » [l'eau coule, les pierres restent] (exemple 54) et « gărla curge, pietrele rămân » « le torrent s'écoule, la pierre reste » (exemple 34), on observe tout d'abord que le jeu lexical concerne les noms « apa » [l'eau] et « gărla » [le ruisseau, le torrent]. Ensuite, la valeur sémantique des deux proverbes bénéficie d'explications très amples : « nu spera sau nu crede în străin mai mult decât în pământean, că el fuge la timp de nevoie, dar noi rămânem ; sau năimutul dispăre, turma rămâne » [ne fais pas davantage confiance à un étranger qu'à ton concitoyen, car l'étranger s'enfuit quand les temps sont rudes, tandis que nous restons ; ou le serviteur disparaît, le troupeau reste] (exemple 34), respectivement « cu cei ce trec nu se poate cineva asocia or profita, ca cu cei ce sunt ai locului ; de la străin nu se poate aștepta un bine ca de la cei cu care trăiește cineva ; străinul nu te poate ajuta ca pământeanul » [on ne peut pas s'associer ou avoir du profit avec quelqu'un comme on le fait avec les nôtres ; on ne peut pas attendre un bienfait de la part de l'étranger comme on en attend de la part de ceux avec lesquels on vit ; l'étranger ne peut pas t'aider comme t'aide le tien] (exemple 54). On remarque également l'absence de l'équivalent français dans l'exemple 54.

Pour les trois proverbes qui suivent – « omul propune, Dumnezeu dispune » [l'homme propose, Dieu dispose] (exemple 24), « omul propune și Dumnezeu dispune » « l'homme propose et le Dieu dispose » (exemple 58) et « nu după voia omului, ci după voia Domnului » « l'homme propose et Dieu dispose » (exemple 25) – on indique la même signification. L'examen de l'expression graphique des exemples 24 et 58 nous permet d'observer l'emploi de moyens différents pour focaliser l'attention sur les deux séquences qui configurent le proverbe : dans l'exemple 24, les deux parties sont juxtaposées à l'aide de la virgule, tandis que, dans l'exemple 58, la virgule est remplacée par la conjonction coordinatrice « și » [et]. Par contre, dans l'exemple 25, nous constatons que la valeur sémantique est rendue par d'autres outils lexicaux, comme le montre cette traduction littérale « rien ne se passe selon la volonté de l'homme, mais tout se passe selon la volonté de Dieu » (n.t.). De plus, dans l'exemple 24, le correspondant français n'est pas enregistré, tandis que, dans les exemples 58 et 25, bien que la forme roumaine du proverbe change, le même équivalent français est consigné. Nous observons aussi que seulement dans l'exemple 58 le proverbe bénéficie d'une ample explication du sens : « omul chibzuiește și Dumnezeu otărăște ; nu este cum va omul, ci cum va Domnul » [l'homme réfléchit et Dieu décide ; les choses ne se passent pas selon la volonté de l'homme, mais selon la volonté de Dieu].

Les unités polylexicales « vocea popului este vocea lui Dumnezeu » « la voix du peuple est la voix de Dieu » (exemple 56) et « vocea popului, vocea lui Dumnezeu » « la voix du peuple est la voix de Dieu » (exemple 65) sont deux proverbes qui ont le même sémantisme. La différence consiste dans la manière dont ils peuvent être rendus à l'écrit. Si, dans l'exemple 56, on a une équation du type « x est y », c.-à.-d. une structure métaphorique d'identification, dans l'exemple 65, suite à la suppression du verbe « être » et à son remplacement par une virgule, la métaphore devient « révélatrice ». On observe également que les deux articles contiennent des explications détaillées du sens du proverbe : « cu un înțele mai obicinuit, simțământul general este fondat pe veritate » [plus communément, le raisonnement général est fondé sur la vérité] (exemple 56), respectivement « când toată lumea se învoiește a crede generalmente un lucru, cată a se crede că toată acea lume este pe calea rațiunii » [quand tout le monde se met d'accord à croire une chose, il faut penser que tous ces gens suivent le chemin de la raison] (exemple 65).

Les entrées « Țeară » [Pays] et « Profetu » [Prophète] contiennent elles aussi ce proverbe, mais avec une petite différence : « nimeni nu este profet în țara lui » « nul n'est prophète dans son pays » (exemple 60) et « nimeni nu poate fi profet în țara lui » [nul ne peut être prophète en son pays] (exemple 57). Dans l'exemple 60, entre le pronom négatif sujet

<sup>5</sup> Bien que « la mission d'un dictionnaire soit de refléter l'usage », l'analyse du corps des articles met en évidence le manque du caractère systémique dans le LB<sup>e</sup>, absence qui s'explique par la longue durée de l'élaboration (plus de 30 ans) et l'implication de plusieurs érudits transylvaniens dans sa rédaction (S. Mîcu, P. Maior, I. Teodorovici et d'autres).



et l'attribut du nom, il y a le présent de l'indicatif du verbe « être » à la forme négative, à valeur déclarative, tandis que, dans l'exemple 57, le présent de l'indicatif du verbe « être » à la forme négative est remplacé par le présent de l'indicatif du verbe de modalité « a putea » [pouvoir] suivi par l'infinitif du verbe « a fi » [être], déclenchant ainsi une connotation supplémentaire – l'impossibilité. On constate que les deux proverbes sont accompagnés par l'explication de leur sens : « mai puțin considerat este cineva în locul său decât aiure » [on est moins apprécié chez soi qu'ailleurs] (exemple 57), respectivement « este prea dificil a-și face cineva reputația în țara lui, ca la străini » [il est plus difficile de construire sa réputation chez soi qu'à l'étranger] (exemple 60), mais seul l'exemple 60 contient le correspondant français aussi.

Les proverbes « finele coronează lucrul » « la fin couronne l'œuvre » (exemple 31) et « sfârșitul încoronează lucrul » « la fin couronne l'œuvre » (exemple 38) véhiculent le même sens, même s'ils emploient des moyens lexicaux différents, plus exactement, soit des mots différents appartenant à la même série synonymique (« sfârșitul », « finele » [la fin], dont le dernier est aperçu comme néologique), soit la forme littéraire « încoronează », d'un côté, et la variante obsolète « coronează » du même verbe « couronner », d'autre côté. En ce qui concerne l'explication fournie, nous observons que le proverbe de l'exemple 31 est expliqué par une structure similaire à celle du proverbe de l'exemple 38, c.-à.-d. « sfârșitul încoronează fapta » [la fin couronne l'œuvre] (exemple 31), tandis que le proverbe de l'exemple 38 bénéficie d'un énoncé détaillé : « nu e destul începutul bun, trebuie a fi și sfârșitul bun » [il ne suffit pas d'un bon début, encore faudra-t-il que la fin soit bonne] (exemple 38).

Dans deux entrées distinctes, « Caldu » [Chaud] et « Fieru » [Fer], apparaît le même proverbe roumain sans subir aucune modification : « bate fierul până este cald » rendu en français par « il faut battre le fer pendant qu'il est chaud » (exemple 17) et « il faut battre le fer tandis qu'il est chaud » (exemple 30). Pourtant, l'examen du correspondant français met en évidence des choix différents pour rendre la préposition « până » [jusqu'à ; tant], qui, dans ce cas, fait partie de la locution à valeur de conjonction temporelle « până când » qui peut être transposée en français par les locutions conjonctives « jusqu'à ce que », « tant que », « pendant que » ou « tandis que ».

Les proverbes consignés dans les entrées « Pătră, Piétră » [Pierre] et « Mușchiu » [Mousse], d'un part, « Miere » [Miel] et « Oțet » [Vinaigre], d'autre part – « picatra ce se rostogolește nu prinde mușchi » « pierre qui roule n'amasse pas de mousse » (exemple 46), « peatra ce rostogolește nu prinde mușchi » « pierre qui roule n'amasse point de mousse » (exemple 54), « mai multe muște se prind cu o lingură de miere decât cu o butie de oțet » « on attrape plus des mouches avec du miel qu'avec du vinaigre » (exemple 43) et « mai multe muște se prind cu o lingură de miere decât cu o bute de oțet » « on prend plus de mouches avec un peu de miel qu'avec un tonneau de vinaigre » (exemple 47) –, se ressemblent d'un point de vue formel. Ainsi, les deux premiers se distinguent aussi bien par un trait d'ordre graphique et implicitement phonétique (l'alternance vocalique [é=ea] / [ié=iea] dans le nom « piatra » [la pierre]) que par l'absence du pronom réfléchi « se » [se] de la structure du verbe pronominal dans l'exemple 54 ; tous les deux bénéficient d'un commentaire du sens : « cel ce schimbă des locul sau profesiunea nu se-mbogățește » [la personne qui change souvent de place ou de profession ne s'enrichit pas] (exemple 46), « cel ce schimbă meseria nu face stare » [la personne qui change son métier ne s'enrichit pas] (exemple 54). Dans le cas des deux autres proverbes, nous observons d'abord la présence de l'explication de leur sens : « mai lesne-și ajunge cineva scopul cu blândețe decât cu rigoare și mânie » [une personne atteint plus facilement ses fins par la bonté que par la sévérité et la colère] (exemple 43), respectivement « cu o manieră dulce reușește cineva mai bine decât cu o manieră de mândrie sau de asprime » [on obtient davantage par la douceur que par un comportement orgueilleux ou âpre] (exemple 47). Nous observons ensuite l'emploi aussi bien de la forme littéraire « bute » [tonneau] dans l'exemple 47 que de la forme régionale « butie » [tonneau] dans l'exemple 43. Pour ce qui est de l'équivalent français, on remarque le choix du rédacteur de rendre le deuxième élément de la négation soit par « point » dans l'exemple 54, soit par « pas » dans l'exemple 46, tandis que, pour les exemples 43 et 47, les choix lexicaux du rédacteur oscillent entre les verbes « prendre » (exemple 47) et « attraper » (exemple 43), d'un part, et entre l'emploi de l'article partitif (exemple 43) ou de l'adverbe « un peu de » (exemple 47), d'autre part. On note également, l'absence du nom « tonneau » dans le correspondant français de l'exemple 43.

Les derniers proverbes que nous analyserons sont : « greșeala confesată este pe jumătate iertată » « une faute confessée est à demi pardonnée » (exemple 21) et « păcatul mărturisit este pe jumătate iertat » « une faute confessée est à demi pardonnée » (exemple 26). Ces deux proverbes sont traduits en français par la même expression et ne bénéficient pas d'une explication. Même si le rédacteur choisit de les rendre par des mots de la même série synonymique – d'une part, des noms, « greșeala » [la faute] (exemple 21) et « păcatul » [le péché] (exemple 26), et, d'autre part, des adjectifs « confesată » [confessée] (exemple 21) et « mărturisit » [avoué] (exemple 26) –, il convient de noter que le second proverbe a, de nos jours, une connotation religieuse.

Et la liste pourrait bel et bien continuer.

Du point de vue sémantique, il faut également mentionner que trois proverbes apparaissent dans les deux dictionnaires (voir les exemples 1 et 19, 6 et 32, 8 et 39). L'examen de la première paire – « mi-i mai aproape cămeșa decât țintra » [la chemise m'est plus proche que la bure] (exemple 1) et « mai aproape mi-e cămeșa decât mantaua » « notre peau nous est plus près que notre chemise » ou, dans une traduction littérale, « la chemise m'est plus proche que le manteau » (n.t.) (exemple 19) – permet d'observer que la variation lexicale porte sur les noms « țintra » [la bure] (exemple 1) et « mantaua » [le manteau] (exemple 19). Bien que tous les deux appartiennent à la sphère sémantique des vêtements, le registre d'emploi est différent : « țintra » [bure] désigne une sorte d'habit spécifique au monde paysan et est un mot régional de Transilvanie et de Banat, tandis que le mot « mantauă » [manteau], considéré de nos jours comme une variante régionale du mot « manta » [manteau], s'utilisait plutôt dans le milieu urbain. Nous remarquons aussi que, dans l'exemple 1, grâce à la structure initiale, « pronom personnel au datif + verbe au présent de l'indicatif + adverbe de mode au comparatif de supériorité », l'attention est focalisée sur l'actant, tandis que, dans l'exemple 19, la structure est renversée, la position initiale étant occupée par « l'adverbe de mode au comparatif de supériorité + pronom personnel au datif + verbe au présent de l'indicatif », ce qui conduit à une focalisation spatiale.



L'analyse de la deuxième paire – « de unde nu e foc, nu iese fum » [là où il n'y a point de feu, il n'y a point de fumée] (exemple 6) et « nu e foc fără fum, sau nu iese fum de unde nu este foc » « il n'y a point de feu sans fumée, ou il n'y a pas de fumée sans feu » (exemple 32) – indique la présence de deux variantes du proverbe de l'exemple 32 : la première variante est une seule proposition « nu e foc fără fum » [il n'y a point de feu sans fumée], tandis que l'autre est la réflexion d'une phrase, proposition principale + proposition subordonnée circonstancielle de lieu « nu iese fum de unde nu este foc » [il n'y a pas de fumée là où il n'y a pas de feu]. Cette dernière variante est repérable dans le proverbe de l'exemple 6, mais l'ordre des mots est renversé : proposition subordonnée circonstancielle de lieu + proposition principale.

La comparaison des éléments de la troisième paire – « trebuie să ne întindem pre cât ne ajunge țolul, cerga » [on doit s'étendre tant que le permet la couverture, le plaid] (exemple 8) et « lungește picioarele pe cât îți este pătura » [étends tes pieds tant que te permet la couverture] (exemple 39) – nous permet de noter la présence dans l'exemple 8 de la série synonymique « țolul » [la couverture] et « cerga » [le plaid], des mots employés surtout dans le patois, tandis que, dans l'exemple 39, on retrouve un seul mot généralement connu « pătura » [la couverture]. Si, par l'incipit de l'exemple 8, on exprime une contrainte, une obligation « trebuie să ne întindem » [on doit s'étendre], par celui de l'exemple 39, on formule un ordre, « lungește picioarele » [étends tes pieds] ; de plus, les deux verbes « a întinde » et « a lungi » [étendre] appartiennent à la même série synonymique.

Un autre aspect qui a retenu notre attention concerne l'origine du mot vedette dans l'article duquel le proverbe est enregistré. Aussi bien dans le cas du LB<sup>e</sup> que du VRF, nous avons observé que la plupart des entrées sont héritées du latin : rou. *cămașă* [chemise] < lat. *camisia* (exemples 1 et 19) ; rou. *fum* [fumée] < lat. *fumus* (exemple 6) ; rou. *întinde* [étendre] < lat. *intendo*, -ere (exemple 8) ; rou. *lup* [loup] < lat. *lupus* (exemple 9) ; rou. *mână* [main] < lat. *manus* (exemple 10) ; rou. *oaie* [brebis] < lat. *ovem* (exemple 11) ; rou. *rândunea* [hirondelle] < lat. *hirundinem* (exemple 14) ; rou. *barbă* [barbe] < lat. *barba* (exemple 16) ; rou. *cald* [chaud] < lat. *caldus* (exemple 17) ; rou. *dispunere* [disposer] < lat. *disponere* (exemple 24) ; rou. *domn* [seigneur] < lat. *dominus* (exemple 25) ; rou. *iertat* [pardonné] < lat. *libertare* (exemple 26) ; rou. *facere* [faire] < lat. *facere* (exemple 27) ; rou. *fier* [fer] < lat. *ferrum* (exemple 30) ; rou. *foc* [feu] < lat. *focus* (exemple 32) ; rou. *lucire* [luire] < lat. *lucire* (exemple 37) ; rou. *lucru* [travail] < lat. *lucrum* (exemple 38) ; rou. *miere* [miel] < lat. *\*melem* (exemple 43) ; rou. *mur* [mur] < lat. *murus* (exemple 45) ; rou. *mușchiu* [mousse] < lat. *\*musculus* (exemple 46) ; rou. *oală* [pot] < lat. *olla* (exemple 48) ; rou. *pământ* [terre] < lat. *pavimentum* (exemple 49) ; rou. *pește* [poisson] < lat. *piscem* (exemple 53) ; rou. *piatră* [pierre] < lat. *petra* (exemple 54) ; rou. *țară* [pays] < lat. *terra* (exemple 60) ; rou. *voce* [voix] < lat. *vocem* (exemple 65), etc.

D'autres entrées sont des créations roumaines ayant une base latine : rou. *râios* [galeux] < rou. *râie* (< lat. *aranea*) + suf. -os (exemple 13) ; rou. *omenie* [humanité] < rou. *omen* (< lat. *homines*) + suf. -ie (exemple 12) ; rou. *lungire* [étendre] < rou. *lung* (< lat. *longus*) + suf. -ire (exemple 39) ; rou. *păcătos* [pêcheur] < rou. *păcat* (< lat. *peccatum*) + suf. -os (exemple 51), etc.

Un nombre considérable d'entrées sont des emprunts à différentes langues :

- au vieux slave : rou. *oțet* [vinaigre] < sl. *ocitŭ* (exemple 47) ;
- au bulgare : rou. *gârlă* [ruisseau] < bg. *gârlo* (exemple 34) ;
- au serbe : rou. *haină* [habit] < sb. *haljina* (exemple 35) ;
- au grec moyen : rou. *călugăr* [moine] < mgr. *καλόγερος* (peut-être par la voie du sl. *kalugerŭ*) (exemple 18) ; rou. *mărgăritar* [perle] < ngr. *μαργαριτάρι* (exemple 41) ;
- à l'italien : rou. *populu* [peuple] < it. *populu* (exemple 56) ; rou. *medic* [médecin] < it. *medico* (exemple 42) ;
- au français : rou. *confesare* < fr. *confesser* (exemple 21) ; rou. *general* < fr. *général* (exemple 33), etc.

Certains mots vedettes ont une origine incertaine : rou. *groapă* [trou], cf. alb. *gropë* (exemple 7), ou ont pénétré en roumain par une voie multiple : rou. *fine* < it. *fine*, fr. *fin* (exemple 31) ; rou. *profet* < lat. *propheta*, fr. *prophète* (exemple 57) ; rou. *propunere* < lat. *proposare*, fr. *proposer* (exemple 58), etc.

Les sphères sémantiques couvertes par les mots vedettes renvoient à l'homme, aux vêtements, aux animaux, aux parties du corps humain et du corps animal, aux notions abstraites, à la configuration du terrain, aux actions, aux états, aux qualités, etc. Pourtant, les sphères sémantiques configurées par l'ensemble des proverbes renvoient à la moralité, aux bienfaits, à l'amitié, à la confiance, à l'appréciation, à la volonté, etc., donc aux vérités d'ordre moral.

Un autre aspect mis en évidence par notre corpus est l'universalité des proverbes. En examinant les deux dictionnaires, nous avons observé que, dans le cas du LB<sup>e</sup>, 7 de ces proverbes sont rendus en latin, en hongrois et en allemand (voir les exemples 1, 6, 7, 8, 10, 12 et 13), tandis que 3 proverbes (voir les exemples 5, 9 et 11) ne contiennent pas le correspondant en hongrois et un seul proverbe (exemple 14) n'a pas de correspondant allemand mentionné. Dans le cas du dernier proverbe, il convient de souligner la présence des correspondants italien et français. En ce qui concerne le VRF, à l'exception de quelques articles dont le correspondant français est absent (voir les exemples 24, 33, 39, 48, 54, 57, etc.), pour tous les autres proverbes, un équivalent français est donné.

En fait, tous ces correspondants démontrent l'universalité des valeurs humaines, des valeurs morales. Nous pouvons le voir dans le cas de quelques proverbes présents dans notre corpus et de leurs équivalents dans d'autres langues romanes :

- « bate fierul până este cald » (voir les exemples 17 et 30) : « fr. il faut battre le fer pendant qu'il est chaud ; it. batti il ferro quando è caldo ; port. quando o ferro estiver acendido, então é que há de ser batido ; esp. cuando el hierro está encendido entonces ha de ser batido » (cf. Gheorghe 1986 : 104) ;
- « mi-i mai aproape cămeșă decât țăndra » (voir les exemples 1 et 19) : « fr. ma chemise me touche de plus près que mon habit ; it. tocca più la camicia che il giubbone ; port. sinto mais e é-me mais precisa a pele que a camisa ; esp. más cerca está la camisa que el sayo » (cf. Gheorghe 1986 : 119-120) ;



- « care sapă groapă altuia cade însuși într-însa » (voir l'exemple 7) : « fr. tel qui creuse une fosse à un autre, y tombe souvent lui-même ; it. chi scava la fossa agli altri, vi cade dentro egli stesso ; port. quem para os outros abre buraco, nele cai ; esp. quien lazo me armó en el cayó » (cf. Gheorghe 1986 : 144) ;
- « trebuie să ne întindem pre cât ne ajunge țolul, cerga » (voir les exemples 8 et 39) : « fr. il faut étendre ses pieds selon ses draps ; it. bisogna stendersi quanto il lenzuolo è lungo ; port. cada qual estende a perna até onde tem coberta ; esp. cada uno extiende la pierna como tiene la cubierta » (cf. Gheorghe 1986 : 167) ;
- « nu face haina pe călugăr » (voir les exemples 18, 27 et 35) : « fr. l'habit ne fait pas le moine ; it. l'abito non fa il monaco ; port. o hábito não faz o monge, mas fã-lo parecer de longe ; esp. el hábito no hace al monje » (cf. Gheorghe 1986 : 218) ;
- « de unde nu e foc, nu iese fum » (voir les exemples 6 et 32) : « fr. il n'est point de fumée sans feu ; it. dove si fa fuoco nasce del fumo ; port. donde fogo não há, fumo se não levanta ; esp. donde fuego no ha, humo no sal » (cf. Gheorghe 1986 : 311).

Et la liste pourrait continuer.

### 3 Conclusion

Cette analyse comparative nous a permis de constater que, statistiquement, il y a une forte différence entre les deux dictionnaires, différence repérable aussi bien au niveau de la nomenclature (plus de 12.000 articles en 771 pages dans le LB<sup>e</sup>, respectivement environ 28.000 articles en 1.332 pages dans le cas du VRF) qu'au niveau de la présence des unités phraséologiques avec le marquage « proverbe » : 0,12% dans le LB<sup>e</sup> (soit 15 unités phraséologiques en 14 articles) et 0,46% dans le VRF (soit 129 unités phraséologiques en 113 articles). Malgré la présence de la marque « proverbe », dans les deux dictionnaires, les rédacteurs ont consigné sous ce marquage des expressions idiomatiques, des syntagmes et des collocations : 20% (soit 3 unités phraséologiques en 3 articles) dans le LB<sup>e</sup>, respectivement 51% (soit plus de 65 unités phraséologiques en 60 articles) dans le VRF. Les proverbes proprement-dits de notre corpus sont enregistrés dans le LB<sup>e</sup> sous des mots vedettes appartenant à la catégorie grammaticale du nom (8 articles, soit 72,72%), suivie par celle du verbe (2 articles, soit 18,18%) et de l'adjectif (1 article, soit 9,09%), tandis que, dans le VRF, les mots vedettes renvoient tour à tour aux catégories grammaticales du nom (43 articles, soit 70,49%), du verbe (la forme longue de l'infinitif) (11 articles, soit 18,03%) et de l'adjectif (7 articles, soit 11,47%) et ils sont placés dans la proximité du sens de base. Ils fonctionnent comme un exemple dans le cas du LB<sup>e</sup>, pendant que, dans le VRF, s'ils y fonctionnent comme des exemples aussi, ils servent cette fois-ci à illustrer des sens dérivés.

Un seul proverbe est accompagné par l'explication de son sens et par les équivalents en latin, en italien, en français et en hongrois, 7 proverbes (soit 63,63%) enregistrés du LB<sup>e</sup> apparaissent accompagnés seulement par leurs correspondants latin, hongrois et allemand, et 3 proverbes (soit 27,27%) sont accompagnés seulement par leur correspondants latin et allemand. Par contre, dans le VRF, 15 articles (soit 24,59%) contiennent des proverbes accompagnés par leur correspondant français, 10 articles (soit 16,39%) – des proverbes bénéficiant d'une explication et 5 articles (soit 8,19%) – des proverbes sans explication ni correspondant français. Tous les autres (31 articles, soit 50,81%) contiennent à la fois une explication et un correspondant français des proverbes.

Aussi bien dans le LB<sup>e</sup> que dans le VRF, nous avons identifié des situations où le même proverbe apparaît dans deux ou trois articles, ainsi que des variations lexicales et / ou syntaxiques au sein du même proverbe ; nous avons remarqué aussi des proverbes dont le sens est rendu par des expressions distinctes dans chacune des entrées.

Pour ce qui est de leur circulation, nous constatons une dynamique plutôt faible ; ainsi, nous avons identifié trois proverbes qui se trouvent à la fois dans le LB<sup>e</sup> et dans le VRF, le premier sous le même mot vedette, les autres sous deux mots vedettes différents.

Tous ces éléments nous permettent de tirer la conclusion qu'au XIX<sup>e</sup> siècle la présence des proverbes dans des dictionnaires de langue joue un double rôle : d'une part, un rôle culturel, car, à travers les siècles, ils représentent les témoignages d'un système de valeurs, d'une mentalité, de certaines normes morales et sociales, etc., et, d'autre part, un rôle didactique et pédagogique, car, fonctionnant à titre d'exemples, ils servent à l'apprentissage de la langue, à l'acquisition et à l'assimilation de nouvelles structures et de significations appartenant tant à la langue source qu'à la langue cible.

### 4 Bibliographie

- Aldea, M. (2018). Lexicographie et Terminologie au XIX<sup>e</sup> Siècle : Vocabularu Romano-Francesu [Vocabulaire Roumain-Français] de Ion Costinescu (1870). In J. Cibej et al. (eds.) Proceedings of the XVIII EURALEX International Congress : Lexicography in Global Contexts. Ljubljana : Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press Faculty of Arts, pp. 789-798. Consulté sur <https://euralex.org> [24.01.2020].
- Aldea, M. (2016). Un Proiect Accompli : le Lexicon de Buda (1825) en Édition Électronique. In T. Margalitadze, G. Meladze (eds.) Proceedings of the XVII EURALEX International Congress. Tbilisi : Ivane Javakhishvili Tbilisi University Press, pp. 856-862. Consulté sur <https://euralex.org> [24.01.2020].
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford : Oxford University Press.
- Bogdan Oprea, H. (2019). Elipsa paremiologică – implicații sintactice, lexico-semantice și stilistice în variația proverbelor din limba română. In C. Ușurelu (ed.) Studii lingvistice. In memoriam Cristina Călărașu. București : Editura Universității din București, pp. 33-77.



- Čermák, F. (2019). *Lexical and Semantic Aspects of Proverbs*. Reviewed by Hana Bouzková, Bohdana Divišová. Charles University in Prague : Karolinum Press.
- Conenna, M. (1988). Sur un lexique-grammaire comparé de proverbes. In *Langages*, no 90, pp. 99-116. Consulté sur [www.persee.fr](http://www.persee.fr) [03.07.2020].
- Conenna, M. (2000). Structure syntaxique des proverbes français et italiens. In *Langages*, no 139, pp. 27-38.
- Costinescu, I. (1870). *Vocabularu romano-francesu, lucratu dupe Dicționarulu Academiei Francese, dupe alu lui Napoleone Landais și alte Dicționare latine, italiene, etc.* Vol. I-II. Bucuresci : Tipographia Națională Antreprenor C.N. Rădulescu.
- Cowie, A.P. (ed.) (1998). *Phraseology : Theory, Analysis and Applications*. Oxford : Clarendon Press.
- Gheorghe, G. (1986). *Proverbele românești și proverbele lumii romanice. Studiu comparativ*. București : Albatros.
- Granger, S., Meunier, F. (eds.) (2008). *Phraseology. An Interdisciplinary Perspective*. Amsterdam : John Benjamins Publishing Company.
- Kleiber, G. (1999). Les proverbes : des dénominations d'un type « très très spécial ». In *Langue française*, no 123, pp. 52-69.
- Kleiber, G. (2010). Proverbes : transparence et opacité. In *Meta*, vol. 55, issue 1, pp. 136-146. DOI <https://doi.org/10.7202/039608ar> [24.01.2020].
- Kleiber, G. (2019). Une métaphore suit-elle toujours le même chemin ? Analyse des expressions idiomatiques et des proverbes métaphoriques. In *Langue française*, no 204, pp. 87-100.
- Kosch, I.M. (2016). Lemmatisation of Fixed Expressions : The Case of Proverbs in Northern Sotho. In *Lexikos*, 26, pp. 145-161. Consulté sur <http://lexikos.journals.ac.za> [24.01.2020].
- Langages*, no 90. Consulté sur [www.persee.fr](http://www.persee.fr) [03.07.2020].
- LB<sup>c</sup>. Pour l'édition électronique de M. Aldea (ed.) (2013). *Lesicon Romanescu-Latinescu-Ungurescu-Nemtescu quare de mai multi autori, in cursul a trideci, si mai multoru ani s'au lucratu. Seu Lexicon Valachico-Latino-Hungarico-Germanicum quod a pluralibus auctoribus decursu triginta et amplius annorum elaboratum est*. Ediție electronică realizată de Maria Aldea, Daniel-Corneliu Leucuța, Lilla-Marta Vremir, Vasilica Eugenia Cristea, Adrian Aurel Podaru. Cluj-Napoca. Consulté sur <https://doi.org/10.26424/lexiconuldelabuda> [09/01/2020].
- Marello, C. (2010). Multilexical Units and Headword Status. A Problematic Issue in Recent Italian Lexicography. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. 6-10 July 2010. Leeuwarden / Ljouwert : Fryske Akademy – Afûk, pp. 1347-1354. Consulté sur <https://euralex.org> [24.01.2020].
- Milică, I. (2013). *Lumi discursive. Studii de lingvistică aplicată*. Iași : Junimea.
- Moon, R. (1998). *Fixed Expressions and Idioms in English. A Corpus-Based Approach*. Oxford : Clarendon Press.
- Munteanu, C. (2013). *Frazeologie românească. Formare și funcționare*. Iași : Institutul European.
- Negreanu, C. (1983). *Structura proverbelor românești*. București : Editura Științifică și Enciclopedică.
- Pelea, A. (2020). *La Traduction : crapaud ou Prince charmant ?! Aspects culturels de la traduction du conte merveilleux*. Cluj-Napoca : Casa Cărții de Știință. Consulté sur <https://play.google.com/books/reader?id=1RDHJwAAAEAJ&pg=GBS.PA0> [04.07.2020].
- Seche, M. (1966). *Schiță de istorie a lexicografiei române, vol. I. de la origini pînă la 1880*. București : Editura Științifică.
- Svensén, B. (1993). *Practical Lexicography. Principles and Methods of Dictionary-Making*. Translated from Swedish by John Sykes and Kerstin Schofield. Oxford / New York : Oxford University Press.
- Ștefănescu, M. (2019). *Studii de lexicografie și semantică interculturală*. Cluj-Napoca : Casa Cărții de Știință.
- Tabarcea, C. (1982). *Poetica proverbului*. București : Editura Minerva.
- Tamba, I. (2000). Le sens métaphorique argumentatif des proverbes. In *Cahiers de praxématique* 35. Montpellier : Pulm, pp. 39-57.
- Visetti, Y.-M., Cadiot, P. (2006). *Motifs et proverbes. Essai de sémantique proverbiale*. Paris : Presses Universitaires de France.
- Vîrva, D. (2019). *Lexicul proverbelor românești. Studiu asupra Proverbelor românilor, de Iuliu A. Zanne*. Teză de doctorat. Universitatea Babeș-Bolyai, Cluj-Napoca, România.
- VRF. Pour l'ouvrage de Costinescu (1870).
- Zgusta, L. (1971). *Manual of Lexicography*. The Hague-Paris : Mouton.

## Annexe

- (1) Cămesia [Chemise], *f. pl.* [...] *subst.* [...] mi-[i] mai aproape cămeșă decât țundra : *prov. tunica pallio propior, indusium toga propius* : közelebb az ing a menténél : das Hemde ist näher als der Rock [la chemise m'est plus proche que la bure].
- (2) Cioroboru [Querelle], sau ciorlu, morlu, *m.* [...] ciorobor pentru-un topor : *prov. rixari de umbra asini*, Plaut. *vel de lana caprina* : egy haszontalanúságért szóvátkozni : um nichtswürdige Dinge zanken [se disputer pour rien].
- (3) Cornu [Corne]. [...] *subst.* [...] o minciună cu coarne : *prov. mendacium magnum, et impudens* : egy iszonyú nagy hazugság : eine große Lüge [un gros mensonge].
- (4) Credu, credere, credzutu [Croire]. [...] *verb. act.* [...] vezi cui crezi : *prov. fide, sed cui vide* : trau, schau wem [fais confiance, mais prends garde à qui].
- (5) Dapiru, are, atu [Bobiner]. [...] *verb. act.* [...] cu o mână te apără, cu alta te dapără : *prov. altera manu panem fert, altera lapidem ostentat* : vorn lecken, hinten kratzen [dans une main, il porte une pierre, dans l'autre, un morceau de pain].



- (6) Fumu [Fumée], *m. pl.* [...] 1) [...] de unde nu e foc, nu iese fum : *prov. scintilla ignis indicium est* : nincsen tűz füst nélkül : wo Funken aus dem Kamin fliehen, da ist Feuer [où il n'y a point de feu, il n'y a point de fumée].
- (7) Grópá [Fosse], *f. pl.* [...] *subst.* [...] care sapă groapă altuia cade însuși într-însa : *prov. incidit in foveam quam fecit* : aki másnak vermet ás, maga esik belé : wer einem andern eine Grube gräbt, fällt selbst hinein [qui creuse la tombe d'autrui tombe lui-même dedans].
- (8) Intându, dere, ténsu [Étendre]. [...] *I. verb. activ.* [...] *II. reciproc.* [...] 2) [...] trebuie să ne întindem pre cât ne ajunge țolul, cerga : *ne ultra pedem calceus* : csak addig nyújtózzál, ameddig a takaró ér : man muß sich nach der Decke strecken. *proverb.* [on doit s'étendre tant que le permet la couverture, le plaid].
- (9) Lupu [Loup], *pl.* [...] *subst.* 1) [...] lupul își schimbă părul, dar nu-ș lasă năravul : *provb. naturam expellas furca, tamen usque recurret, vel lupus pilum mutat, non animus* : der Wolf wird älter, aber nicht frömmer, oder der Fuchs läßt Haare, aber nicht die Tücke [le loup change de poil, mais rien de ses coutumes]. – lupul mănă și din oile cele numărate : *provb. lupus numerum non curat* : der Wolf frißt auch die gezählten Schafe [le loup mange même les brebis comptées].
- (10) Mână [Main], *f. pl.* [...] *subst.* 1) [...] o mână spală pre alta : *proverb. manus manum lavat* : egy kéz mossza a másikat : eine Hand wäscht die andere [une main lave l'autre].
- (11) Óe [Brebis], *f. pl.* [...] *subst.* [...] o óe reîosă imple totă turma : *proverb. morbida facta pecus totum corrumpit ovile* : ein räudiges Schaf steckt die ganze Herde an [une brebis galeuse gâte le troupeau].
- (12) Omenie [Humanité], *f.* [...] *sub.* [...] omenia omenie așteaptă : *proverb. gratia gratiam parit, officium officio provocatur* : a becsület becsületet kíván vizontag : ein Dienst erfordert den andern [fais le bien et tu trouveras le bien].
- (13) Riiosu, séu rîiosu [Galeux] [...] *adj.* [...] *prov.* o (una) óue rîiosă imple totă turma : *morbida facta pecus, totum corrumpit ovile* : egy rühes juh megrontja az egész nyáját : ein räudig Schaf steckt die ganze Herden [une brebis galeuse gâte le troupeau].
- (14) Rînduné, séu rînduné, *f. pl.* [...] Rînduré [Hirondelle] [...] *subst.* [...] *proverb.* una (o) rîndunea nu face primăvară : *una hirundo non facit ver* : *Ital.* una rondine non fa primavera : *Gall.* une hirondelle ne fait pas le printemps, – însemnează că o întâmplare nu face regulă [signifie qu'un évènement ne fais pas la règle] : egy fecske nem teszen tavaszt [...].
- (15) Acū [Aiguille]. *s.etr.* [...] *prov. fig.* A căuta un ac în carul cu fân : a căuta un lucru foarte greu de găsit. *Chercher une aiguille dans une botte de foin* [chercher quelque chose de difficile à trouver] [...].
- (16) Barbă [Barbe]. *s.f.* [...] *A face drum prin barbă* : a repeta mai de multe ori, a obicinui des ca ceea ce la întâia încercare s-a permis, s-a iertat, s-a trecut cu vederea ca greșală. *Faire chemin dans la barbe* [répéter à plusieurs reprises, avoir l'habitude de pardonner, d'ignorer la première erreur]. *Cine are barbă să aibă și peptene* : cine are un lucru cată să-l aibă complect cu toate accesoriile lui spre a nu se împrumuta, ori a ruga pe altul ca să i le îlesnească. *Celui qui a la barbe, doit aussi avoir un peigne* [quand on achète quelque chose il faut le faire avec ses propres moyens pour ne pas dépendre d'une autre personne]. – *A căuta cui va în barbă* : a stima, a respecta pe cineva ; a-și stăpâni furia către el pentru vârsta sau oarecare virtuți ce posedă. *Chercher dans la barbe de quelqu'un* [respecter quelqu'un ; contrôler sa furie envers l'autre en raison de son âge ou de ses qualités]. Toate aceste expresiuni sunt în stilul fam. sau prov. [Toutes ces expressions sont en registre familier ou proverbial] [...].
- (17) Caldū [Chaud]. *adi.* [...] *prov. fig.* Bate fierul până este cald : folosește-te de circumstanțele favorabile, nu scăpa timpul și ocaziunea favorabilă. *Il faut battre le fer pendant qu'il est chaud* [profite des circonstances favorables, ne perds pas le temps et l'occasion favorable].
- (18) Călugărū [Moine]. *s.m.* [...] *prov.* Nu face haina pe călugăr : nu trebuie să judeci pe cineva după cele din afară. *L'habit ne fait pas le moine* [il ne faut pas juger quelqu'un selon son aspect].
- (19) Cămășă [Chemise]. *s.f.* [...] *prov. fig.* Mai aproape mi-e cămeșă decât mantaua : să vedem întâi de noi, de interesele noastre până la ale altora. *Notre peau nous est plus près que notre chemise* [qu'on s'occupe tout d'abord de nos affaires et ensuite de celles des autres].
- (20) Ciokine [Trousse]. *s. plr.* [...] *fam. prov.* A lega, a atârna, a spânzura ceva la ciochine : a părăsi, a abandona, a lepăda de bună voia o datorie, o funcțiune, o însărcinare. [*Jeter aux oubliettes* : abandonner, renoncer à un devoir, à une fonction].
- (21) Confessare [Confesser]. *v.s.* [...] *prov.* Greșeala confesată este pe jumătate iertată. *Une faute confessée est à demi pardonnée.*
- (22) Corbū [Corbeau]. *s.m.* [...] *prv.* Negru ca corbul. *Noir comme un corbeau.*
- (23) Culcare [Coucher]. *v.s.* [...] *iro. prov.* Culcă-te pe urechea aceia : în dar nu speri sau aștepti că nu vei avea ceea ce gândești ori ceri. *Attendez-moi sous l'orme, Attendez-vous-y* [tu as beau espérer ou attendre car tu n'obtiendras pas ce que tu penses obtenir ou ce que tu demandes] [...].
- (24) Dispunere [Disposer]. *v.s.* [...] *prov.* Omul propune, Dumnezeu dispune [L'homme propose, Dieu dispose] [...].
- (25) Domnū [Seigneur]. *s.m.* [...] *prov.* Nu după voia omului, ci după voia Domnului. *L'homme propose et Dieu dispose.*
- (26) Ertatū [Pardonné]. *sup.* și *adi.* [...] *prov.* Păcatul mărturisit este pe jumătate iertat. *Une faute confessée est à demi pardonnée.* [...].
- (27) Facere [Faire]. *v.s.* [...] *prov. fig. fam.* Cu încetul se face oțetul : încet și cu răbdare se săvârșește ceva sigur. *Petit à petit l'oiseau fait son nid* [la persévérance vient à bout de tout]. – *prov. fig.* Nu face haina pe călugăr : să nu judecăm pe oameni după aparență ; și nu este sânt cel ce se arată a fi. *L'habit ne fait pas le moine* [qu'on ne juge pas les gens selon leur aspect ; et il n'est pas saint celui qui fait semblant de l'être]. *prov. fig.* A face țințarul armăsarului : a exagera. *Faire d'une mouche un éléphant* [exagérer] [...].
- (28) Fînū [Foin]. *s.etr.* [...] *prov. fig.* A căuta un ac în carul cu fân : a căuta un lucru ce nu se mai poate găsi. *Chercher une aiguille dans une botte de foin* [chercher une chose qui ne peut pas être trouvée].
- (29) Feréstră [Fenêtre]. *s.f.* [...] Se zice *prov. fig.* de un obraznic de care nu se poate cineva scăpa [On dit *prov. fig.* d'une personne insolente qu'on ne peut pas éviter] : Îl dai afară pe ușă și intră pe fereastră, sau gonește-l pe ușă că el îți vine pe fereastră. *Si vous le faites sortir par la porte, il rentrera par la fenêtre, ou chassez-le par la porte, il rentrera par la fenêtre.*
- (30) Ferū [Fer]. *s.etr.* [...] *prov.* Nu e cineva de fier : spre a arăta că ostenește, obosește cineva până în sfârșit. *On n'est pas de fer* [pour montrer qu'une personne peut devenir fatiguée après un temps] ; cu asemenea înțeles [avec le même sens], Cată să fie cineva de fier ca să poată duce în asemenea ostenele. *Il faudrait être de fer pour résister à de telles fatigues.* [...] *prov.* Bate fierul până este cald : stăruiește într-o împregiurare, într-o afacere până ocaziunea este favorabilă. *Il faut battre le fer tandis qu'il est chaud* [persévère dans



une circonstance, dans une affaire jusqu'à ce que l'occasion favorable se présente] [...].

(31) Fine [Fin]. *s.m.* [...] *prover. Finele coronează lucrul : sfârșitul încoronează fapta. La fin couronne l'oeuvre* [la fin couronne l'œuvre] [...].

(32) Foc [Feu]. *s.etr.* [...] *prov. A arunca foc în paie : a întârâta, a face să izbucnească, să se declare ura între persoane ce nu se pot suferi. Mettre le feu aux étoupes* [déclencher, susciter la haine entre des personnes qui ne se supportent pas]. – *Nu e foc fără fum, sau nu iese fum de unde nu este foc* : că vorbele ce s-au făcut sunt mai deseori bazate pe un adevăr, că nu se fac vorbe fără motive, nu se zic calomnii când nu e vreun defect. *Il n'y a point de feu sans fumée*, ou *il n'y a pas de fumée sans feu* [on parle davantage d'une chose quand il y a une vérité derrière, on n'adresse pas d'injures quand il n'y a pas de défaut] [...].

(33) Generală [Général]. *adi.* [...] *prov. Nu este regulă generală fără excepțiune* [*Il n'y a pas de règle générale sans exception*] [...].

(34) Gîrlă [Ruisseau]. *s.f.* [...] *prov. Gârla curge, pietrele rămân : nu spera sau nu crede în străin mai mult decât în pământean, că el fuge la timp de nevoie, dar noi rămânem ; sau nămitul dispăre, turma rămâne. Le torrent s'écoule, la pierre reste* [ne fais plus confiance à un étranger qu'à ton concitoyen, car l'étranger s'enfuit quand les temps sont rudes, mais nous restons ; ou le serviteur disparaît, le troupeau reste].

(35) Haină [Habits]. *s.f.* [...] *prov. Nu face haina pe călugăr : să nu considerăm persoana după aparințe, după cele de afară. L'habit ne fait pas le moine* [ne pas juger une personne selon son aspect, d'après ce qui se voit à l'extérieur].

(36) Lebedă [Cygne]. *s.f.* [...] Se zice proverbialemente unei persoane ce are părul sau fața foarte albă *că este albă ca o lebedă* [On dit proverbialement d'une personne qui a les cheveux ou le visage très blancs], *qu'elle est blanche comme un cygne* [...].

(37) Lucire [Luire]. *v.s.* [...] *Soarele lucește pentru toată lumea : soarele dă lumina sa tuturor, sau, în proverbial, sunt niște lucruri de care oricine are dreptul a se bucura. Le soleil luit pour tout le monde* [le soleil brille pour tous ou proverbiallement il y a des choses dont chacun a le droit de bénéficier] [...].

(38) Lucru [Travail]. *s.etr.* [...] *prv. Sfârșitul încoronează lucrul : nu e destul începutul bun, trebuie a fi și sfârșitul bun. La fin couronne l'oeuvre* [il ne suffit pas d'un bon début, encore faudra-t-il que la fin soit bonne] [...].

(39) Lungire [Étendre]. *v.s.* [...] *prov. Lungește picioarele pe cât îți este pătura : nu te întinde cu cheltuielile mai mult decât îți este venitul. Étendre tes pieds selon ta couverture* [ne dépense pas plus que tu ne gagnes].

(40) Mânică [Manche]. *s.f.* [...] *fig. prov. Aceasta e altă mână : este altă treabă. C'est une autre paire de manches* [c'est une autre affaire].

(41) Mărgăritar [Perle]. *s.m.* [...] *prov. A arunca mărgăritarul porcilor : a da cuiva un lucru pe care nu știe să-l prețuiască, a spune cuiva lucruri pe care nu știe să le înțeleagă. Jeter des perles devant les pourceaux* [donner une chose à une personne qui ne sait pas l'apprécier, dire des choses à une personne qui ne peut pas les comprendre]. – *Nu se aruncă mărgăritarul porcilor. Il ne faut pas jeter les Marguerites devant les pourceaux.*

(42) Medic [Médecin]. *s.m.* [...] *prov. Nu face haina pe medic : nu este titlu care face pe om învățat. La robe ne fait pas le médecin* [ce n'est pas le titre qui rend l'homme érudit]. – *prov. Medice, lecuiește-te pe tine însuși* : se zice la o persoană care având defecte critică defectele altora, sau de acela care având lipsă de capacitate consiliază pe alții cum să fie capabili. *Médecin, guéris-toi, toi-même* [on dit d'une personne qui malgré ses défauts critique les défauts des autres, ou qui malgré son caractère défectueux donne des conseils].

(43) Miere [Miel]. *s.* [...] *prov. fig. Luna de miere : cea dintâi lună a căsătoriilor. La lune de miel* [le premier mois des jeunes mariés]. – *Mai multe muște se prind cu o lingură de miere decât cu o butie de oțet* : mai lesne-și ajunge cineva scopul cu blândețe decât cu rigoare și mânie. *On attrape plus des mouches avec du miel qu'avec du vinaigre* [une personne arrive plus facilement à ses fins par la bonté que par la sévérité et la colère].

(44) Minune [Merveille]. *s.f.* [...] *prov. Aceasta este una din șapte minuni ale lumii. C'est une des sept merveilles du monde* [...].

(45) Mur [Mur]. *s.m.* [...] *prov. Murii au urechi : pereții au urechi, ceea ce zici o să se afle. Les murs ont des oreilles* [les murs ont des oreilles, ce que tu dis va s'entendre].

(46) Mușchiu [Mousse]. *s.m.* [...] *prov. Pietra ce se rostogolește nu prinde mușchiu* : cel ce schimbă des locul sau profesiunea nu se-mbogățește. *Pierre qui roule n'amasse pas de mousse* [la personne qui change souvent de place ou de profession ne devient pas riche].

(47) Oțet [Vinaigre]. *s.etr.* [...] *prover. Mai multe muște se prind cu o lingură de miere decât cu o bute de oțet* : cu o manieră dulce reușește cineva mai bine decât cu o manieră de mândrie sau de asprime. *On prend plus de mouches avec un peu de miel qu'avec un tonneau de vinaigre* [on obtient davantage par la douceur que par un comportement orgueilleux ou âpre] [...].

(48) Ôlă [Pot]. *s.f.* [...] *prov. În oala acoperită nu dă gunoarie* : cel modest este cruțat de calomnii ; cel ce nu se expune nu se periculează. *Dans le pot couvert n'entrent pas les ordures* [les gens modestes ne sont pas calomnieux ; celui qui ne s'expose pas n'est pas en péril].

(49) Pământu [Terre]. *s.etr.* [...] *prover. Cine pământ are, pace n-are* : cine are proprietate este supus a avea procese. *Qui terre a, guerre a* [celui qui a des biens matériels risque des procès] [...].

(50) Pată [Tache]. *s.f.* [...] *fig. și prov. A găsi pete în soare* : a găsi defecte în lucrurile cele mai perfecte. *Trouver des taches dans le Soleil* [trouver des défauts dans les choses les plus parfaites] [...].

(51) Pecătosu [Pécheur]. *adi. și s.* [...] *pro. Dumnezeu nu voiește moartea păcătosului* : trebuie să iertăm greșelile altuia. *Dieu ne veut pas la mort du pécheur* [il faut pardonner les fautes de l'autre] [...].

(52) Pescuire [Pêcher]. *v.s.* [...] *prov. A pescui în apă turbure* : a profita de dezordine, a face să se întâmple dezordine spre a profita, atrage avantaje. *Pêcher en eau trouble* [profiter du désordre, déclencher du chaos pour en profiter, pour en tirer des avantages].

(53) Pesce [Poisson]. *s.m.* [...] *prov. pop. A fi ca peștele în apă* : a trăi undeva bine. *Être comme le poisson dans l'eau* [vivre bien]. – *A trăi sau a se bate ca peștele pe uscat* : a trăi foarte greu ; a nu putea fi sau trăi unde dorește cineva. *Être comme le poisson hors de l'eau* [vivre dans des conditions difficiles ; ne pas pouvoir vivre où l'on veut]. – *Peștele mare înghite pe cel mic* : cei mai tari apasă pe cei mai slabi. *Les gros poissons mangent les petits* [les forts pèsent sur les faibles] [...].

(54) Pêtră, Piétră [Pierre]. *s.f.* [...] *prov. fig. Interesul este peatra de-ncercare a amicitiei. L'intérêt est la pierre de touche de l'amitié*. [...] *Peatra ce rostogolește nu prinde mușchi* : cel ce schimbă meseria nu face stare. *Pierre qui roule n'amasse point de mousse* [la personne qui change son métier ne devient pas riche]. – *Apa curge, pietrele rămân* : cu cei ce trec nu se poate cineva asocia or profita,



ca cu cei ce sunt ai locului ; de la străin nu se poate aștepta un bine ca de la cei cu care trăiește cineva ; străinul nu te poate ajuta ca pământeanul [*L'eau coule, les pierres restent* : on ne peut pas s'associer ou avoir du profit avec quelqu'un comme on le fait avec les nôtres ; on ne peut pas attendre un bienfait de la part de l'étranger comme on en attend de la part de ceux avec lesquels on vit ; l'étranger ne peut pas t'aider comme t'aide le tien].

(55) Pițigoii [Pinson]. *s.m.* [...] *prov. Vesel ca un pițgoi* : foarte vesel. *Gai comme un pinson* [très gai].

(56) Populă [Peuple]. *s.m.* [...] *prov. Vocea popului este vocea lui Dumnezeu* : cu un înțeles mai obicinuit, simțământul general este fondat pe veritate. *La voix du peuple est la voix de Dieu* [plus communément, le raisonnement général est fondé sur la vérité].

(57) Profetă [Prophète]. *s.m.* [...] *prov. Nimeni nu poate fi profet în țara lui* : mai puțin considerat este cineva în locul său decât aiure [*Nul ne peut être prophète en son pays* : une personne est moins appréciée chez soi qu'ailleurs].

(58) Propunere [Proposer]. *v.s.* [...] *prov. Omul propune și Dumnezeu dispune* : omul chibzuiește și Dumnezeu otărăște ; nu este cum va omul, ci cum va Domnul. *L'homme propose et le Dieu dispose* [l'homme propose, Dieu dispose ; les choses ne se passent pas selon la volonté de l'homme, mais selon la volonté de Dieu].

(59) Sardoă [Herba sardoniana]. *s.f.* [...] D-aici acea locuțiune proverbială : *râs sardonian* [D'où la locution proverbiale : *rire sardonique*].

(60) Țără [Pays]. *s.f.* [...] *prover. fig. Nimeni nu este profet în țara lui* : este prea dificil a-și face cineva reputația în țara lui, ca la străini. *Nul n'est prophète dans son pays* [il est plus difficile de construire sa réputation chez soi qu'à l'étranger] [...].

(61) Tonă [Ton]. *s.etr.* [...] *Tonu* se întrebuițează în multe expresiuni particularii și proverbiale [le mot *ton* s'emploie dans plusieurs expressions spécifiques et proverbiales]. *A fi după tonul cuiva* : a avea conformitate în idei, în expresiuni, în gusturi. *Être au ton de quelqu'un* [penser, parler et avoir les mêmes goûts qu'une autre personne]. – *A face pe cineva să cânte pe alt ton* : a-l obliga să-și schimbe limbajul, manierele. *Faire chanter quelqu'un sur un autre ton* [obliger quelqu'un de changer son langage, son comportement] [...].

(62) Tradatoră [Traître]. *sub.* [...] *proverb. Trădător ca Iuda. Traître comme Juda*.

(63) Tragere [Tirer]. *v.s.* [...] *proverr. A trage pe dracul de coadă* : a subsista, a o duce cu multă strâmtorare. *Tirer le diable par la queue* [avoir des soucis financiers].

(64) Turbure [Trouble]. *adi.* [...] *prover. A pescui în apă turbure* : a profita din dezordinile publice sau particularii. *Pêcher en eau trouble* [profiter du chaos public ou particulier].

(65) Voce [Voix]. *s.f.* [...] *proverb. Vocea popului, vocea lui Dumnezeu* : când toată lumea se învoiește a crede generalmente un lucru, cată a se crede că toată acea lume este pe calea rațiunii. *La voix du peuple est la voix de Dieu* [quand tous les gens se mettent d'accord à croire une chose, il faut penser que tous ces gens suivent le chemin de la raison].



# The interaction of argument structures and complex collocations: role and challenges in learner's lexicography

Giacomini L.<sup>1,2</sup>, DiMuccio-Failla P., Lanzi E.<sup>2</sup>

<sup>1</sup> University of Hildesheim, Germany

<sup>2</sup> University of Heidelberg, Germany

## Abstract

This contribution focuses on the status of complex collocations in pattern-based learner's dictionaries, reporting on findings of the ongoing corpus-based project *Pattern-based learner's lexicography* (Hildesheim University/Heidelberg University). After comparing recursively built complex collocations with argument-related complex collocations, the paper concentrates on the latter type and its functions. On the one hand, complex collocations displaying argument complementarity efficiently support the identification and formulation of sense patterns. On the other hand, they can serve different purposes within the microstructure of a pattern-based dictionary, namely as semantic types of sense patterns or as lexicographic items in a subordinate treatment unit. Argument-related complex collocations are phraseological lexicalisations of the conceptual scenes provided by sense patterns, and are therefore of key importance to language learners. The challenges related to the extraction of complex collocations from corpora are also addressed in the paper, and proposals are made for improving time efficiency, coverage, and quality of extracted candidates in future research.

**Keywords:** learner's lexicography; sense pattern; complex collocation; argument structure; cognitive lexicography; lexicogrammar

## 1 Introduction

Studies on the treatment of collocations in lexicography have been relatively frequent in recent decades, partly following the constant development of new approaches and tools in corpus linguistics. However, the traditional view of collocations as simple, binary combinations still dominates contemporary paper and electronic dictionaries. In particular, the potential of complex collocations for learner's lexicography has remained largely unexplored in relevant literature. This paper deals with the topic of complex collocations from the specific angle of their interaction with argument structures of verbs, specifically aiming to illustrate their advantages for learner's lexicography. This topic is related to the ongoing corpus-based project *Pattern-based learner's lexicography* carried out at Hildesheim University and Heidelberg University, and aimed at the compilation of an electronic pattern-based learner's dictionary of Italian.

The primary theoretical background of the project is provided by the tradition of linguistic approaches covering the interplay between lexis and grammar (cf. among others, Halliday 1992; Gross 1994; Herbst 2016/2017; Herbst et al. 2014), as well as by cognitive lexicography (Geeraerts 2007; Ostermann 2015). At the core of the proposed microstructural model is the notion of *normal patterns of usage* (Sinclair 1996/2004; Hanks 2013), or *sense patterns*, as the true lexical units of language. A sense pattern (e.g. EN *to follow someone going somewhere*, DE *jdn./etw. mit den Augen verfolgen*, or IT *accompagnare qualcuno in un luogo/in un percorso*) is a syntactic-semantic entity given by a combination of syntactic and semantic arguments, semantic types, and semantic roles, uniquely identifies one meaning of a word, and has a largely phrasal nature (Sinclair 2004)<sup>1</sup>. In the pre-lexicographic stage of the project, the focus lies on exploring theoretical models which can help us cover the interplay between sense patterns and collocations in lexicography, on devising a semi-automated, corpus-based method for identifying and formulating sense patterns, as well as on designing the microstructure of verb entries. In performing these tasks, we take advantage of the contrastive analysis of verb patterns in Italian, German, English, and French. In this contribution, we will present data extracted from Italian and German corpora<sup>2</sup>.

As pointed out in Giacomini & DiMuccio-Failla (2019), the method for sense pattern identification has changed over time, maximizing the use of corpora in the process of pattern-related data retrieval. In particular, the method changed from initial concordance analysis as proposed in the context of Corpus Pattern Analysis (CPA, Hanks 2004) and of CPA-oriented approaches (cf. Renau & Nazar 2016), to the analysis of collocates. This radical methodological change originates from the observation of patterns originally identified by manual analysis of concordances, a large amount of which appears to be of a phraseological nature. The use of collocations for detecting the correlation between argument structures and meanings of a verb has proven to be a reliable and highly accurate method, in which both simple and complex collocations play a key role. In doing this, we take into account corpus-based studies exploring possible associations between collocations and constructions: besides those directly inspired by Sinclair's *idiom principle* (cf. Stubbs 1995), also the ones carried out in the context of Pattern Grammar (cf. Hunston & Francis 2000), of Construction Grammar (cf. collostructional analysis, Stefanowitsch & Gries 2003, and the idea of language as a Collostruction in

<sup>1</sup> A detailed discussion of the semantic and phraseological features of sense patterns is provided by DiMuccio-Failla & Giacomini (2017a/2017b).

<sup>2</sup> Examples will be followed by an English translation. For the sake of clarity, the abbreviations IT, DE, and EN for the three languages will be used throughout the paper.



Herbst 2018), as well as of Frame Semantics (cf. Almela-Sánchez 2019).

Section 2 of this paper discusses state-of-the-art approaches to complex collocations, focusing on their formation by recursive expansion or by argument complementarity. Section 3 concentrates on the advantages of complex collocations with argument complementarity for pattern-based learner's lexicography, discussing the challenges related to their extraction from corpora, as well as their possible metalexicographic and lexicographic applications. Section 4 provides some insights into initial tests for the evaluation of complex collocation coverage within the planned dictionary, and hints at future research developments aimed at the optimization of complex collocation extraction from corpora in the context of the lexicographic process.

## 2 The formation of complex collocations

The focus of our paper lies in the description of the metalexicographic and lexicographic use of complex collocations of verbs in a pattern-based learner's dictionary. In the context of lexicography- and computationally-oriented studies on collocations, complex collocations have generally been described or defined as collocations involving more than two content words, largely drawing on the traditional view of a collocation as a primarily binary word combination. While the general understanding of collocations as n-ary combinations has been gaining some popularity on the basis of corpus evidence in the recent past, the nature of complex, i.e. larger than binary, collocations appears not to have been explored in depth and to generally conform to a single description model.

### 2.1 Complex collocations built by recursive expansion

Complex collocations have been named in different ways, ranging from *grammatically extended collocations* (Tutin & Kreif 2016), *collocation chains* (Alonso Ramos & Wanner 2007), and *nested collocations* (Seretan 2011), up to *collocations of collocations* (Wehrli et al. 2010). Unfortunately, qualitative and quantitative lexicogrammatic approaches on which we rely for sense pattern identification (cf. Section 1) do not provide a specific focus on complex collocations. The topic of complex collocations sporadically appears in the context of very specific empirical studies, without being directly embedded in a distinct theoretical framework.

In his account of lexical combinatorics and challenges posed by the acquisition and application of collocational information, Heid (1994: 231) writes: "An additional problem of the interaction between syntactic and collocational description is the recursive nature of collocational properties: the components of a collocation can be again collocational themselves: next to the German collocation *Gültigkeit haben* (EN *to be applicable*) (n+v), we have *allgemeine Gültigkeit haben* (EN *to be generally applicable*), with the collocation *allgemeine Gültigkeit* (EN *general applicability*) (n+a) as a component. These cases have sometimes been analyzed as different from collocations, but there is no reason for such treatment."

The *recursive* nature of collocations implies that a core collocational phrase is progressively expanded by the addition of new collocates, like in the abovementioned example *allgemeine Gültigkeit haben* (Heid 1994: 231). As a result, simple collocations are embedded in complex ones (cf. also Seretan 2013). The typical case is the expansion by means of adjectival or adverbial modifiers. Simple collocations in this sense are potentially open to any extension, the only limit being the collocational range (McIntosh 1966) of the progressively added components. This natural limit will cause complex collocations to usually consist of a limited number of elements.

Heid (1994) also highlights the importance of a formal account of this kind of collocation, for instance for machine translation. In our opinion, the importance of complex collocations goes far beyond the issue of their formalization, and equally affects syntactic, semantic, and lexicological analysis. As later pointed out by Zinsmeister & Heid (2003) in the context of adjective-noun-verb triples, a frequent case is the combination of two collocations with the same base, e.g. in the German examples *eine klare Absage + Absage erteilen* (EN *a clear refusal + to give a refusal*), or *absolute Mehrheit + Mehrheit erreichen* (EN *absolute majority + to obtain a majority*). The same principle applies to other languages, for example to English (cf. *spark strong emotions* mentioned in Gouws 2015: 172).

Lexicography possibly provides the perfect environment for observing complex collocations, not only from a theoretical point of view but also from the empirical perspective of their extraction from corpora and their presentation to end users such as foreign language learners or translators.

### 2.2 Complex collocations built by argument complementarity

In the context of our lexicographic project, the application of the new pattern-based approach to meaning representation coincides with a new perspective on complex collocations. Our focus has been mainly on verbs as the substantial carrier of sense pattern structures, and it is exactly during our study of verb argument structures that we identified the significant role complex collocations can play in a learner's dictionary. In extracting, validating, and sorting collocations, we noticed that a different type of complex collocation can be identified and described for the purpose of verb pattern treatment. This type of complex collocation directly involves the level of argument structures: collocations extracted for at least two different arguments of a verb are often syntactically and semantically complementary to each other in such a way that the native speaker perceives them as a syntactic but also semantic continuum. In the case of simple collocations mapping onto verb argument structures, we cannot speak of a recursive feature but rather of *complementarity*, since complementary argument-specific collocations simultaneously combine with each other.

This is, for instance, the case of some usual constructions of the Italian verb *accompagnare* (EN *to accompany*) such as *il padre accompagna la sposa all'altare* (EN *the father walks the bride down the aisle*), *il cantante è accompagnato al pianoforte* (EN *the singer is accompanied on the piano*), or *piatti tradizionali accompagnati da ottimi vini* (EN *traditional cuisine accompanied by excellent wines*). It is clear that the verb and its arguments build a coherent *scene*



(intended as a conceptual entity in the sense of Fillmore 1975) in which each component fulfils a cognitively functional role. Unlike simple collocations, these scenes can better reflect cultural specificities, with complex collocations, e.g. Italian *la madre accompagna i bambini a scuola* (EN *the mother takes her children to school*) possibly matching free, unremarkable word combinations in other languages.

Here, the main source of extension restriction is the valency of a verb, i.e. the number of elements within a complex collocation primarily depends on the number of arguments of the verb. The collocational range of simple collocations for each specific argument naturally becomes a further criterion for restriction.

Literature on collocations sometimes contains examples for complex collocations with argument complementarity without distinguishing it from the recursive expansion type. In describing German and Spanish specialised collocations in the field of investment funds, for instance, Ana Caro Cedillo (2004: 78) points out that “Die Zwei-Konzepte-Kollokation ist die Grundform. Einfache Kollokationen können aber weiter von anderen Elementen bestimmt werden. Sie können sich miteinander verketteten und komplexe, aus mehr als zwei Konzepte bestehende Kollokationen bilden”<sup>3</sup>. Besides examples of recursively expanded collocations (e.g. *einen Wertzuwachs von x% erzielen*, EN *to achieve a x% increase in value*, *ibid.*: 215), the book also provides examples such as *dem Anteilwert einen Ausgabeaufschlag hinzurechnen* (EN *to add an issuance fee to the unit value*) as a complex collocation of *Ausgabeaufschlag* (EN *issuance fee*), which is a simple collocation in the form of a compound (*ibid.*: 223). All elements of the complex collocations are paired with arguments of the verb *hinzurechnen* (EN *to add*), which has a three-argument structure covering the syntactic functions subject, direct object, and indirect object).

### 2.3 Comparing modalities of complex collocation formation

The two modalities of complex collocation formation can be compared along four salient features which have been summed up in Table 1.

Formation:	Recursive expansion	Argument complementarity
a) Formation process:	Expansion of a binary collocation through the progressive addition of words	Concatenation of simple collocations of a verb matching two or more of its arguments
b) Semantic core:	Content word	Verb taking at least two arguments
c) Restriction rationale:	Collocational range	Valency
d) Conceptual range:	Phrase level	Sentence level

Table 1: Comparison between different modalities of complex collocation formation.

a) The *formation process* by which a complex collocation is created differs depending on the semantic core: a simple collocation can be gradually expanded by the addition of new lexical items or, alternatively, it can concatenate with other collocations corresponding to further arguments of the core, typically a verb.

b) The *semantic core* of a complex collocation built by recursive expansion can be any type of content word, for instance a noun modified by an adjective, a verb by an adverb, or an adjective by an adverb. In the case of complex collocations built by argument complementarity, the semantic core is typically constituted by a verb taking at least two arguments, e.g. a subject and a direct object<sup>4</sup>. The status of a word as a base or a collocate (cf. Hausmann 1985: 119) is not taken into account, since any element of a simple collocation can serve as the semantic core of the new collocation.

c) As previously mentioned in this section, the two formation modalities are also characterised by crucial differences in the *restriction rationale*. The primary principle behind the possibility that a simple collocation builds a complex one by recursive expansion is the collocational range of its elements, whereas in the case of argument complementarity a restraint is imposed by the number of arguments of the semantic core.

d) The *conceptual range* of a complex collocation is the syntactic level at which its concept is encoded. From this perspective, complex collocations built by recursive expansion have the same characteristics as the simple collocations from which they originate. A noun phrase, for instance, is expanded into a larger noun phrase by the addition of an adjectival modifier, or a verb phrase is expanded into a larger verb phrase by the addition of an adverbial modifier: in both cases, the concept encoded by the complex collocation is still specified at the phrase level. Concepts covered by argument-related complex collocations, on the contrary, are embedded at sentence level. This level is also able to identify complex *scenes* as shown in Section 2.2.

## 3 Complex collocations in pattern-based learner’s lexicography

Argument complementarity in complex collocations is crucial for meaning description in learner’s lexicography, providing learners with a consistent, non-fragmented view of the phraseological templates typical of a language. This is particularly true of dictionaries focusing on sense patterns based on argument structures, as described in Section 1. The Italian verb *inseguire* (EN *to chase, to pursue*), for instance, counts among its sense patterns the pattern *inseguire qualcuno (che cerca di non farsi raggiungere)* (EN *to chase someone (who is trying not to be caught up)*). Both obligatory

<sup>3</sup> “The two-concept collocation is the basic form. However, simple collocations can be further determined by new elements. They can be linked together to form complex collocations consisting of more than two concepts” (our translation).

<sup>4</sup> Content words other than verbs can also constitute semantic cores, as long as they have their own arguments. This is, for example, the case of the Italian nouns *libertà + di parola* (EN *freedom + of speech*) or *rispetto + delle regole* (EN *adherence + to the rules*).



arguments, subject phrase and object phrase, subsume a number of simple collocations. From the perspective of the sense pattern as a cognitively founded unit of meaning, however, some simple collocation pairs appear to be syntactically and semantically linked to each other and build complex collocations, e.g. *il cacciatore insegue la selvaggina* (EN *the hunter chases the game*).

We are now going to discuss some issues related to the extraction of this type of complex collocation from corpora, and their use in the microstructure of a pattern-based learner's dictionary.

### 3.1 Extracting complex collocations from a corpus

The analysis of significance scores, and possibly of parsed data, is useful for validating collocation candidates and could potentially be applied to any kind of collocation. However, widespread corpus query systems primarily concentrate on the extraction of simple collocation candidates, while the retrieval of complex collocations is usually left to the linguistic and technical skills of the lexicographer.

For the identification of complex collocations we employ different methods and tools. We first use Sketch Engine (Kilgariff et al. 2004), collecting collocations of a node verb from the Italian Web 2016 corpus through the Word Sketch tool. This tool extracts binary word combinations, sorting them according to the specific grammatical relations defined by a *sketch grammar*. The corpus needs to be POS-tagged and lemmatised, whereas no parsing is required. Whenever relevant, Word Sketch displays the most frequent representation of a binary combination, possibly revealing some complex collocation. Table 2 shows some results for the Italian verb *inseguire* (EN *to chase, to pursue*), with the example of a complex collocation candidate based on argument complementarity:

Search word:	Collocation candidate:	Most frequent form in the corpus:
<i>inseguire</i> (EN <i>to chase, to pursue</i> )	subject of <i>inseguire</i> : <i>squadra</i> (EN <i>team</i> )	<i>due squadre si inseguono</i> (EN <i>two teams chase each other</i> ) (binary candidate)
	subject of <i>inseguire</i> : <i>notte</i> (EN <i>night</i> )	<i>la notte insegue sempre il giorno</i> (EN <i>the night always follows the day</i> ) (complex candidate with argument structure: subject + direct object)

Table 2: Search for complex collocations in the Italian Web 2016 corpus through word sketches (Sketch Engine).

This is by no means an efficient solution for identifying complex collocations in the corpus. In order to systematically look for argument-related complex collocations, we use the Sketch Engine's multiword sketch function, which extends the search for collocation candidates to further collocates of the original word sketches. This expansion enables the detection of complex collocations, as exemplified by the Italian noun *braccio* (EN *arm*) and the verb *accompagnare* (EN *accompany*) in Table 3:

Search word:	Collocation candidate:	Complex collocation candidate:
<i>braccio</i> (EN <i>arm</i> )	verbs with <i>braccio</i> as object: <i>tendere</i> (EN <i>stretch</i> )	modifiers of <i>tendere</i> + <i>braccio</i> : <i>destro</i> (EN <i>right</i> ) (recursively built complex candidate: <i>tendere il braccio destro</i> , EN <i>to reach out your right arm</i> )
<i>accompagnare</i> (EN <i>to accompany</i> )	object of <i>accompagnare</i> : <i>visitatore</i> (EN <i>visitor</i> )	subject of <i>accompagnare</i> + <i>visitatore</i> : <i>guida</i> (EN <i>guide</i> ) (argument-related complex candidate: <i>la guida accompagna il visitatore</i> , EN <i>the guide accompanies the visitor</i> )
		prepositional phrase with <i>accompagnare</i> + <i>visitatore</i> : <i>lungo il percorso</i> (EN <i>along the route</i> ) (argument-related complex candidate: <i>accompagnare il visitatore lungo il percorso</i> , EN <i>to accompany the visitor along the route</i> )

Table 3: Search for complex collocations in the Italian Web 2016 corpus through multiword sketches (Sketch Engine).

Candidates are then validated by considering frequency and score of each collocation, and by introspection, in particular through the analysis of collocation contexts within GDEX-filtered corpus samples (Kilgariff et al. 2008) and the comparison with data from general and collocation dictionaries. The procedure that needs to be followed in order to find and validate complex collocates requires a considerable amount of time, which is mainly due to the fact that searches have to be separately performed on each simple collocation candidate. Experiments carried out with German corpora provided by the DWDS Wortprofil tool (<http://dwds.de/d/wortprofil>) substantiate these observations. Moreover, there seems to be no correlation between the availability of complex collocations at the level of sense patterns and the semantic specificity of a verb: in fact, we did not notice any particular difference between semantically generic verbs and more specific verbs.

Another method for retrieving complex collocations is the use of corpus query languages to formulate complex queries, including multiple arguments of a verb. We carried out concordance searches employing the Corpus Query Language



option in Sketch Engine, and also performed some tests on different corpora using the Corpus Query Processor provided by the IMS Corpus Workbench (Evert & Hardie 2011). The main limits of this method lie in the mandatory predefinition of the specific argument structure of a verb, as well as in the lack of a specific evaluation frame for collocation significance.

To the best of our knowledge, the topic of extraction of complex collocations from corpora has been treated only marginally in relevant literature. In the case of the extraction of adjective-noun-verb combinations, Zinsmeister & Heid (2003) pleaded for a parsing procedure with a lexicalised probabilistic grammar instead of a simple pattern-matching on part-of-speech shapes. A parsing-oriented, syntax-based approach to collocational data extraction is also discussed by Seretan (2011), who proposes the pre-processing of extracted bigrams in order to automatically infer longer collocations (ibid.: 103 ff.). This method identifies recursively built nested collocations such as *treaty on the non-proliferation of weapons of mass destruction* (ibid.: 104). Despite general usefulness of this type of complex collocation, the results are not sufficient for the purpose of sense pattern description.

The method proposed by Kraif & Diwersy (2014) also relies on a parsed corpus, from which *lexicograms*, i.e. models for the main syntactic collocates of a given node together with association measures, are extracted and can be recursively employed to find increasingly longer combinations. The definition of specific syntagmatic structures allows, in this case, for more precise results in terms of argument structures. For the French node noun *respect* (EN *respect*) in a verb-object relation, for instance, the output would include *inspirer un profond respect* (EN *to command deep respect*) and *imposer le respect des normes* (EN *to enforce compliance*), which would match the argument structure of the input noun. However, it is not clear to what extent extraction from a corpus can be carried out in a systematic way for the complete range of arguments of any search word.

From a lexicographic standpoint, the automated extraction of argument-related complex collocations from corpora still presents general problems in terms of time efficiency, coverage, and quality of results. Section 4 will mention some interesting perspectives for future research on this topic.

## 3.2 Complex collocations in a pattern-based learner's dictionary

After candidate validation, complex collocations are employed for different tasks in the dictionary making process. Not only do they play a significant role in the metalexicographic activity of sense pattern formulation, as illustrated in Section 3.2.1, they are also intended to be recorded as lexicographic data in dictionary entries. At the level of the dictionary's microstructure, in fact, the presentation of complex collocations can primarily take place in the following mutually exclusive ways:

- complex collocations as semantic types or semantic roles (Section 3.2.2);
- complex collocations as subordinate treatment units (Section 3.2.3).

### 3.2.1 Complex collocations as a base for sense pattern formulation

Complex collocations extracted from a corpus and manually validated are systematically employed for the formulation of sense patterns. This essential metalexicographic function is fulfilled in two ways. On the one hand, analysing and grouping syntactically and semantically homogeneous collocations supports the identification of a specific sense pattern, i.e. of a distinct argument structure associated with a distinct meaning. For instance, the complex collocations *il poliziotto insegue il ladro* (EN *the policeman pursues the thief*) and *il malintenzionato insegue la vittima* (EN *the ill-intentioned person pursues the victim*) help identify the sense pattern *inseguire qualcuno (che cerca di non farsi raggiungere)* (EN *to chase someone (who is trying not to be caught up)*)<sup>5</sup>.

On the other hand, the analysis of paradigmatic structures matching verb arguments supports the selection of appropriate semantic types and semantic roles needed for the formulation of sense pattern. Semantic types, together with semantic roles, are the fundamental meaning components in sense patterns: in a hierarchy of concepts, they lexically represent the least common subsumer of all lexical items matching a specific verb argument in a specific verb meaning (Hanks 2004, DiMuccio-Failla & Giacomini, 2017a). The generic noun *vehicle*, for instance, is the suitable semantic type for a large cluster of lexical items such as *car*, *truck*, *bicycle*, *train*, or *ship*.

Observing collocate paradigms such as the ones at the subject and direct object levels in *il poliziotto/l'agente/la pattuglia/... insegue il ladro/il criminale/il sospettato/...* (EN *the policeman/the cop/the patrol/... pursues the thief/the criminal/the suspect/...*) or *il malintenzionato/il criminale/... insegue la vittima* (EN *the ill-intentioned person/the criminal pursues the victim*) is central for this process. The direct object of *inseguire* (EN *to pursue*) in this particular meaning, for instance, has been associated with the semantic type *qualcuno (che cerca di non farsi raggiungere)* (EN *someone (who is trying not to be caught up)*), which displays the most suitable level of generalization for subsuming all available collocates.

Of course, adequate semantic types do not always match the collocate of a verb, and sometimes they can even coincide with quite uncommon concepts, as exemplified by the English verb *toast*, with *breadstuff* as a semantic type for usual direct objects such as *bread*, *bun*, or *sandwich* (see discussion in DiMuccio-Failla & Giacomini, 2017a).

The following examples illustrate sense patterns with some of their subsumed complex collocations:

<sup>5</sup> Complex collocations extracted from corpora may involve adjuncts, e.g. *a piedi* (EN *on foot*) in *il poliziotto insegue il ladro a piedi* (EN *the policeman pursues the thief on foot*). However, it needs to be pointed out that in our lexicographic model, sense patterns of verbs typically cover syntactic and semantic arguments, but no adjuncts. Adjuncts are therefore not used for pattern formulation.



- i) IT *seguire (la direzione indicata da) una cosa* (EN *to follow (the indication given by) something*)  
 --subsumes--> *seguire la propria vocazione* (EN *to follow one's vocation*) (complex collocation)  
 --subsumes--> *seguire il richiamo della foresta* (EN *to follow the call of the wild*) (complex collocation)
- ii) IT *pedinare una persona (che sta andando da qualche parte)* (EN *to tail someone (who is going somewhere)*)  
 --subsumes--> *il poliziotto pedina il sospettato* (EN *the policeman is tailing the suspect*) (complex collocation)  
 --subsumes--> *il malintenzionato pedina la vittima* (EN *the attacker is stalking the victim*) (complex collocation)
- iii) IT *guidare un veicolo in un certo luogo* (EN *to drive a vehicle to a particular place*)  
 --subsumes--> *guidare la macchina fino al parcheggio* (EN *to drive the car to the parking lot*) (simple collocation only: *guidare la macchina*)  
 --subsumes--> *guidare una nave in porto* (EN *to steer a ship into harbor*) (complex collocation)
- iv) DE *einen Befehl (insb. einer Autorität) befolgen* (EN *to obey an order (esp. from an authority)*)  
 --subsumes--> *den Befehl des Vorgesetzten befolgen* (EN *to obey one's superior's orders*) (complex collocation)  
 --subsumes--> *der Soldat befolgt das Kommando* (EN *the soldier obeys the command*) (complex collocation)

### 3.2.2 Complex collocations as semantic types or semantic roles

In some cases verb collocates already reaching the most suitable level of generalisation can be directly elevated to semantic types, as shown in the following sense pattern examples for Italian and German<sup>6</sup>:

- v) IT *guidare un veicolo* (EN *to drive a vehicle*) (simple collocation: verb – direct object)
- vi) IT *tallonare un avversario in una classifica* (EN *to tail an opponent in a ranking*) (complex collocation: verb – direct object – adverbial)
- vii) DE *eine Aktivität wird von (dem Klang einer/eines) Stimme/Musikinstrument(s) begleitet* (EN *an event is accompanied by (the sound of) a voice/music instrument*) (complex collocation: subject – verb – direct object)

The main advantage of presenting semantic types or roles of sense patterns by means of complex collocations lies in the fact that dictionary users are simultaneously provided with typical scene-like structures and typical phraseological units. As is always the case for semantic types, complex collocations with this function may encompass more specific lexical items, among which are further collocates of the verb:

- viii) DE *eine Aktivität wird von (dem Klang einer/eines) Stimme/Musikinstrument(s) begleitet* (EN *an event is accompanied by (the sound of) a voice/music instrument*)  
 --subsumes--> *der Sonnenaufgang wird vom Gesang der Vögel begleitet* (EN *the sunrise is accompanied by the song of the birds*) (simple collocation only: *Gesang der Vögel*)  
 --subsumes--> *der Sänger wird am Klavier begleitet* (EN *the singer is accompanied on the piano*) (complex collocation)

### 3.2.3 Complex collocations as subordinate treatment units

Complex collocations that are not selected for presentation within a sense pattern, e.g. the subsumed collocations mentioned in examples i-iv and viii, can still be allocated in a subordinate treatment unit (cf. Gouws 2015) with a distinct search zone for each sense of the verb<sup>7</sup>. In the entry of the verb *pedinare* (EN *to tail*), the sense pattern *pedinare una persona (che sta andando da qualche parte)* (EN *to tail someone (who is going somewhere)*) as a treatment unit can be presented as in Figure 1<sup>8</sup>.

Both the given sense pattern and its subpattern *una persona fa pedinare un'altra persona da qualcuno* (EN *someone has someone else being followed by a person*) build independent microstructural treatment units that include a subordinate component dedicated to collocations. Being compositional phraseological units, complex and simple collocations do not require a meaning paraphrase.

<sup>6</sup> The sense pattern is underlined in each example.

<sup>7</sup> We deliberately use the adjective *subordinate* instead of *secondary* to distinguish it from the traditional vision of the lemma as the *primary* treatment unit (Wiegand 1996, Gouws 2015), highlighting at the same time the dependence of collocations on sense patterns as the superordinate and key microstructural element of our pattern-based dictionary model.

<sup>8</sup> Features illustrated in Figure 1 and Figure 2 are excerpts from a prototype of the planned pattern-based dictionary.



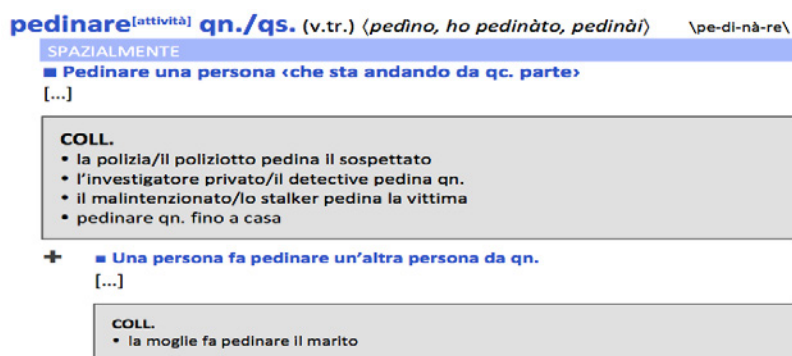


Figure 1: Excerpt from the entry for the Italian verb *pedinare* (EN *to tail*) with a sense pattern and a sense subpattern.

In discussing the insertion of complex collocations in dictionaries, Gouws (2015: 184-185) states the following:

The inclusion of complex collocations remains important and lexicographers should negotiate the best possible way of presenting them and of making users aware of their existence. This could be done either as guiding elements in a subcategory of the search zone for collocations or in a more implicit way as part of the treatment of single collocations, typically within an example sentence illustrating the use of the single collocation but also its occurrence as component of a complex collocation.

Whereas recursively built complex collocations can be easily seen as a subcategory of simple collocations, we think that argument-related complex collocations should be treated as a superordinate category subsuming simple collocations. The planned data representation in XML format allows for a hierarchical distribution of microstructural items (Figure 1) and for the attribution of argument-related collocates in the collocation treatment unit to the arguments of the corresponding sense pattern (Figure 2). Dictionary users will then be able to search for complex collocations matching all arguments or simple collocations matching specific arguments of a verb pattern. Adding thematic roles to argument representation further increases the degree of semantic detail. Whenever relevant, results can be expanded to adjunct-related collocations, e.g. *pedinare qualcuno fino a casa* (EN *to tail someone all the way to their home*). Figure 2 shows all validated collocates of *pedinare* identified for the selected sense pattern: collocates are sorted according to the related argument and, in the case of complex collocations, are linked to further collocates, e.g. in *il poliziotto pedina il sospettato* (EN *the policeman is tailing the suspect*).

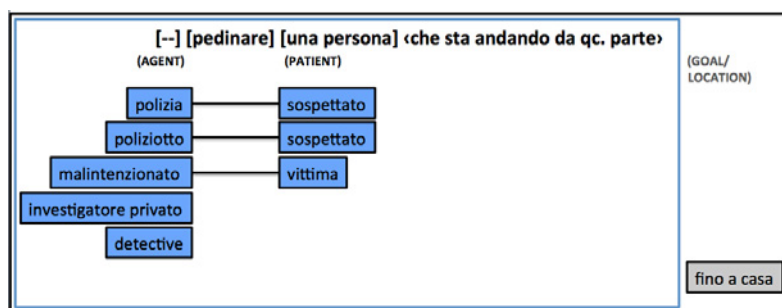


Figure 2 – Prototype visualization of results for search queries combining argument structures, thematic roles and collocates of the Italian verb *pedinare* (EN *to tail*): simple and complex argument-related collocates are highlighted in blue, adjuncts in grey.

This data model highlights complex collocations as the phraseological building blocks of sense patterns, and simple collocations as their basic constituents. Moreover, any search parameter can be employed here by the user to perform a query and be combined with other parameters, e.g. to retrieve all complex collocations for a given sense pattern, all complex collocations matching specific arguments, or all collocates matching a specific thematic role.

Embedding argument-related complex collocations in lexicographic examples provided to illustrate the use of a sense pattern is an option we generally do not take into account, since the information conveyed by this solution might be too implicit for language learners.

#### 4 Conclusions

This contribution has focused on a type of complex collocation involving the level of argument structures, in particular of verbs, showing how they can support multiple metalexicographic and lexicographic functions in a pattern-based learner's dictionary. Initial tests for the evaluation of complex collocation coverage within the planned dictionary were recently carried out at Heidelberg University by MA students in translation with Italian or German as their native (L1) or second/foreign language (L2/FL). Pattern-specific collocations of selected polysemous verbs were employed for performing a text reception and an active translation task. Participants were asked (1) to identify sense patterns of the same verb in a L2/FL text by analysing the available collocations, and (2) to translate L1 sentences containing different



senses of the same verb into the L2/FL by finding the right sense pattern through collocations. First results show that simple collocations of verbs are usually exhaustive enough to support sense disambiguation during text reception, whereas sense disambiguation for text production in (and translation into) the foreign language is improved by the availability of complex, argument-related collocations. These results seem to confirm the status of argument-related complex collocations as phraseological lexicalisations of conceptual scenes, and thus their relevance for language learners. Further tests on new datasets are planned for the future.

Some relevant issues have been highlighted in the paper regarding the extraction of complex collocations from corpora. Our future research on this topic will explore current findings in the field of terminology extraction, which can possibly yield some insights into methods for the detection and validation of n-grams and term variants in corpora. An interesting perspective in this context is also the investigation of the possibilities opened up by neural word embeddings, in particular in the field of analogy recovery (cf. Goldberg 2017). As pointed out in Section 3, reliable results are needed in terms of time efficiency, coverage, and quality of extracted data. However, we are convinced that only a solid underlying theory on complex collocations can support the development of new lexicographic-oriented procedures.

## 5 References

- Almela-Sánchez, M. (2019). Collocation and Selectional Preferences: A Frame-based Approach. *Journal of English Studies*, 17 (2019), pp. 3-41.
- Alonso Ramos, M. & Wanner, L. (2007). Collocation chains: how to deal with them? In K. Gerdes, T. Reuther, L. Wanner (Eds.), *Proceedings of the Third International Conference on Meaning-Text Theory* (pp. 11-20). Wiener Slawistischer Almanach, Sonderband 69. Munich.
- Cedillo, A. C. (2004). *Fachsprachliche Kollokationen: Ein übersetzungsorientiertes Datenbankmodell Deutsch-Spanisch* (Vol. 63). Tübingen, Germany: Gunter Narr Verlag.
- DiMuccio-Failla, P.V. & Giacomini, L. (2017a). Designing an Italian learner's dictionary based on Sinclair's lexical units and Hanks's corpus pattern analysis. In *Proceedings of the Fifth eLex Conference Electronic Lexicography in the 21st Century*. Leiden, Netherlands.
- DiMuccio-Failla, P.V. & Giacomini, L. (2017b). In M. Mitkov (ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017*, LNAI 10596. Springer, pp. 290-305.
- Evert, S. & A. Hardie (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. <http://cwb.sourceforge.net/index.php>
- Fillmore, C. J. (1975). An alternative to checklist theories of meaning. In: *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society, pp. 123-131.
- Geeraerts, D. (2007). Lexicography. In: D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics*. OUP, pp. 1160-1174.
- Giacomini, L. & DiMuccio-Failla, P. (2019). Investigating Semi-Automatic Procedures in Pattern-Based Lexicography. In: *Proceedings of the eLex 2019 conference. Electronic lexicography in the 21st century*. Sintra, Portugal.
- Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.
- Gouws, R. H. (2015). The presentation and treatment of collocations as secondary guiding elements in dictionaries. *Lexikos*, 25, pp. 170-190.
- Gross, G. (1994). Classes d'objets et description des verbes. *Langages*, pp. 15-30.
- Halliday, M. A. K. (1992). Some lexicogrammatical features of the zero population growth text. *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*. Benjamins, pp. 327-358.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.
- Hanks, P. (2004). Corpus Pattern Analysis. In: *Proceedings of the XI EURALEX International Congress*, Vol. 1, 87-98.
- Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch: Ein Beitrag zur Theorie des lexikographischen Beispiels. In: H. Bergenholtz & J. Mugdan (Eds.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch* 28. - 30. 6. 1984. Niemeyer, pp. 118-29.
- Heid, U. (1994). On ways words work together – research topics in lexical combinatorics. In: *Proceedings of the VI EURALEX International Congress*, Amsterdam, pp. 226–257.
- Herbst, T. (2017). Menschliche Sprache: Ein Netzwerk aus Mustern genannt Konstruktionen. *Sprachwelten: Vier Vorträge. Erlanger Universitätstage*, pp. 105-147.
- Herbst, T., Schmid, H. & Faulhaber, S. (Eds.) (2014). *Constructions Collocations Patterns*. De Gruyter Mouton.
- Herbst, T. (2016). Wörterbuch war gestern. Programm für ein unifiziertes Konstruktikon! In: S. Schierholz, R. H. Gouws, Z. Hollós & W. Wolski (Eds.), *Wörterbuchforschung und Lexikographie*. De Gruyter, pp. 169-206.
- Herbst, T. (2018). Is language a Collostruction? – A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions. In: P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications*. Springer, pp. 1-22.
- Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English* (Vol. 4). John Benjamins Publishing.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: *Proceedings of the XIII EURALEX International Congress*, Barcelona, pp. 425-431.
- Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 The sketch engine. *Information Technology*, pp. 105-116.
- Kraif, O. & Diwersy, S. (2014). Exploring combinatorial profiles using lexigrams on a parsed corpus: a case study in the lexical field of emotions. In: P. Blumenthal, I. Novakova & D. Siepmann (Eds.), *Les émotions dans le discours*.



- Emotions in discourse*. Peter Lang, pp. 381-394.
- McIntosh, A. (1966). Patterns and ranges. *Language* (37), pp. 325-337.
- Ostermann, C. (2015). *Cognitive lexicography: A new approach to lexicography making use of cognitive semantics*. de Gruyter.
- Renau, I. & Nazar, R. (2016). Automatic Extraction of Lexical Patterns from Corpora. In T. Margalitazde & G. Meladze (eds.) In *Proceedings of the XVII EURALEX International congress. Lexicography and linguistic diversity*, pp. 823-830.
- Seretan, V. (2013). A multilingual integrated framework for processing lexical collocations. *Computational Linguistics*. Springer, pp. 87-108.
- Seretan, V. (2011). *Syntax-based collocation extraction*. Springer Science & Business Media.
- Sinclair, J. McH. (1996). The search for units of meaning. *Textus: English Studies in Italy* 9(1), pp. 75-106.
- Sinclair, J. McH. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), pp. 209-243.
- Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language* 2, pp. 23-55.
- Tutin, A. & Kraif, O. (2016). From binary collocations to grammatically extended collocations: Some insights in the semantic field of emotions in French. *Mémoires de la Société néophilologique de Helsinki, Helsinki: Société néophilologique de Helsinki, 2016, Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, hal-01337486, pp. 245-266.
- Wehrli, E., Seretan, V. & Nerima, L. (2010). Sentence Analysis and Collocation Identification. In: *COLING 2010, Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, Beijing, pp. 28-36.
- Wiegand, H. E. (1996). Das Konzept der semiintegrierten Mikrostrukturen. Ein Beitrag zur Theorie zweisprachiger Printwörterbücher. In: H. E. Wiegand (Ed.), *Wörterbücher in der Diskussion II. Vorträge aus dem Heidelberger Lexikographischen Kolloquium*. Tübingen (Lexicographica. Series Maior 70), pp. 1-82.
- Zinsmeister, H., & Heid, U. (2003). Significant triples: Adjective+ noun+ verb combinations. In *Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003)*, Budapest.







# Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek

Stamou V.<sup>1</sup>, Malli M.<sup>2</sup>, Takorou P.<sup>2</sup>, Xylogianni A.<sup>2</sup>, Markantonatou S.<sup>1</sup>

<sup>1</sup> Institute for Language and Speech Processing, Greece

<sup>2</sup> Department of French Language Literature, Greece

## Abstract

We report on issues concerning the use of association measures and linguistic knowledge (Part-of-Speech sequences) with the environment MWETOOLKIT (Ramisch et al. 2010) for discovering all types of verb multiword expressions (VMWE) in corpora of Modern Greek (MG) literature. "MWE discovery" refers to detecting new MWEs in a corpus for lexicographic purposes (Constant et al. 2017). We are interested in boosting lexicographic work with (semi-)automatic facilities, in particular, the development of the VMWE database IDION (Markantonatou et al. 2019).

**Keywords:** verb multiword expression discovery; association measures; lexicography

## 1 Introduction

We discuss the behaviour of 5 association measures as regards the discovery of verb multiword expressions (VMWEs) on literature corpora of Modern Greek (MG). We used the environment MWETOOLKIT3 for this purpose. MWETOOLKIT was selected because (i) it can be applied and adapted to any language provided that lemma and pos tag information are available, (ii) the level of linguistic analysis can be specified by the user, (iii) it is a complete pipeline (includes also evaluation module), (iv) it contains lexical association measures, (v) it has been applied to several languages: English (Ramisch et al. 2010), French (Dubremetz & Nivre 2014), Brazilian Portuguese (Duran et al. 2011), Latvian and Lithuanian (Mandravickaite & Krilavičius 2017). The tool exploits either n-grams extracted from a corpus or morphosyntactic patterns defined using regular expressions in order to produce a list of candidate MWE phrases that are subsequently filtered into a significance list with the use of AMs (association measures).

MWETOOLKIT3 has already been used for the Modern Greek language to discover nominal MWEs (Linardaki et al. 2010) and light verb constructions (Stripeli et al. in prep). Both MWE types are typically bigrams. Neither approach discusses the behaviour of the AMs in relation to the grammatical patterns used probably because the short length of the particular MWEs reduces the number of possible patterns.

Some scepticism has been expressed in the literature concerning the appropriateness of AMs for discovering strings longer than bigrams (Constant et al. 2017). VMWEs quite often contain more than two words and, perhaps this is the reason why studies on the behavior of AMs as regards all types of VMWEs are sparse.

In Section 2 below we present a short review of the literature on the evaluation of the AMs with respect to MWE discovery and/or identification and then move to present our work: the corpus we developed and used is described in Section 3, Sections 4 and 5 report on two experiments and, finally, in Section 6 we summarize our conclusions.

## 2 A brief review of the relevant literature

Most of the literature related to the evaluation of the AMs used either for identification or for extraction points is about bigrams and attributes the observed differentiations to corpora diversity (size, type, etc.) or to the type and characteristics of the studied MWEs. This implies that no AM always achieves best performances and that a combination of AMs (Pecina and Schlesinger 2006) is required to guarantee best results. Furthermore, many studies emphasize the importance of manual validation as regards the exploration of the functionality of AMs in MWE discovery/identification tasks.

In Krenn and Evert (2001) AMs, namely *mutual information (MI)*, *dice coefficient*,  *$\chi^2$  measure*, *log-likelihood* and *t-score*, are utilized for detecting PP-verb collocations in the German language (newspaper and newsgroup corpus); *t-score* yielded the best precision score, while the remaining AMs yielded results below the baseline (random selection) and the simple co-occurrence frequency. Evert and Krenn (2001) evaluate lexical association measurements in German corpora against a gold set of manually detected cases. They worked with *Adj+Nouns* and *Preposition+Noun+Verb* sequences; the later sequences actually form triples, but AMs were implemented as P(reposition)N(oun),V(erb). The following AMs were used: *MI*, *LL*, *t-test* and  *$\chi^2$ -test*. For *Adj+Nouns*, *log-likelihood* and *t-test* were observed to receive higher precision values on the sets of 100 and 500 candidates, while the *MI* measure obtained the lowest precision value. In addition, co-occurrence frequency, which was also examined outperformed both  *$\chi^2$ -test* and *MI*. In the case of PN,V pairs, *t-score* and *frequency* were better than *LL*. Next, they considered different frequency thresholds (low-high values) and observed that *t-score* achieves the best results in high frequency data for the PN,V pairs, while for the low frequency data *LL* was found to get higher values. Summarizing the findings of this study, it was shown that the AMs' functionality in filtering phrases highly depends on



the MWE type, and the role of word frequency in the resulting candidate list was highlighted. English prepositional verbs (e.g. ‘*come across*’) are discussed by Baldwin (2005), in terms of extraction and methodology evaluation. Among the proposed methods for accepting a verb preposition sequence as a PV or not, along with linguistic tests, statistical criteria were used as well as *the co-occurrence frequency*, *the dice coefficient*, *the PMI*,  $\chi^2$  and *the LL*. The AMs were tested on Brown corpus, Wall Street Journal and BNC; the *dice coefficient* performed best while the other statistical criteria made better guesses than selection based on pure frequency.

Hoang et al. (2009) attempt to group 82 AMs discussed in Pecina & Schlesinger (2006) into two classes in order to better understand and justify similarities or differences among the ranking scores and guess which AMs fit better to specific MWE classes. Class I (e.g. *PMI*, *t-score*) was defined by considering the tendency of a phrase to form a semantic unit rather than being a random word combination, and Class II (e.g. *entropy*, *distance metrics*) by taking into account the context and therefore non-compositionality. The MWE examined were Verb particle constructions (only bigrams) and Light verb constructions extracted from the Wall Street Journal corpus. Ranking similarity (or ‘rank equivalence’) on the candidate lists allowed to characterize groups of AMs based on their average precision (AP) values. Interestingly, *MI* and *LL* were found to be included into the same group, while in total 5 groups of AMs were detected. Subsequently, the authors proposed a replacement of more complex AMs by simpler ones in the case they return the same AP scores. As regards the MWE type examined, they concluded that AMs belonging to Class II were not appropriate for identifying both MWE types due to the small corpus size.

In Antunes and Mendes (2014), MWEs of different types (14,000) belonging to the COMBINA-PT lexicon generated from a subcorpus of Contemporary Portuguese are used to extract example cases (n-grams:2-5). The example cases were subsequently selected and sorted according to their *MI* values. The authors focused on candidate MWE phrases with *MI* values around 8 and 10, given the previous research findings (Pereira & Mendes 2002). They checked the lists manually considering several criteria (both linguistic and quantitative). Interestingly, often phrases (n-grams) accepted in terms of their linguistic properties received very low *MI* values. *MI* was compared with other AMs such as *t-test* and *LL* with respect to their ranking preferences. Differentiations were observed for the cases ranked in the middle but not for the cases ranked high or low. In general, *LL* was found to be more similar to *MI*.

More recently, Garcia et al. (2019) evaluated twelve AMs (including *raw frequency*) on three types of dependency-based collocations, namely *adjective noun*, *verb object* and *nominal compounds* pairs for English, Portuguese and Spanish. The following AMs were included in the study: *ll*, *t-score*, *z-score*, *MI(PMI)*, *MI2*, *Dice*, *log-likelihood*, and  $\chi^2$ . Precision and Recall values were computed against gold corpora annotated with 1,394 unique collocations. *Frequency* and *t-score* achieved the best performance, while *PMI* obtained the lowest values. Similar average results were obtained for the three languages. Although *frequency* was found to be amongst the best measures for collocation extraction, the authors discuss cases missed due to their low frequency of occurrence and conclude that there is a need of “applying specific AMs for different relations and frequency folds” (Garcia 2019: 56).

### 3 The Corpus

The corpus we developed for our experiments consists of five novels (Table 1), all of them containing contemporary colloquial language rich in VMWEs. The scanned text was corrected manually and tagged/lemmatized with the ILSP tools (Papageorgiou et al. 2000).

Subcorpora	Tokens	Lemmas	Number of sentences	Number of MWEs
Maratos (2007)	84,172	5,937	5,931	1,136
Markaris (1995)	105,575	5,916	8,457	1,877
Markaris (2016)	84,172	4,86	5,931	1,175
Papadaki(2001)	82, 084	6,281	10,311	912
Tachtsis (1970)	104,215	6,701	6,700	1,554

Table 1: The five subcorpora.

Each novel in the corpus was annotated for VMWEs. The annotators read the texts and listed the text extracts they considered as instances of verbal MWEs (VMWEs), in other words, the annotators performed manual VMWE identification. The lists derived from each subcorpus were annotated by a third expert. Interannotator agreement (IA) between the corpus annotators and the third annotator was calculated with the Fleiss kappa coefficient (Table 2). A Golden Standard (GS) of about 3500 VMWEs corresponding to ~2400 (lemmatized) types was formed that contains the text extracts that were considered MWEs by both the annotators; crucially, GS does not necessarily contain all the VMWEs in the corpus.

Literature Texts	$\kappa$
Μαράτος	0.94



Μάρκαρης (Offshore)	0.87
Μάρκαρης (Νυχτερινό Δελτίο)	0.94
Παπαδάκη	0.77
Ταχτσής	0.97

Table 2: Interannotation agreement (Fleiss κ) for each subcorpus.

The annotators used the following criteria to identify VMWEs:

- Native speakers' knowledge of the language.
- The construction does not demonstrate an established sense of the verb head.
- Frequent occurrence of an expression in the corpus.
- Frequent use of the construction in the web.
- Distinguishing between MWEs and literal occurrences, which is always a matter of context (Savary et al. 2019).
- Whether the idiomatic meaning of a potential VMWE was preserved when the construction was translated literally in English (Constant et al. 2017; Salehi et al. 2018).

Figure 1 shows the distribution of VMWE lengths in the corpus. Most MWEs contain more than two fixed words. This is a potential problem for the AMs used in MWE detection that normally perform on bigrams. Actually, a large part of the literature on VMWE detection is about particle verbs and light verb constructions that are naturally modelled as bigrams (Stevenson et al. 2004). Instead, we try to identify the best method of discovering MG VMWEs of any length.

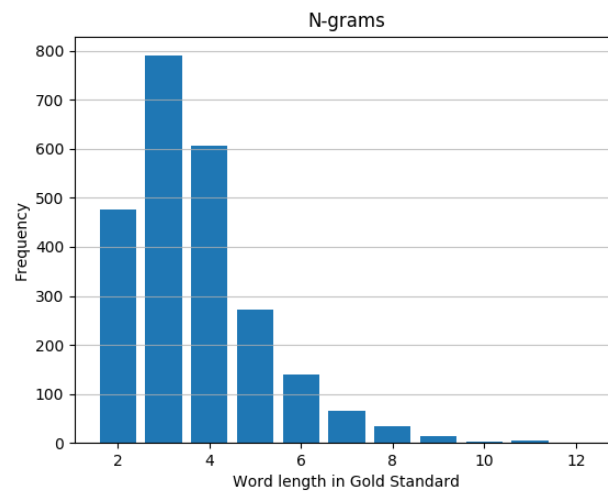


Figure 1: Distribution of VMWE lengths in terms of words in the GS.

#### 4 The First Experiment

We have already mentioned that MWETOOLKIT3 (Ramisch et al. 2010) allows for both n-grams and syntactic patterns. We defined 6 syntactic patterns, shown below as Parole (Labropoulou et al. 1996) tag sequences (Table 3), aiming at modelling a large portion of the VMWE phrase types in the corpus. We also used flat ngrams (nmin:2grams-nmax:5grams).

Patterns
(Pn)+(Vb)+Vb+(Ad)+(At)+(Aj)+No+(Pn)
Vb+Cj+Vb
(At)+No+(Pt)+(Pn)+(Pn)+(Vb)+Vb
(Pn)+No+(Pt)+(Pn)+(Vb)+Vb
(Pn)+Pn+(Vb)+Vb
(Pn)+(Vb)+Vb+(At)+(No)+(Ad)+AsPp+(At)+No

Table 3: VMWE patterns (first experiment).

In Table 3, brackets indicate optionality. The first pattern practically captures verb plus noun sequences, the second pattern subordination and the next two patterns noun plus verb sequences and the next one pronoun plus verb sequences. The last pattern captures verb plus prepositional phrase sequences. Several tenses in Modern Greek are formed periphrastically with auxiliary verbs and this fact is captured with the (Vb)+Vb sequences. Pronouns at the beginning of the strings capture the



very frequent use of dative genitives.

Candidate strings were filtered with the following AMs: *dice coefficient*, *log likelihood*, *relative frequency (mle)*, *PMI (pointwise mutual information)* and *t-score*. We report on the manual and automatic evaluation of a list of candidate VMWEs composed of the top 3000 highest ranking candidates proposed by each AM, in total 15,000 candidates; in this list, all the expression tokens sharing the same lemma form were merged under a single entry (the lemma).

We followed the same manual evaluation method as with the corpus annotation. The ranking of the AM scores by the annotators in terms of decreasing reliability is shown in Table 4.

Association Measures	$\kappa$
Dice	0.72
Log likelihood	0.69
Mle	0.67
T-score	0.60
Pmi	0.53

Table 4: AM scores ranked (by Fleiss  $\kappa$ ).

Figure 2 shows the intersection among the sets of the top best ranking 3000 expressions returned by *ll*, *mle* and *t-score*: *mle* and *t-score* overlap significantly, a phenomenon also observed in Linardaki et al. (2010).

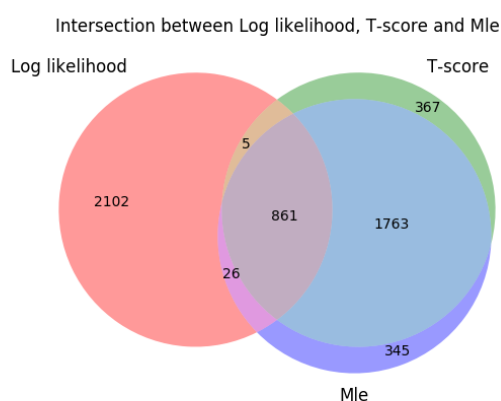
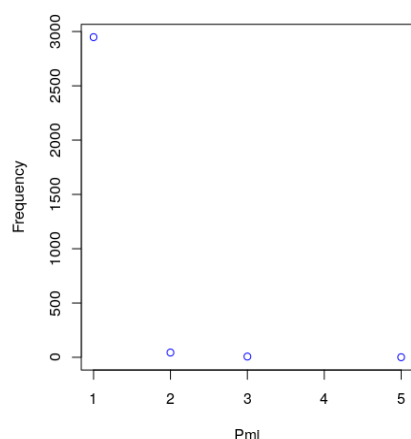
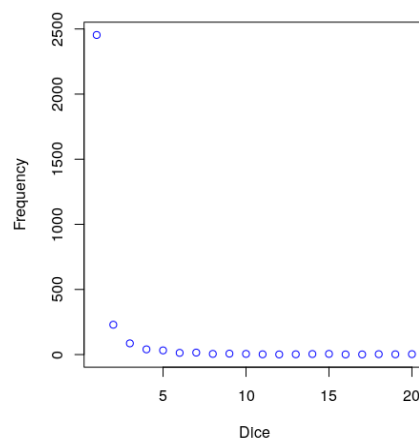


Figure 2: Intersection among the sets returned by *log-likelihood*, *mle* and *t-score*.

Qualitatively speaking, *dice* and *pmi* returned VMWEs not found by the other AMs. The plots of the frequency values for the top 3000 phrases returned by *pmi* and *dice* (Figures 3 & 4) show that most of the retrieved phrases were hapax legomena.



Figure 3: Histogram for the top 300 candidates of *pmi*.Figure 4: Histogram for the top 3000 candidates of *dice*.

Despite the annotators' agreement on *dice*'s superiority, the automatic comparison of the AM best scores against the GS returned the following decreasing list of predictions: *t-score* (302), *mle* (282), *log likelihood* (221), *dice* (172) and *pmi* (42). The Precision score with GS was 2%. Precision P is defined as  $P = TP / (TP + FP)$  where TP=true positives, FP=false positives. We also used n-grams in order to have a back-off mechanism for cases that might not be captured by the 6 patterns. We ran each n-gram (2:5) separately and evaluated the output against the GS. The best guesses were made by the *t-score* for 2grams (44 phrases) and 4grams (71 phrases). In the second experiment we did not use n-grams.

We employed a sequence matching technique using the python class *SequenceMatcher* to compare strings in the GS with the 829 manually identified "True Positives" (TP), i.e., the lemmatized VMWEs in the AMs output that were shared by all the annotators who annotated each set of results.

We applied additional computations on the results because of the limitations of the lemmatizer, for instance it misses verbal types with apostrophes (example 1) and because of the non-fixed words in a VMWE such as the possessive pronouns that do not turn up in the lemmatized version of the VMWE (example 2); e.g.

(1) του 'ψηνε το ψάρι στα χείλη

Lit. to.him roasted.3<sup>rd</sup> the fish on.the lips

'he made his life extremely difficult'

(2) χάνω το χρώμα μου

Lit. lose.1<sup>st</sup> the colour mine

'I become pale'.

We checked manually the common phrases with a ratio above 0.8 and identified 408 phrases. Therefore, an enriched GS2 was created from GS with the addition of 421 new VMWEs (TPs were 829 in total). GS2 contained ~3000 lemmatized phrases.

An automatic evaluation of the AM results against GS2 returned an improved Precision score of 4,76%. Recall values are not reported since we do not know the precise size of the VMWE population in our corpus.

## 5 The Second Experiment

In the hope of improving the results of the first experiment, we increased the amount of linguistic knowledge and adjusted the patterns to include adverbs, adjectives, double prepositional phrases and more complex conjunction patterns. The enriched patterns are shown in Table 5 where the boldfaced PoS indicate the additions to the patterns of Table 3. We followed the same evaluation method.

Patterns
(Pn)+(Vb)+ <b>Vb</b> +(Ad)+(At)+(Aj)+(At)+ <b>No</b> +(Pn)+(Aj)
<b>Vb</b> +(At)+(No)+Cj+(Pn)+(At)+(No)



<b>Vb +Cj+(Pt)+(Pt)+Vb</b>
<b>(Pn)+Pn+(Pt)+(Pt)+(Vb)+Vb+(Pn)+No+(Pt)+(Pn)+(Vb)+Vb</b>
<b>(Pn)+(Vb)+Vb+(At)+(No)+(Ad)+AsPp+(Aj)+(At)+No</b>
<b>(Pn)+(Vb)+Vb+(Cj)+(Ad)+(At)+(No)+AsPp+(Ad)+(At)+No+(At)+(No)</b>
<b>Vb+AsPp+(At)+No+AsPp+(At)+No</b>
<b>(Pn)+(At)+(No)+(Pt)+Pn+(Vb)+Vb</b>
<b>(Pn)+(Vb)+Vb+Ad+(Ad)+(Pn)</b>

Table 5: VMWEs patterns (second experiment).

As regards the manual evaluation, we computed IA agreement on the phrases proposed by each score. The ranking of the AMs by reduced reliability is: *dice*, *mle*, *pmi*, *t-score*, *log likelihood*, while the Fleiss  $\kappa$  value ranges from 0.72 to 0.86. The automatic evaluation against the GS2 resulted in a different order, as in the first experiment: *t-score* (22,9%), *mle* (21%), *ll* (12%), *dice* (7,6%) and *pmi* (0,4%).

A similar disagreement between manual and automatic evaluation is reported by Linardaki et al. (2010) for nominal MG MWES and Gurrutxaga et al. (2011) for Basque VMWEs. This differentiation could perhaps be attributed to the fact that native speakers can easily detect hapax legomena in the outputs of the AMs and promote AMs sensitive to hapax legomena, while the automatic evaluation relies on the contents of the definitely incomplete GS; therefore, AMs that bring more frequent VMWEs fare better.

True Positives for the new AM results were 1455, of which 619 in the GS2. GS3, containing ~4000 VMWE lemmata was obtained with the addition of 836 new phrases to GS2 and returned P=5,2%.

## 6 Conclusions

This work aimed at evaluating the contribution of AMs to VMWE discovery for lexicographic purposes, therefore our conclusions refer to the discovery of new VMWEs, which often contain more than two words.

So, if the lexicographic goal is the discovery of new VMWEs, our experiments have shown that as far as literature corpora of Modern Greek are concerned, human annotation is indispensable. However, the combination of linguistic knowledge with AMs can significantly improve the results received with human annotation: in our experiments human annotation returned 2400 VMWEs and the application of patterns and AMs increased their number to 4000 but only when the output of AMs was evaluated by human experts. Here, we underline the fact that automatic evaluations against a GS failed to discover less frequently used VMWEs as was discussed in Section 2.

Furthermore, patterns rich in linguistic knowledge have produced better results than leaner patterns or n-grams: experiment two used richer phrasal patterns than experiment one and returned about 26% more VMWEs.

As regards the AMs, our experiments indicated that not all the AM outputs have to be evaluated because there are significant overlaps. So, evaluation of the results of *pmi*, *dice*, *t-score* and *ll* by the annotators is necessary in order to identify interesting VMWEs because:

1. It has been observed that the results of *mle* and *dice* overlap significantly (Figure 2); therefore, only the results of one of the two AMs could be evaluated. We recommend the evaluation of the results returned by *dice*, because it fared better in both the manual evaluations (as we have already mentioned, manual evaluation of the AM results is strongly recommended for VMWE discovery).
2. *Pmi* and *dice* return VMWEs that are rare in corpora on which the AMs have been applied, as can be observed in Figures 3 & 4 (and has also been reported in literature, Church & Hanks 1990; Pereira and Mendes 2002).

Overall, if human effort has to be saved, the results of *dice* and *pmi* should be evaluated as *dice* scored always first in the IA agreement lists and *pmi* returned a greater number of rare VMWEs than any other AM.

As regards the automatic evaluation, *t-score* was found to perform best with Modern Greek VMWEs; similar findings have been reported for other languages as well (Linardaki et al. 2010; Gurrutxaga & Alegria 2011 among others). The role of the GS in the automatic evaluation can only be strongly emphasized. However, the development of an adequate GS is not an easy issue as it has been widely acknowledged and was shown by this work as well.

## 7 References in Greek

- [Maratos (2007)] Μαράτος, Τ. (2007). Οι τυφώνες ήταν γένους θηλυκού. Βιβλιοπωλείον της “Εστίας”, Ι.Α. ΚΟΛΛΑΡΟΥ & ΣΙΑΣ.
- [Markaris (1995)] Μάρκαρης, Π. (1995). Νυχτερινό Δελτίο, ΓΑΒΡΙΗΛΙΔΗΣ.
- [Markaris (2016)] Μάρκαρης, Π. (2016). Offshore. ΓΑΒΡΙΗΛΙΔΗΣ.
- [Papadaki (2001)] Παπαδάκη, Α. (2001). Βαρκάρισσα της χίμαιρας. Καλέντης.



[Tachtsis (1970)] Ταχτσής, Κ. (1970). Το τρίτο στεφάνι. ΕΡΜΗΣ.

## 8 References in English

- Antunes, S. & Mendes, A. (2014). An evaluation of the role of statistical measures and frequency for MWE identification. In *Proceedings of The Ninth International Conference on Language Resources and Evaluation–LREC 2014*, 26-31 May 2014. Reykjavik, Iceland, pp. 4046–4051.
- Baldwin, T. (2005). Looking for prepositional verbs in corpus data. In *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in computational linguistics formalisms and applications*, Colchester, UK, pp. 180-189.
- Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, 16(1), pp. 22–29.
- Constant, M., Eryigit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. & Todirascu, A. (2017). Survey: Multiword expression processing: A Survey. In *Computational Linguistics*, 43(4), pp. 837-892.
- Dubremetz, M. & Nivre, J. (2014). Extraction of nominal multiword expressions in French. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, April. Association for Computational Linguistics. Gothenburg, Sweden, pp. 72–76.
- Duran, M., Ramisch, C., Aluisio, S. & Villavicencio, A. (2011). Identifying and Analyzing Brazilian Portuguese Complex Predicates. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE2011)*. Association for Computational Linguistics, 23 June 2011. Portland, Oregon, USA, pp. 74-82.
- Evert, S. & Krenn, B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 188-195.
- Garcia, M., García-Salido, M. & Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWEWN 2019)*. Association for Computational Linguistics, Florence, pp. 49-59.
- Gurrutxaga, A. & Alegria, I. (2011). Automatic extraction of NV expressions in Basque: Basic issues on cooccurrence techniques. In *Proceedings of the workshop on multiword expressions: from parsing and generation to the real word*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 2-7.
- Hoang, H. H., Su N. K. & Kan M.-Y. 2009. A re-examination of lexical association measures. In *Proceedings of the Workshop on Multiword Expressions*, Singapore, pp. 31–39.
- Krenn, B. & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, Toulouse, France, pp. 39-46.
- Labropoulou, P., Mantzari, E. & Gavrilidou, M. (1996). *Lexicon-morphosyntactic specifications: Language specific instantiation (Greek)*. PP-PAROLE, MLAP report, pp. 63–386.
- Linardaki, P., Ramisch C., Villavicencio, A. & Fotopoulou, A. (2010). Towards the construction of language resources for greek multiword expressions: Extraction and evaluation. In *Proceedings of the international conference on language resources and evaluation*, May. Valetta, Malta, pp. 31-40.
- Mandravickaite, J. & Krilavičius, T. (2017). Identification of multiword expressions for Latvian and Lithuanian: hybrid approach. *Proceedings of the 13th Workshop on Multiword Expressions (MWE '17)*, Valencia, Spain, Association for Computational Linguistics, Stroudsburg, PA, pp. 97-101.
- Markantonatou, S., Minos, P., Zakis, G., Moutzouri, V., & Chantou, M. (2019). Idion: A database for Modern Greek multiword expressions. In *Proceedings of joint workshop on multiword expressions and wordnet (mwe-wn 2019)*, workshop at acl 2019. Toulouse, France: Association for Computational Linguistics, pp.130-134.
- Papageorgiou, H., Prokopidis, P., Giouli, V. & Piperidis, S. (2000). A unified POS tagging architecture and its application to Greek. In *Proceedings of the second international conference on language resources and evaluation (LREC'00)*, May 2000. Athens, Greece: European Language Resources Association (ELRA), pp. 1455-1462.
- Pecina, P., & Schlesinger, P. (2006) Combining association measures for collocation extraction. In *Proceedings of the 21th international conference on computational linguistics and 44th annual meeting of the association for computational linguistics (COLING/ACL 2006)*. Sydney, Australia, pp. 651–658.
- Pereira, L. & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. In *Proceedings of the 10th International Congress of the European Association for Lexicography*. Copenhagen, Denmark, vol. II, pp. 841-849.
- Ramisch, C., Villavicencio, A. & Boitet, C. (2010). mwetoolkit: a Framework for Multiword Expression Identification. In N. Calzolari et al. (Eds.), *Proceedings of LREC 2010*. Valetta, Malta: ELRA, pp. 662–669.
- Salehi, B., Cook, P. & Baldwin, T. (2018). Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary and Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, Berlin: Language Science Press, pp. 343–373.
- Savary, A., Cordeiro, S.R., Lichte, T., Ramisch, C., Iñurrieta, U. & Giouli, V. (2019). Literal Occurences of Multiword Expressions: Rare Birds That Cause a Stir. In *The Prague Bulletin of Mathematical Linguistics*, 112, pp. 5-54.
- Stevenson, S., Fazly, A. & North, R. (2004). Statistical measures of semi-productivity of light verb constructions. In *Proceedings of the workshop on multiword expressions: Integrating processing*. Barcelona, Spain: Association for Computational Linguistics, pp. 1-8.
- Stripeli, E., Prokopidis, P. & Papageorgiou, H. (in prep.). Multiword expressions in Greek, deltio epistimonikis orologias ke neologismou. *Academy of Athens*, 15(4), pp. 75-95.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Historical and Scholarly Lexicography and Etymology**







# Creating a DTD template for Greek dialectal lexicography: the case of the Historical Dictionary of the Cappadocian dialect

Karasimos A.<sup>1</sup>, Manolissou I.<sup>1</sup>, Melissaropoulou D.<sup>2</sup>

<sup>1</sup> Academy of Athens, Greece

<sup>2</sup> Aristotle University of Thessaloniki, Greece

## Abstract

This article reports on the compilation of a full dictionary, both print and digital, of Cappadocian Greek, one of the major Modern Greek dialects. This bilingual (Cappadocian Greek-Standard Modern Greek) dictionary is one of the products of the ‘DiCaDLand’ dialectological project, funded by the Hellenic Foundation of Research and Innovation (<http://cappadocian.upatras.gr/en>). Its compilation is based on the powerful professional dictionary editing software TLex Suite, after extensive parameterization in order to meet the needs and the particularities of both the project and the dialectal variety in study. More specifically, we present a sophisticated and state-of-the art e-lexicographic annotation template capable of handling and describing the complex data of an obsolescent and “aberrant” dialect, without written tradition, heavily influenced by language contact (with Turkish), and presenting considerable variation and serve as a model for future approaches to Greek digital lexicography.

**Keywords:** e-lexicography; dialectology; historical e-dictionary; Cappadocian Greek

## 1 Introduction

An increasing number of scholarly publications and media reports have described, in the last few years, an ongoing decline in linguistic variation on a cross-linguistic level, a phenomenon known as language death. Similar gloomy prognostics emerge for the majority of Modern Greek dialects. The status of geographically determined linguistic variation is changing rapidly under the influence of mobility, migration, and the role of mass media. Thus, the documentation and analysis of existing variation is becoming crucial from the point of view of both linguistic analysis and the history of culture. Furthermore, documentation and analysis of non-standard forms within a language continuum allows for further predictions concerning the notion of language and the limits of linguistic variation and change. In this spirit, the preservation of the Asia Minor Greek linguistic heritage through innovative products has become vital. The term “innovative products” refers here to the digital tools, procedures and methodologies which, although gradually starting to be implemented in Cultural Heritage preservation, have not yet become entrenched in this domain in Greece, nor has a scientifically significant role been acknowledged for them. The combination of state-of-the-art Digital Humanities tools with linguistic research on unique and endangered cultural data forms the core of the project presented. The use of digital tools and methodologies is of major importance for the preservation of the Greek dialectal landscape as well, for which no extended use of digital humanities tools has been reported (see however Galiotou, Karanikolas, Manolissou et al. 2014; Galiotou, Karanikolas & Ralli 2018; Melissaropoulou et al. 2015; Themistokleous et al. 2012). The project DiCaDLand, funded by the HFRI (ΕΛΙΔΕΚ), aims at the thorough documentation and study of an Asia Minor Greek linguistic variety, more specifically the Cappadocian dialect. Apart from primary linguistic research, one of the main objectives of the project is to also produce two state-of-the-art major reference works, namely an interactive electronic dialectal atlas (which will constitute the first such effort in the domain of Greek Linguistics) and a comprehensive historical dictionary of the Cappadocian dialects (again lacking until now). The purpose of this paper is to describe the electronic Historical Dictionary of the Cappadocian Dialects (background, aims, methodology, implementation).

The article is structured as follows: Section 2 describes the shift of Greek dialectology in the digital era and its major steps from a lexicographic perspective. Section 3 offers a brief description of the corpus on which the Dictionary is based. Section 4 provides an analysis of the infrastructure of the project, i.e. the DWS used to compile it and the parameterized DTD especially constructed for the purposes of the DiCaDLand project. Section 5 summarizes our results.

## 2 Background

### 2.1 Greek dialectology in the digital era

#### 2.1.1 Dialectal lexicography and e-lexicography in Greece

In current practice, there is an operational-/pre-theoretical distinction between Modern Greek dialects proper, and Modern Greek patois, depending on the degree of “aberrance” from Standard Modern Greek, and on the consequent possibility of mutual intelligibility between standard and dialect speakers. To the first category belong the following dialects: Pontic, Cappadocian (including Pharasa and Silli), South Italian (Grico-Grekanico), and Tsakonian, and to the



second all the other dialectal varieties of Greek (Cypriot, Cretan, Northern, Cycladic, Old Athenian etc.). Academic-level dialectal dictionaries exist for all major dialects *except* Cappadocian, namely: Papadopoulos (1955-1958) for Pontic, Karanastasis (1984-1992) for S. Italian, and Kostakis (1986-1987) for Tsakonian, while for other dialects of Greek no lexicographical works of similar high academic quality exist, except the still on-going *Historical Dictionary of Modern Greek* of the Academy of Athens (Manolessou & Bassea-Bezantakou 2017; for an overview of MG dialectal lexicography, see Katsouda 2012). This means that Cappadocian is the only major MG dialect that has not yet been investigated on a lexicographical level, something which constitutes an important lack that is constantly felt in the field of MG dialectology. Up to now, this gap is being filled by the smaller “glossaries” used as an appendix to grammatical descriptions of Cappadocian as a whole (Dawkins 1916: 580-663) or of individual Cappadocian dialects (e.g. Andriotis 1948; Costakis 1963; Mavrochalyvidis & Kesisoglou 1960), or by amateur lexicographical endeavours (e.g. Kotsanidis 2006). However, none of them includes the very rich dialectal material that has only recently surfaced, from current 3<sup>rd</sup> generation native speakers, thanks to the research by M. Janse and D. Papazachariou (see below 2.2). Furthermore, these glossaries do not adopt a unified system of presentation (graphematic representation – principles of lemmatization – system of semantic analysis and sense ordering), and can only give a fragmentary picture of the dialect, accessible only with great difficulty even to the specialist reader (scattered in many publications, with non-obvious cross-glossary correspondence between entries).

In general, Greek dialectal lexicography, until the end of the '80s, was mostly in the hands of non-professional linguists, and its aims were more cultural-folkloristic than purely academic-linguistic. There was also a special focus on the diachronic dimension, as it was felt that evidence of the “archaicity” of a dialect and of a closer connection to Ancient Greek would add validation and prestige to an otherwise culturally threatened and undervalued variety. Synchronic and electronic dialectal lexicography are in their first steps, although they could provide an answer to the danger of extinction faced by many Greek dialectal varieties, as well as the lack of adequate funding for large-scale print lexicographical projects.

In the domain of dialectal e-lexicography, most attempts are limited to the digitisation of already extant print dictionaries (retro-digitisation) through scanning, with, and more usually without OCR (due to the special difficulties presented by the Greek alphabet and its various symbols, see below section 4.3.1). Several of these older print dictionaries or glossaries can be found, for example, in digital depositories like the Internet Archive ([www.archive.org](http://www.archive.org)) or especially for Modern Greek, Anemi (<https://anemi.lib.uoc.gr/>). Simple digitisation of a print dictionary, however, is a practice which is still quite distant from the basic principles and presupposition of true electronic lexicography. As more complex attempts in the direction of dialectal e-lexicography, all very recent, one may mention the following (see Karasimos 2019 for an overview): a) the online dictionary “Syntyshe” for the Cypriot dialect, offered by the University of Cyprus (<http://lexcy.library.ucy.ac.cy/sintixies.aspx>, see Katsoyannou & Armotistis 2019), which offers a digital lemma list with search capabilities, as well as sound clips for each word, created through speech synthesis b) the digital tridialectal dictionary of Pontic, Cappadocian and Aivaliot offered by the University of Patras (Karanikolas et al. 2013; Ευδόπουλος, Δημελά, Μελισσαροπούλου, Παπαναγιώτου & Πάλλη 2015), which offers a sample of 7500 entries from 3 Asia Minor dialects, with complex search capabilities and a wealth of lexicographic information supported by digitised written or oral documentation and c) the retro-digitisation of the *Historical Dictionary of Modern Greek*, both of the Standard Language and the Dialects, offered by the Academy of Athens (<http://repository.academyofathens.gr/kendi/index.php/gr>, see Manolessou & Katsouda forthcoming), which again offers multiple search capabilities for all the entries (A-D) of the print dictionary, as well as the possibility to download a pdf version of a page or a whole volume.

## 2.2 Cappadocian Greek: a short overview

The Cappadocian dialect constitutes a special case within Modern Greek dialectology. First of all, it is the most highly differentiated dialectal variety of Greek, due to the very long time of separation from evolutions involving the rest of the Greek-speaking world (11<sup>th</sup> c.), and to the very strong influence of Turkish. The dialect is often employed in the literature as a prototypical example of ‘heavy borrowing’ in terms of Thomason & Kaufman’s borrowing scale, with reference to ‘overwhelming long-term cultural pressure’ (Thomason & Kaufman 1988: 50). As a result, it presents a high number of unsolved theoretical and diachronic-historical problems, of great linguistic interest, and has attracted international attention (see e.g. Thomason & Kaufman 1988: 215-222; Johanson 2002:104; Winford 2010; Ralli 2009). Secondly, it is a dialect which was until recently considered extinct, after the end of the life-span of the 1<sup>st</sup> generation speakers, relocated in mainland Greece following the compulsory exchange of populations in 1923. However, in recent years it has been discovered (see Janse 2009) that the dialect is still very much alive, retained by 3<sup>rd</sup> generation speakers in several villages in Thessaly, Macedonia and Thrace and therefore, a wealth of data concerning it, not available to earlier research and not taken into consideration in previous studies of the dialect, is for the first time available. The combination of the factors described above entail that research on Cappadocian is at the forefront of modern dialectology in Greece and abroad (see e.g. Karatsareas 2013; Melissaropoulou 2016; 2019a; 2019b; Ralli 2009; Galiotou et al. 2014). Cappadocian has also formed part of two recent major research projects undertaken by the University of Patras, which had as their aim to collect and preserve as much of the old (written) and new (oral) material as possible (Ralli 2015).

Cappadocian had been under Turkish influence from the 11<sup>th</sup> century until 1923, namely until the exchange of populations that followed the treaty of Lausanne, when spoken in Asia Minor (today's central Turkey), in an area that covered approximately thirty-two communities. From that period, it was spoken in a situation of regressive bilingualism, since Turkish was the dominant language of the political authorities and was spoken by the overwhelming majority of the population in all aspects of life (cf. Vryonis 1971: 457–59). The dialect is subdivided into two basic groups, North and



South Cappadocian (cf. Dawkins 1916) and an intermediate one, namely Central Cappadocian (cf. Janse forthcoming) showing intra-dialectal divergence. The different zones reflect, following Dawkins (1916: 209–211), different degrees of Turkish influence, which can be attributed to the large extension of the area in combination with other demographic and geographic factors, the most prominent of which would have been the presence of the Turkish population in each different community as well as the existence of Greek schools.

### 3 Corpus of data

The sources of the DiCadLand dictionary are twofold: on the one hand, all the available written sources starting from the 19th century onwards (dictionaries, glossaries, linguistic descriptions, collections of primary texts such as folk-tales, songs, narrations, riddles etc.), most of which were collected and digitized thanks to an earlier project, AMiGre, and are available online (<http://amigredb.philology.upatras.gr/>) were taken into account, in order to allow the systematisation of all existing intra-Cappadocian variation. Considerable effort was expended for the homogenization of this material, which came in a multiplicity of transcription systems, especially the older sources. On the other hand, the corpus was considerably enlarged through the addition of oral recordings. A few date back to the 1930s (available online through the depository Gallica of the Bibliothèque Nationale de France), but most of them derive from the very recent and rich dialectal material from current 3<sup>rd</sup> generation native speakers (descendants of Cappadocian refugees) collected the last decade. Special emphasis was placed on the exploitation of this new oral material, so that the dictionary under preparation will not simply be a depository of already available but disparate data, but an opportunity for the presentation of new data. This allows also for a “diachronic” examination of the evidence, as we have the possibility to examine side-by-side data which may be divided by more than 100 years. However, the dichotomy between written and oral sources has given rise to a major problem: whereas older sources are roughly equally distributed with respect to geographical provenance (i.e. data is available for almost all Cappadocian settlements, ca. 20 in number), oral data, from current speakers, are available only from 2-3 major communities, and mostly from that of Misti, which was the largest. This creates an imbalance in the lexicographical treatment of words, phenomena and senses. Another issue requiring special attention is the fact that the older material was in part collected by amateurs, or at a time when linguistic science had not yet been sufficiently developed, and therefore it is to a certain extent less reliable than the oral material, containing many inaccuracies which can no longer be assessed.

### 4 TLex software parameterization in the service of Greek dialectal lexicography

The electronic availability of well-organized lexical material is of quintessential importance for the transformation of dialectal comparative linguistics into a quantitative and collaborative field of research. To this end, the format of the Dicadland dictionary attempts to conform to standards of state-of-the-art academic-level Dictionary Writing Systems, after careful evaluation of available options.

Its realization requires adherence to the most recent advances in the domain of electronic lexicography and dialect mapping on the one hand (see e.g. Granger & Paquot 2012), and historical and dialectal lexicography on the other (see e.g. Reichmann 2012; Manolessou 2016), filtered by digital humanities methodologies.

#### 4.1 DWS and selection criteria

The Historical Dictionary of Cappadocian Greek is being built using the powerful professional dictionary editing software TLex Suite, one of the most widely-used state-of-the-art DWS internationally (for an overview of DWSs see Abel 2012 and for the main features of TLex and its application to dialectal data see Joffe, McLeod & de Schryver 2008; de Schryver 2011). The DTD construction is aided by the experience acquired by the research team in smaller-size and scope dialectal e-dictionaries, such as the tri-dialectal Asia Minor dictionary (Galiotou, Karanikolas, Manolessou et al. 2014; Galiotou, Karanikolas & Ralli 2018). From a typological viewpoint, this dictionary is being structured as a bilingual one due to the Cappadocian dialect’s considerable distance from the standard form of Modern Greek (among others, Geeraerts 1989: 294-295; Bejoint 2000: 39; Marellò 2004:351).

One may consider that there are three available DWS construction options (Krek, Abel, Tiberius 2015), namely: (a.) purchasing a commercial off-the-shelf software/app/platform, (b.) using a free app or open-source web-based platform, and (c.) building from scratch a tailored app/software for one’s own needs. As far as the first option is concerned, the most widely used and tested applications are the following, although a few of them are currently no longer available (see Abel 2012): IDM DPS (Digital Publishing System) <https://www.idmgroupp.com/content-management/dps-info.html> (Grundy & Rawlinson 2016), TLex (TshwaneLex) <https://tshwanedje.com/tshwanelex/> (de Schryver 2007; 2011), iLex (Erlandsen Media Publishing) <http://groupbanker.dk/generic-en/index.htm> (Erlandsen 2010), and ABBYY Lingvo Content <http://www.lingvo.ru/content/> (Kuzmina & Rylova 2010). In the second category, one may find quite a large number of freely available programs online, but most are meant for relatively simple lexicographical projects and therefore cannot match the potential and range of professional commercial applications. Some of the free programs with the best capabilities are: Lexonomy (<https://www.lexonomy.eu/>), Dictionary Editor and Browser (<https://deb.fi.muni.cz/index.php>), Matapuna (<https://sourceforge.net/projects/matapuna/>), Dictionary System DWS (<http://dictionary-system.hvalur.org/index.php?lang=en>), and WeSay <https://software.sil.org/wesay/> (Perlin 2012). As for the third category, many academic and research institutions have opted for the construction of a custom-made DWS, usually based on an already extant general database construction program (e.g. Oracle/MySQL/Filemaker) or a general XML editor (e.g. Oxygen, XMetal), parameterized for lexicographic use. Examples include the DWS of the Institute of



Dutch Lexicology for its collection of historical and local dictionaries (Tiberius, Niestadt & Schoonheim 2014), and the DWS of Institute of Czech language at the Academy of Sciences of the Czech Republic (Barbierik et al. 2014).

#### 4.1.1 Selection criteria

The selection criteria between the three alternative options may be summarised as follows:

- a) Cost: Freely available online programs prove superior on the basis of this criterion, as their cost is literally zero. Commercial projects are obviously the most expensive ones, but custom-made programs also come at considerable cost, since the general programs on which they are based are also commercial ones. Cost also depends on the number of licences to be purchased, as well as the option between personal or institutional use. Some programs (such as TLex) offer a special discount for lexicographic work on endangered languages or varieties.
- b) Time: commercial programs fare the best with respect to the time factor, as the long-term experience behind them and their large professional support teams have already solved most of the challenges in the domain of e-lexicography. At the other end of the spectrum, custom-made programs are the most time-consuming, since everything needs to be rebuilt from scratch. Free online programs constitute an intermediate solution, since, as they are designed to serve a wide variety of research aims with various specifications, they may require a high degree of parameterization.
- c) Capabilities: Commercial programs are the ones to offer the widest range of capabilities as compared to free programs. This concerns not only available functions and greater degree of customisation, but also higher storage capacity, easier interconnection with other software programs, and better support (manuals, tutorials, updates). The only great advantage of free programs is that they are web-based, something which allows parallel processing by several users without need for a server or special installations. Custom-made programs are obviously the ones which allow for the widest range of specialised capabilities, limited only by the time and the funds one is willing to invest in their construction.
- d) User-friendliness: with respect to this criterion, each category has its own advantages and disadvantages. More specifically, free programs, being in general simpler and with fewer capabilities, usually come with a relatively uncomplicated and intuitive interface and fewer interactive buttons than commercial programs, which may have quite a complex GUI. On the other hand, commercial programs may be able to perform automatically a great number of routines which may either be impossible in a free program, or may need to be executed manually with multiple repetitions, or with special programming scripts. As a simple example, for text formatting commercial programs offer an environment similar to that of specialised word-processing applications, with a wide variety of options for fonts, sizes, colours, symbols etc., whereas free programs usually have a built-in predetermined and limited range of options, or require the construction of a Cascading Style Sheet (CSS). Furthermore, professional programs standardly allow for multiple viewing alternatives of the dictionary under construction: in database format, in XML format, or, more importantly, as final exported text or front-end interface (with a WYSIWYG window), so the dictionary compiler immediately realises the impact of his choices. In free programs one usually sees only the back-end interface in database view, and the final (usually html) outcome becomes visible only after the completion of the project. Finally, as also mentioned above, free programs cannot offer the strong user support provided by commercial applications, as they are usually unable to go beyond a FAQ page or a users' forum. As far as custom-made programs are concerned, anything is possible, but previous experience has demonstrated that the more special functions a program has, the more complex it becomes, and without the services of a professional software engineer the final product has few chances of being user-friendly.

#### 4.1.2 Proposal

On the basis of the above, it was deemed that the construction of a custom-made program *ab initio* should be avoided. On the one hand it would be extremely time-consuming, given that the DiCaDLand project needs to be completed in a specific time-frame (3 years), and on the other it would force the research team in a direction beyond the main aims of the project, which are primarily linguistic rather than technical. Furthermore, given that the main issues and requirements of electronic lexicography are similar worldwide, any solution arrived at would most probably be a "re-discovery" of already solved problems (a 're-invention of the wheel' in the terms of de Schryver 2007; 2011). Also, the construction of a custom-made electronic tool would hardly be a cost-effective option, as on the one hand the general programs required for its construction (database server, xml writer etc.) would not come free of charge and on the other a technical specialist would need to be hired. After extensive testing, the possibility of using a freely available online program was also rejected, as the options they offered could not cover the range of needs and specifications of the DiCaDLand dictionary.

#### 4.2 Why TLex?

After comparative evaluation, it was decided to construct the Historical Dictionary of the Cappadocian dialects using the powerful TLex suite, which presents several advantages (cf. De Schryver & De Pauw 2007; De Schryver 2011). More specifically, it has already been tried and tested on more than 40 national and academic lexicographic projects, it offers a surprisingly wide range of parameterization for even the most "non-standard" linguistic varieties (including endangered languages), and it offers the possibility of direct export to Microsoft Word (in .rtf format), or, even better, to professional desktop publishing programs such as Indesign and QuarkXpress. Additionally, it offers the possibility of import/export from spreadsheets such as Excel, allowing us to automatically import material previously collected in such formats and very importantly, it provides its own integrated Corpus Query system, which can be used for example extraction on the basis of the transcribed oral corpus.



Furthermore, as de Schryver (2011) has already pointed out, and as discussed above, the TLEX suite, as any other dedicated DWS software as compared to custom-made solutions, guarantees reduced project completion time, thanks to (amongst others): various levels of automation (e.g. automatic cross-reference tracking and updating of homonym and sense numbers, easy entry of any phonetic symbol through macros, fast full-dictionary text search, automatic checking for various dictionary errors, immediate WYSIWYG, full Unicode support, customisable styles (font, colour, etc.) for every field in the dictionary and language of the metalanguage (cf. De Schryver & Joffe 2005a), increased consistency in the treatment of articles, thanks to features such as the article filter, and finally improved teamwork and easy multi-user adaptation.

### 4.3 Creating and parameterizing a DTD for the Cappadocian Dictionary

The DTD of the Cappadocian Dictionary under implementation was constructed through extensive parametrization and customization of the TLEX software, presented in the following subsections. Although the TLEX Suite comes with a built-in template both for monolingual and bilingual dictionaries, it was deemed necessary for our project's needs to construct a new DTD, which would be better adapted to the needs of modern Greek dialectology. The construction and compilation of this dictionary benefits from the established research and the expertise of research conducted at the Research Centre for Modern Greek Dialects of the Academy of Athens for the publication of print dialectal dictionaries, with special adaptations in order to meet the needs of a specific Asia Minor dialectal group as well as digitization.

Element types	Class	List item	Internal ID	List item	Internal ID
Dictionary	Document	Ανακού	44	Βαγρ.	73
Language	Section/L	Αξός	29	Cost.	70
Lemma	Entry	Αραβανί	42	Dawk.	62
Etymology	User	Αραβισσός	48	Dawk.Song.	63
Sense	User	Αφσάρι	49	Greg.	74
Example	User	Γαριπτσός	50	Karats.	75
Subentry	Subentry	Γούρδονος	43	Lag.	64
Definition	User	Δίλα	51	Αλεκτ.	77
TE (Translation Eq...	User	Κίσκα	53	Αναστασ.	78
References	SmartRef	Καπαδοκία	41	Ανδρ.	67
Form	User	Καρατζάβιραν	52	ΑΠΥ-Βαγρ.	129
POS	User	Καρατζάρεν	130	ΑΠΥ-ΑΠΘ	138
Example_phrase	User	Μαλακοπή	20	ΑΠΥ-ΕΝΔ	127
Example_Paraim	User	Μισθί	18	ΑΠΥ-Καρατσ.	128
Example_Riddle	User	Ουλαγάς	28	Αρχέλ.	79
Example_Song	User	Ποτάμια	26	ΕΚΠΑ 2142	95
Period	User	Σύλλα	45	Ελευθερ.	80
Synonym	User	Σατί	54	Θεοδ.Παραδ.	81
Comment	User	Σεμέντρα	47	Θεοδ.Τραγ.	82
Forms	User	Σινασσός	21		
Senses	User	Τελμηςσός	27		
		Τζαλέλα	56		

Figure 1: Controlled vocabularies from the proposed DTD schema including geodata and sources

In more detail, all major Greek dialectal dictionaries share the same tripartite structure, following the model of the *Historical Dictionary of Modern Greek* (ILNE) of the Academy of Athens, the earliest and largest Greek scientific lexicographic endeavour (Manolessou & Bassea-Bezantakou 2017; Manolessou & Katsouda forthcoming). This structure comprises, apart from the headword, (i) a Forms field, where the variant dialectal forms are set out, with phonetic transcription and geographical distribution (ii) an Etymology field, where the word's origin (native/loanword) and dating are recorded and (iii) a Senses field, with definitions, examples, quotations and documentation from oral and written corpora, and also including "special" types of examples, such as proverbs, songs and riddles.

Additionally, parameterization involved also issues of alphabetic representation and phonetic transcription, as well as the incorporation of extensive bibliographic references and cross-references, as is expected in a fundamentally academic publication. Finally, the parameterization needed to take into consideration the necessity of a parallel print and digital publication and the combination with an online digital dialectal atlas also under preparation by the same project (<http://cappadocian.upatras.gr/en/node/10>), which among other things entailed the switching of the software's meta-language to Greek and the preparation of alternative attribute lists (full vs. abbreviated).

#### 4.3.1 Alphabet, script, font and encoding

The Greek alphabet and the multiple non-standard alphabetic symbols used in Greek dialectal literature for the encoding of phonetic features absent from Standard Modern Greek (e.g. σ = [ʃ], α̃ = [æ]; for an overview see Manolessou, Beis & Bassea-Bezantakou 2012). From a typographical/orthographic point of view, Standard Modern Greek can (and is) very easily rendered through the Greek alphabet, which has a standardized orthography and suffices to represent all the sounds of the standard phonological system. Correspondingly, when it comes to electronic transcription/encoding, Standard







free-text form, the word's etymological provenance (inherited, loanword etc.), its dating (ancient, medieval, modern etc.), its morphological analysis (stem, suffix etc.) and possibly, bibliographic references and cross-references, as well as quotations from old (ancient, medieval) textual sources for documentation purposes. TLex does not provide the possibility to “embed” XML objects within a free text.

#### 4.3.4 Senses

As in the case of “Forms” above, the element “Senses” has the option of multiple embedded sub-senses, with several attributes, following standard lexicographical practice. This possibility was already provided for in the built-in TLex templates. Similarly, the standard templates included basic fields such as “definition” and “example”. The parameterization implemented for our dictionary centered around the presentation of examples. The simple “example” was expanded through several sub-categories. These were “phrases” (idioms and collocations), “proverbs”, “songs” and “riddles”. Each of these categories required special treatment: apart from their literal translation, they also need an extra field for their interpretation (meaning of the proverb, solution of the riddle etc.), and some required special formatting (verses, in the case of songs and occasionally riddles). Furthermore, each definition and each example is followed, as in the case of forms, by geographical information. Finally, all senses are dated as to their first appearance (ancient, medieval etc.) through a controlled list.

The screenshot displays the TLex interface for a lemma entry. The main window shows a list of senses for the lemma 'Senses'. Each sense is numbered and includes a definition, examples, and geographical information. The interface is divided into several panes: a left sidebar with navigation options, a central pane for the lemma entry, and a right pane for detailed examples and references. The central pane shows the lemma 'Senses' with a list of senses. The right pane shows a detailed view of a sense, including its definition, examples, and geographical information. The interface is designed to be user-friendly and accessible, with clear navigation and a structured layout.

Figure 3: Lemma sample with an extensive use of different kinds of examples to document the senses

## 5 Concluding Remarks

The construction of the electronic Historical dictionary for the Cappadocian is meant to contribute significantly to the thorough documentation and study of this dialect, which holds a prominent position not both in Greek dialectology and in international language contact studies. The proposed template provides contributions in several scientific fields. More specifically, in Greek Linguistics, this study constitutes a much-needed holistic approach to the Cappadocian dialect. In *Lexicography*, it produces an online, freely accessible dictionary, following the latest international standards for electronic lexicography (DWS platform, .xml output) and dialectal lexicography, for the first time in Greece. It provides a standard system of transcription, fully compatible with the Unicode standard while at the same time following the principles of the IPA; it implements, for the first time in Greece, new methodologies in dialectal lexicography, applying in them on a dialect possessing many special or unique features rendering lexicographical treatment extremely difficult (e.g. loss of gender, morphological opacity of the vocabulary due to long-term isolated evolution, high level of foreign influence, unreliable primary sources requiring constant re-evaluation). Finally, this effort constitutes an important step for the development of Digital Humanities in Greece and salvages a significant amount of cultural data from an endangered dialect, which is of great value for Greek civilisation and historical memory.

## 6 References

Abel, A. (2012). Dictionary Writing Systems and Beyond. In S. Granger and M. Paquot (Eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 83-106.



- Abel, A., Krek, S., Tiberius, K. (2015). *Bibliography – Dictionary Writing Systems (DWS) & related software*. ENEL Cost Action documentation, available at [http://www.elexicography.eu/wp-content/uploads/2015/04/Bibliography\\_DWS\\_CQS\\_v7\\_web.pdf](http://www.elexicography.eu/wp-content/uploads/2015/04/Bibliography_DWS_CQS_v7_web.pdf) [accessed 05/11/2018]
- Andriotis, N. P. (1948). *Το γλωσσικό ιδίωμα των Φαράσων* [The dialect of Pharasa]. Athens: Ikaros.
- Armstrong S., Christodoulou K., Katsoyannou, M. & Themistocleous, C. (2014). Addressing writing system issues in dialectal lexicography: the case of Cypriot Greek, In Dyck *et al.* (Eds), *Dialogue on Dialect Standardization*. Cambridge: Cambridge Scholars Publishingpp, pp. 23-38.
- Barbierik, K., Dēngeová, Z., Holcová Habrová, M., Jarý, V., Liška, T., Lišková, M., Virius, M. (2014). Simple and Effective User Interface for the Dictionary Writing System. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pp. 125-136.
- Béjoint, H. (2000). *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.
- Bowers, J. & L. Romary (2016). Deep Encoding of Etymological Information in TEI. *Journal of the Text Encoding Initiative [Online]*, Issue 10 | 2016. URL: [<http://journals.openedition.org/jtei/1643>] [accessed 03/08/2019]
- Chambers, J.K. & P. Trudgill. (1980). *Dialectology*. Cambridge: CUP.
- de Schryver, G-M. (2007). *Oxford Bilingual School Dictionary: Northern Sotho and English*. Cape Town: Oxford University Press Southern Africa.
- de Schryver, G-M. (2011). Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters. In K. Akasu and S. Uchida (Eds.), *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the Seventh ASIALEX Biennial International Conference*. Kyoto, Japan, August 22-24, 2011 Kyoto: The Asian Association for Lexicography, pp. 647–656.
- de Schryver, G-M & De Pauw, G. (2007). Dictionary writing system (DWS) + corpus query package (CQP): The case of Tshwane Lex. *Lexikos*, 17, pp. 226–246.
- de Schryver, G-M, Joffe, D., & M. MacLeod (2008). The TshwaneLex Electronic Dictionary System (Software Demonstration). In E. Bernal and J. Decesaris (Eds.), *Proceedings of the XIII Euralex International Congress*. Barcelona: IULA, Documenta Universitaria, pp. 421-424.
- Erlandsen, J. (2010). iLEX, a general system for traditional dictionaries on paper and adaptive electronic lexical resources. *Proceedings of the 14th EURALEX International Congress*, pp. 306-316.
- Galiotou E, N. Karanikolas, I. Manolessou *et al.* (2014). Asia Minor Greek: Towards a computational processing. In *Procedia- Social and Behavioral Sciences 147*, pp. 458 – 466. Elsevier.
- Galiotou E., Karanikolas N. & A. Ralli. (2018). Preservation and Management of Greek Dialectal Data. In Ioannides, M., E. Fink *et al.* (Eds.) *Digital heritage: progress in cultural heritage: documentation, preservation and protection. 7th International Conference Euromed 2018*. Berlin/Heidelberg: Springer, pp. 752- 761.
- Dawkins, R. M. (1916). *Modern Greek in Asia Minor: a Study of the Dialects of Silli, Cappadocia and Phárasa with Grammar, Texts, Translations and Glossary*. Cambridge: Cambridge University Press.
- Geeraerts, D. (1989). Principles in monolingual lexicography. In F. J. Hausmann, O. O. Reichmann, H. E. Wiegand & L. Zgusta (Eds.), *Wörterbücher/ Dictionaries/ Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie*. Berlin: de Gruyter, pp. 287-296.
- Granger, S. & M. Paquot. (2012). *Electronic Lexicography*. Oxford: Oxford University Press.
- ILNE, (1933-). *Ιστορικόν Λεξικόν τῆς Νέας Ἑλληνικῆς, τῆς τε κοινῆς ὁμιλουμένης καὶ τῶν ιδιωμάτων* [Historical Dictionary of Modern Greek, both of the Standard language and the dialects], Athens: Academy of Athens.
- Grundy, V. & D. Rawlinson. (2016). The Practicalities of Dictionary Production; Planning and Managing Dictionary Projects; Training of Lexicographers. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 561-578.
- Janse, M. (forthcoming). Καππαδοκικά [Cappadocian]”. In C. Tzitzilis (ed.), *Ἡ ελληνική γλώσσα καὶ οἱ διάλεκτοί της* [Greek and its dialects]. Thessaloniki: Institouto Neoellinikon Spoudon (Manolis Triantafyllides Foundation).
- Johanson, L. (2002). *Structural factors in Turkic language contacts*. London: Curzon.
- Karanastasis, A. (1984-1992). *Ιστορικόν Λεξικόν τῶν Ἑλληνικῶν ιδιωμάτων τῆς Κάτω Ἰταλίας* [Historical Dictionary of the Greek dialects of South Italy]. Athens: Academy of Athens.
- Karasimos, A. (2019). Οι νεοελληνικές διάλεκτοι στον ψηφιακό χάρτη: ανασκόπηση και προτάσεις. [Modern Greek dialects in digital maps: a review and suggestions]. In G. Karla, I. Manolessou, N. Pantelidis (eds.) *Λέξεις / Τιμητικός τόμος για την Χριστίνα Μπάσσα-Μπεζαντάκου* [Words | Studies in honor of Christina Bassea-Bezantakou]. Athens: Kardamitsa publications, pp. 109-142.
- Karatsareas, P. 2013. Understanding diachronic change in Cappadocian Greek: the dialectological perspective. *Journal of Historical Linguistics* 3(2), pp. 192–229.
- Katsoyannou M. & S. Armstrong (2019). Διαλεκτική λεξικογραφία στην Κύπρο: Προκλήσεις και προοπτικές [Dialectal Lexicography in Cyprus: Challenges and prospects], In G. Karla, I. Manolessou, N. Pantelidis, (eds.) *Λέξεις / Τιμητικός τόμος για την Χριστίνα Μπάσσα-Μπεζαντάκου* [Words | Studies in honor of Christina Bassea-Bezantakou] Athens: Kardamitsa publications, pp. 187-210.
- Katsouda, G. 2012. Διαλεκτική Λεξικογραφία: Επισκόπηση [Dialectal Lexicography: An overview]. *Lexicographikon Deltion* 26, pp. 77-159.
- Kostakis A. (1963). *Le parler grec d'Anakou*. Athenes: Kentro Mikrasiatikon Spoudon.



- Kostakis, A. (1986-1987). *Λεξικό της Τσακωνικής διαλέκτου*. [Dictionary of the Tsakonian dialect]. Athens: Academy of Athens.
- Kotsanidis, L. (2006). *Το γλωσσικό ιδίωμα του Μιστί Καππαδοκίας* [The Cappadocian dialect of Misti]. Kilikis: Gnomi.
- Kuzmina, V. & A. Rylova (2010). ABBYY Lingvo electronic dictionary platform and Lingvo Content dictionary writing system. In S. Granger & M. Paquet (Eds.) *E-Lexicography in the 21st Century: New Challenges, New Applications*. UCL Presses Universitaires de Louvain, pp. 419-423.
- Manolessou, I. (2016). Ο νέος τόμος του Ιστορικού λεξικού της νέας ελληνικής της Ακαδημίας Αθηνών: διαχρονικές προοπτικές [The new volume of the Historical Dictionary of Modern Greek of the Academy of Athens: diachronic perspectives]. *Studies in Greek Linguistics* 36, pp. 239-249.
- Manolessou, I., S. Beis & Ch. Bassea-Bezantakou (2012). Η φωνητική απόδοση των νεοελληνικών διαλέκτων [The phonetic transcription of Modern Greek Dialects]. *Lexikographikon Deltion* 26, pp. 161-222.
- Manolessou, I. & Ch. Bassea-Bezantakou (2017). "The Historical Dictionary of Modern Greek". In Wandl-Vogt E. & A. Dorn. (eds.) *Dialekt | Dialect 2.0. Langfassungen, 7. Kongress der Internationalen Gesellschaft für Dialektologie und Geolinguistik (SIGD)*. Wien: Praesens Verlag, pp. 13-37.
- Manolessou, I. & Katsouda G. (forthcoming). The making of the Historical Dictionary of Modern Greek: Problems and solutions in the domain of historical and dialectal lexicography. *Proceedings of the 10<sup>th</sup> International Conference on Historical Lexicography and Lexicology*, Fryske Akademy, Leeuwarden 12-14 June 2019.
- Marello, C. (2004). Lexicography in Italy: Specific themes and trends. *International Journal of Lexicography* 17.4, pp. 349-356.
- Mavrochalyvidis, G. & I. Kesiosoglou. (1960). *Τὸ γλωσσικὸ ἰδίωμα τῆς Ἀξοῦ (Le dialecte d'Axo)* [The Axós dialect]. Athens: Institut Français d'Athènes.
- Melissaropoulou, D. (2016). Variation in word formation in the light of the language contact factor: the case of Cappadocian Greek. *Journal of Language Sciences* 55, pp. 55-67.
- Melissaropoulou, D. (2019a). *Γλωσσική ποικιλία και γλωσσική αλλαγή στο σχηματισμό συγκριτικών παραθετικών δομών: δεδομένα από την Καππαδοκική* [Language variation and change in the formation of comparative constructions: evidence from Cappadocian Greek]. In A. Archakis, N. Koutsoukos, G. Xydopoulos & D. Papazachariou (eds.), *Γλωσσική Ποικιλία. Μελέτες αφιερωμένες στην Αγγελική Ράλλη*. Athens: KAPA EKDOTIKI, pp. 329-346.
- Melissaropoulou, D. (2019b). Morphological pattern replication phenomena as instances of typological shift. In G. Karla, I. Manolessou, N. Pantelidis (eds.), *Λέξεις | Τιμητικός τόμος για την Χριστίνα Μπασέα-Μπεζαντάκου* [Words | Studies in honor of Christina Bassea-Bezantakou]. Athens: Kardamitsa publications, pp. 317-340.
- Melissaropoulou, D., Galiotou, E., Dimela, E., Karanikolas, N., Papanagiotou, Ch., Xydopoulos, G. & A. Ralli (2015). Υλοποίηση του πρώτου πολυμεσικού τριδιαλεκτικού διαδικτυακού λεξικού [Implementation of the first multimedia tri-dialectal e-dictionary]. In M. Tzakosta (ed.) *Η διδασκαλία των νεοελληνικών γλωσσικών ποικιλιών και διαλέκτων στην πρωτοβάθμια και δευτεροβάθμια εκπαίδευση. Θεωρητικές προτάσεις και διδακτικές εφαρμογές* [Teaching Modern Greek Dialectal varieties in primary and secondary education]. Athens: Gutenberg, pp. 259-280.
- Papadopoulos, A. A. (1958-1961). *Ιστορικὸν Λεξικὸν τῆς Ποντικῆς Διαλέκτου* [Historical Lexicon of the Pontic Dialect]. Αρχεῖον Πόντου. Athens: Epitropi Pontiakon Meleton.
- Papazachariou D., Vassalou, N., & M. Janse (2016). Methodological Principles of Mišótika Cappadocian Data Collection. In I. Kappa & M. Tzakosta (Eds.), *7th International Conference on Modern Greek Dialects and Linguistic Theory*, Rethymno: University of Crete, pp. 32-32.
- Ralli, A. (2009). Morphology meets dialectology: insights from Modern Greek dialects *Morphology* 19, pp. 87-105.
- Ralli A. (ed.) (2015). *Πρόγραμμα Θαλής, Πόντος-Καππαδοκία-Αἰβαλί. Στα χνάρια της Μικρασιατικής Ελληνικής*. Πάτρα: Πανεπιστήμιο Πατρών.
- Reichmann, O. (2012). *Historische Lexikographie*. Berlin: De Gruyter.
- Themistocleous, C., Katsogiannou, M., Armosti, S., Christodoulou, K. (2012). *Cypriot Greek Lexicography: An Online Lexical Database*. In: *Proceedings of Euralex 2012*, pp. 889-891.
- Tiberius, C., Niestadt, J., Schoonheim, T. (2014). The INL Dictionary Writing System. *Slovenščina* 2.0, 2 (2), pp. 72-93.
- Thomason, S.G. & T. Kaufman. (1988). *Language contact, creolization and genetic linguistics*. Berkeley: University of California Press.
- Trudgill, P. (2003). Modern Greek dialects: A preliminary classification. *Journal of Greek Linguistics* 4(1), pp. 45-63.
- Vryonis, S. (1971). *The Decline of Medieval Hellenism in Asia Minor and the Process of Islamization from the Eleventh through the Fifteenth Century*. Berkeley: University of California Press.
- Winford, D. (2010). Contact and borrowing. In Hickey, R. (ed.), *The Handbook of Language Contact*. Malden, MA/Oxford: Wiley-Blackwell, pp. 170-187.
- Xydopoulos, G.I., Dimela, E., Melissaropoulou, D., Papanagiotou, Ch. & A. Ralli. (2015). ΛΕΠΟΚΑΜ, ένα πολυμεσικό λεξικό των διαλέκτων του Πόντου, της Καππαδοκίας και του Αἰβαλίου: Σχεδιασμός και υλοποίηση [LEPOKAM, a multimodal dictionary of the dialects of Pontus, Cappadocia and Aivali: Designing and Implementation]. In A. Ralli (ed.), *Πρόγραμμα Θαλής: Πόντος, Καππαδοκία, Αἰβαλί. Στα χνάρια της Μικρασιατικής Ελληνικής* [Thalis project: Pontus, Cappadocia, Aivali: in search of Asia Minor Greek. Patras: Laboratory of Modern Greek Dialects, University of Patras, pp. 99-114.



**Acknowledgements**

This research has been co-financed by the General Secretariat for Research and Technology and the Hellenic Foundation for Research and Innovation. Many thanks to Angela Ralli, Metin Bağrıaçık and Petros Karatsareas for providing access to oral dialectal material. Further information available at <http://cappadocian.upatras.gr/en>



# John Pickering's *Vocabulary* (1816) Reconsidered: America's Earliest Philological Exploration of Lexicography

Miyoshi K.

Soka Women's College, Japan

## Abstract

John Pickering is the author of the first dictionary of Americanisms, the *Vocabulary, or Collection of Words and Phrases Which have been Supposed to Be Peculiar to the United States of America* (1816). Allen Read, a masterly scholar of Americanisms, regards the dictionary as “an important landmark in the study of the English language in America”, acclaiming Pickering as “one of the most perceptive linguists America has produced”. However, there seems to be the situation that research on the *Vocabulary* has scarcely been done since the 1950's. Then, has research on the *Vocabulary* been exhausted? My answer to the question is “Never in the least”. When browsing through the *Vocabulary*, we can notice Pickering having finely used quite a few reference materials, thus the body of the *Vocabulary* becoming highly scholarly. As far as I can judge, this fact has not been pointed out to date. My intention in this paper is to clarify Pickering's use of English dictionaries out of such materials. To summarize my analysis, Pickering was versed in wide range of English dictionaries, making the fullest use of them for his investigation on the historical background of Americanisms.

**Keywords:** John Pickering; Americanisms; use of historical dictionaries

## 1 Introduction

The American lexicographer John Pickering is widely known among authorities on Americanisms for his dictionary published in the 1810's under the title *Vocabulary, or Collection of Words and Phrases Which have been Supposed to Be Peculiar to the United States of America* (1816). George Krapp's *English Language in America* (1925: vol. 1, 376) and Albert Baugh and Thomas Cable's *History of the English Language* (2002: 391 and 394), both of which are indispensable reference books for historical research on English in America, discuss Pickering's *Vocabulary*, their authors recognizing the dictionary as “the first dictionary of Americanisms”.

And, Allen Read (2002: 114), a masterly scholar of Americanisms, acclaiming Pickering as “one of the most perceptive linguists America has produced”, he (1947: 271) also regards the *Vocabulary* as “an important landmark in the study of the English language in America”. In addition, for Henry Mencken (1982: 48), the legendary authority who compiled the historic volume *American Language* (1919-1948), the *Vocabulary* was the “first really competent treatise on the subject [Americanisms]”. Such words of high commendation are quite notable when the fact is taken into account that the *Vocabulary* only treats approximately 600 words.

At the same time, however, in spite of such situations, there is the fact that research on the *Vocabulary* has scarcely been done since the 1950's; for instance, in the case of the journal *American Speech* (1925-), this seems not to have carried any papers which mainly treat the *Vocabulary* since 1957, with Henning Cohen's “Drayton's notes on Pickering's list of Americanisms” (1956) as the last one. In this situation, one very rare exception is Julie Andresen's *Linguistics in America, 1769-1924* which appeared in 1990. This book contains quite a few descriptions about the *Vocabulary*, but, regrettably, offers little discussion on the dictionary from the viewpoint of the history of American lexicography, although the book may be a well-documented study on the history of linguistics.

Two reasons are conceivable for such pretermission of the *Vocabulary*. One is the preconception that research on Pickering's *Vocabulary* has already been exhausted before the 1960's; this may indicate the conception among authorities that the *Vocabulary* is merely the beginning of the lexicography of American English and no more. And the other is the judgment that the *Vocabulary* is essentially not a scholarly dictionary but a conservative and normative one. As to the latter, it has usually been a customary practice for authorities concerned, including Mencken and Baugh and Cable, to cite such passages as the followings which is included in Pickering's “Essay”, an introductory material prefixed to the *Vocabulary*: it:

The preservation of the English language in its purity throughout the United States is an object deserving the attention of every American, who is a friend to the literature and science of his country. (Pickering 1816: 2)

[...] it [the language in the United States] has in so many instances departed from the English standard, that our scholars should lose no time in endeavouring to restore it to its purity, and to prevent future corruption. (Pickering 1816: 17)

In this regard, Richard Bailey's “National and regional dictionaries of English” (2009: vol I, 279-301) which comprises one chapter of *The Oxford History of English Lexicography* (2009) edited by A. P. Cowie, a standard book of reference for researchers on English lexicography, is symbolical. Bailey, in this chapter, divides the section related to the history of the dictionary of Americanisms into two parts; one is for the discussion of the generalities of American lexicography and the other for the treatment of “scholarly dictionaries of Americanisms”. He discusses Pickering's *Vocabulary* exclusively



in the former, using the latter for the treatment of William Craigie's *Dictionary of Americanisms* (1938-1942) and Mitford Mathews's *Dictionary of Americanisms on Historical Principles* (1951). Besides, Bailey quotes Pickering's statement in his letter from Read's work, which is "John Pickering [...] wrote to his father from London: 'I find we use several words in America [...] for which there is no authority' [quoted by Read 2002: 16]". There will be no denying that this passage written by Pickering himself strengthens the notion that he is a conservative purist in terms of the use of the language.

Then, has research on Pickering's *Vocabulary* actually thoroughly been done, leaving no room for further research? And, is the *Vocabulary* such strongly subjectively-based and lacking in objectivity? My answer to both of these questions is "Never in the least". This is because there is a decisive vacuum in research on the *Vocabulary* until today. That is, in the volume *Milestones in the History of English in America* (2002), which is the collection of Read's papers, Read refers to Pickering in eight pages, but we cannot see any mention in them concerning the point of what reference materials Pickering used to compile the *Vocabulary*. And in Mencken's fourth edition of the *American Language* (1982), which is an abridged edition annotated by Raven McDavid, where descriptions about the *Vocabulary* are seen in thirty-six pages in all, we see Mencken cite dozens of entries in the dictionary, but can hardly ever find his discussion on Pickering's reference materials for the *Vocabulary*; concerning this point, there is a possibility that McDavid may have deleted Mencken's reference to such materials, but if this is the case, it will corroborate the fact that McDavid did not attach importance to them, with having the notion that the dictionary is essentially subjectively-oriented.

Actually, however, when browsing through 113 entries in the *Vocabulary* whose head-words and head-phrases begin with the letters *J, K, L, M, N, O* and *P*, which comprise approximate 18 % of all entries in the dictionary, we can notice Pickering having used well more than 60 reference materials, in which dictionaries, state papers, periodicals, private letters and the records of lectures and sermons are included. And, concerning these reference materials, Pickering is found to have used them quite finely, thus the body of the *Vocabulary* being highly scholarly and even philological for such materials; I here use the term "philological" based on Tom McArthur's definition of "philology" in his *Oxford Companion of the English Language* (1992: 768): "the study of language, literature, and even national culture". As far as I can judge, this fact seems to have been thoroughly passed over in research on Pickering's *Vocabulary*.

Then, after the preamble so far, my intention in this paper is to clarify Pickering's use of dictionaries out of his reference materials within the range of the 113 entries I have mentioned. In order for this purpose to be fulfilled, I will divide my analysis into four sections. They are "dictionaries consulted by Pickering" (Section 2), "his ways of using dictionaries" (Section 3), "his reference to Webster's and Johnson's dictionaries" (Section 4) and "his comparative observation of dictionaries" (Section 5). Out of these four, the result of analysis in the section "dictionaries consulted by Pickering" is to be a basis of analysis in other three sections.

## 2 Dictionaries Consulted by Pickering

If we are to seek to know what dictionaries Pickering referred to in the compilation process of his *Vocabulary*, the fact comes to be perceived that he used 18 dictionaries by 14 lexicographers within the range of my scope, performing this practice 84 times in 47 entries, which account for 41.6 % of the 113 of my scope. This situation is as shown in the "Table 1" below; here, the editions of the dictionaries, when they are specified, are based on Pickering's indications, and I also specify the frequency of his reference to each of the dictionaries, as well, which is the result of my analysis.

Ash, John, <i>The New and Complete Dictionary</i> (1775 edition): 4 times in 4 entries.	Pegge, Samuel, <i>A Supplement to the Provincial Glossary of Francis Grose</i> : 4 times in 4 entries.
Bailey, Nathan, his dictionary (1721 edition) (title not specified): once.	Perry, William, <i>The Royal Standard Dictionary</i> (1755 edition): once.
-----, his dictionary (1727 edition) (title not specified): once.	-----, <i>The Royal Standard Dictionary</i> (1805 edition): once.
-----, his dictionary (1730 edition) (title not specified): once.	Ray, John, <i>South and East Country Words</i> : 3 times in 3 entries.
Barclay, James, <i>A Complete and Universal English Dictionary</i> (1774 edition): once.	Rees, Abraham, <i>The Cyclopaedia, or Universal Dictionary of Arts and Literature</i> : once.
Entick, John, <i>The New Spelling Dictionary</i> (1764 edition) : once.	Sheridan, Thomas, <i>A General Dictionary of the English Dictionary</i> (1780 edition): 3 times in 2 entries.
-----, <i>The New Spelling Dictionary</i> (1795 edition): once.	Walker, John, <i>A Critical Pronouncing Dictionary and Expositor of the English Language</i> (1791 edition): 2 times in 2 entries.
Grose, Francis, <i>A Provincial Glossary</i> : 7 times in 7 entries.	Webster, Noah, <i>A Compendious Dictionary of the English Language</i> : 22 times in 20 entries.
Johnson, Samuel, <i>A Dictionary of the English Language</i> : 18 times in 18 entries.	
Mason, George, <i>A Supplement to Johnson's English Dictionary</i> : 4 times in 4 entries.	

Table 1: List of Dictionaries Consulted by Pickering.

From an overall perspective, for the present, we may say that the table also reflects the extensive knowledge of Pickering's, a figure in the 1810's America, concerning dictionaries in two respects. One is that he referred to dictionaries of pronunciation, of spelling, of dialects and of arts and literature, along with the general type of dictionary, which may suggest his command of using various types of dictionaries in a highly appropriate way. The other is that he referred to the



different editions of one dictionary, as seen in the cases of the dictionaries of John Entick's and William Perry's, which signifies the fact that his interest extended to the development of dictionaries. In addition to these two points, it is also to be noted that Pickering referred to Noah Webster's and Samuel Johnson's dictionaries quite frequently as compared to other dictionaries. These points are included in those about which I intend to expound in the following sections, with examples cited from the *Vocabulary* and the provision of another table.

### 3 His Ways of Using Dictionaries

I hope I have successfully revealed the generalities of Pickering's use of dictionaries with the use of the "Table 1" in the previous section. Then, in what ways did he use the 18 dictionaries which I pointed out? As to this point, he used the dictionaries in three ways, which I will discuss below one by one.

Firstly, he referred to dictionaries 45 times out of the 84, which I mentioned in the previous section, as well, to determine whether or not a specific word is treated in them and, when such a word is treated, to show whether or not its specific meaning is treated. The following is one example of this case:

(1) From the entry on **meadow**: "[...] it [*meadow*]" is defined by *Bailey* – 'Pasture land yielding grass, hay,' and *Sheridan* (who is followed by *Walker*) also defines it – 'a rich posture ground, from which hay is made.'"

This example may also be regarded as reflecting Pickering's close perusal of each dictionary which he referred to. That is, he, in this example, compares the dictionaries of Nathan Bailey's, Thomas Sheridan's and John Walker's; as to such a situation, I will expound in Section 5 later.

Secondly, he cites the views of lexicographers 31 times regarding the use of words, like the following:

(2) From the entry on **to narrate**: "Walker [...] thus defend the word [*narrate*]: As it is derived from the Latin *narro*, and has a specific meaning to distinguish it from every other word, it ought to be considered as a necessary part of the language."

In this example, it may be worthy of note, as well, that the *Vocabulary*, a dictionary of Americanisms, even makes reference to Latin through Walker's dictionary.

And thirdly, with regard to the remaining 8 cases (84 - 45 - 31), Pickering used dictionaries in the way in which the first and second ways are combined, as the following instance shows:

(3) From the entry on **ledge**: "*Grose* defines it [*ledge*], 'brisk, lively' and says it is used in the *South*."

Here, Pickering's use of Francis Grose's dictionary (*A Provincial Dictionary*) also gives a glimpse of his interest in dialects; it may also be noted that he is objective here, making no critical remarks about the use of the dialect (*ledge*), whose similar situation is observable here and there in the *Vocabulary*.

Such are the three basic ways of Pickering's use of dictionaries, and it may be said that we can confirm the fact that the *Vocabulary* is basically a dictionary of language, rather than an encyclopedic which lays emphasis on explanation about things.

### 4 His Reference to Webster's and Johnson's Dictionaries

Next, the "Table 1" in Section 2 shows that Pickering most often referred to Noah Webster's *Compendious Dictionary of the English Language* (1806) and Samuel Johnson's *Dictionary of the English Language* (edition not specified); as to the former, we should note that Pickering's *Vocabulary* was published twelve years before Webster's renowned *American Dictionary of the English Language* (1828). To be specific, Pickering referred to Webster's *Compendious* 22 times in all in 20 entries and Johnson's *Dictionary* 18 times in all in 18 entries, both within the range of my scope. It can be known how often Pickering referred to the two dictionaries when we recognize the fact that his reference to the third most frequently cited dictionary, namely Francis Grose's *Provincial Glossary* (edition not specified), is limited to 7 times in all in 7 entries.

This can be regarded as the evidence of the fact that Pickering attached special importance to the treatment of Webster's *Compendious* and Johnson's *Dictionary*. In this situation, if we are to clarify Pickering's reference to dictionaries in his *Vocabulary*, a close analysis of his use of the two dictionaries will be unavoidable. I will, in the following, examine how he treated Webster's *Compendious* and Johnson's *Dictionary* in this order.

#### 4.1 His Reference to Webster's *Compendious Dictionary*

As to his reference to Webster's *Compendious*, in order to know how intently Pickering read the dictionary, the beginning of the entry on **prayerless** in the *Vocabulary* is suggestive, which is the following:

(4) "Not praying, not using prayers." *Webst. Dict.* I have never known this word to be used here [America] [...].

Without closely perusing the *Compendious*, he could not have said like this. If I cite one more similar example, Pickering says the following in the entry on **lengthy**:

(5) Mr. *Webster* has admitted it [*lengthy*] into his dictionary; but (as need hardly be remarked) it is not in any of the *English* ones.

The situation is the same not only about the aspect of words as seen in the examples above but also about that of meanings. That is, we can see such examples as the following, which is from the entry on **location**:



(6) “The act of designating or surveying and bounding land; the tract so designated.” *Webst.* This substantive [noun] is in the English dictionaries, but not in this sense.

In this way, Pickering actually cites words and meanings peculiarly treated in Webster’s *Compendious* in 8 entries out of the relevant 20. They are entries on **kentle**, **lengthy**, **location**, **noticeable**, **to packet**, **to parade**, **prairie** and **prayerless**. In most cases, Pickering is critical about Webster, as inferred from the three examples above.

## 4.2 His Reference to Johnson’s *Dictionary*

Similarly to the case of Webster’s *Compendious*, Pickering is thought to have been considerably familiar with the contents of Johnson’s *Dictionary* in respect of its information on the language; this is significant in America at the beginning of the 1800’s, as I will mention at the end of this sub-section. The following passage from the entry on **to progress** in the *Vocabulary* is one example which shows how closely Pickering perused Johnson’s *Dictionary*:

(7) It is true that some authorities may be found for it [*progress*] in English writers, and it is accordingly in *Johnson’s* and other dictionaries; but *Johnson* has noted it as “*not used*.” It seems also, that the accent was formerly placed on the *first* syllable, and not (as we pronounce it) on the last [...].

This entry also suggests the possibility that Pickering had even noticed the fact that Johnson placed an accent mark in the relevant entry-word; although it is not certain which edition of Johnson’s *Dictionary* Pickering referred to, we can see that the entry-word is written as “*To Pro’gress*” in its first edition (1755).

In regard to pronunciation, Pickering, in the entry on **perk**, provides the following information on its relevant word, as well, using Johnson’s *Dictionary*:

(8) It [*perk*] is used in the interior of *New England*; and is commonly pronounced *pear*k, (the *ea* as in *pear*) just as it is written in the passage which Dr. Johnson quotes from *Spenser*.

We can also see Pickering’s description as the following in the entry on **obnoxious**, this time concerning the meaning of the word:

(9) The English formerly used *obnoxious* in the sense of *liable* or *subject to*; and *Johnson* accordingly explains each of these words by the others.

From an overall viewpoint, Pickering, within the range of my scope, referred to Johnson’s *Dictionary* in these entries of his *Vocabulary*: **to jeopardize**, **jeopardy**, **jockeying**, **leanto or lean-to**, **to legislate**, **meadow**, **mean for means**, **mission**, **muggy**, **to narrate**, **navigation**, **near for to or at**, **to notice**, **to notify**, **obnoxious**, **to peak or peek**, **plenty for plentiful** and **to progress**.

It will be notable that all references in such entries concern the substantial contents of Johnson’s *Dictionary*; there had probably been no one or very few persons in America who had perused Johnson’s *Dictionary* so closely from the perspective of the use of the language before Pickering. Concerning the point, it is widely known among authorities on historical English lexicography that in the 1950’s, well more than one hundred years after Pickering’s *Vocabulary*, James Sledd and Gwin Kolb published the epoch-making work *Dr. Johnson’s Dictionary: Essays in the Biography of a Book* (1955). However, even this work, as the word “*biography*” in its subtitle indicates, refers mainly to the historical background of the *Dictionary* and little to its contents; for this reason, the book was to be sternly criticized by William Wimsatt (1956: 308) that it is “shaped somewhat like a doughnut, the hole being the *Dictionary* itself”. Therefore, once again, Pickering’s perusal of Johnson’s *Dictionary* from a linguistic viewpoint may be regarded as notable when we think of the fact that he was an American at the very beginning of the 1800’s.

(If I may add a few words to the above, it will be a basic knowledge and common agreement among authorities on the history of American lexicography that Webster’s *Compendious* only gives very brief definitions for entry words, never providing any lexicographical information, which can be seen if we riffle through the small dictionary. In this sense, it is far from probable, as well as probably absurd to think, that the *Compendious* might have exerted essential influence on the *Vocabulary*. Rather, it may be thinkable that Webster was influenced from the *Vocabulary* in the compilation process of the *American Dictionary* to be published twelve years later.)

## 5 His Comparative Observation of Dictionaries

Thus far, I have mainly analyzed Pickering’s use of each individual dictionary, laying emphasis on his treatment of Noah Webster’s *Compendious* and Samuel Johnson’s *Dictionary*. However, Pickering’s use of dictionaries is not limited to such a sphere. Although I touched upon this point briefly in Section 3, significant is the fact that he quite frequently compared the dictionaries, especially focusing on the contents of Webster’s and Johnson’s. For example, the entry on **jeopardize** reads the following:

(10) It [*jeopardize*] is doubtless a corruption of the ancient verb *jeopard* [...]. But even the verb *to jeopard*, which is in all the dictionaries, Dr. *Johnson* says, is “*obsolete*”; *Ash* says, it is “*not much used*”; and *Barclay*, that it is “*used only in Divinity*”. It is hardly necessary to remark, that *to jeopardize* is not in any of the dictionaries.

In this example, Pickering compares the contents of the dictionaries of Johnson’s, John Ash’s and James Barclay. And we can see the following passage in the entry on **to legislate**:

(11) *Walker* has inserted it [*legislate*] in his dictionary, but (as he remarks) it is “*neither in Johnson nor Sheridan*,” nor is it in *Mason’s Supplement* to Johnson. It was noticed, however, several years ago in *Entick’s* dictionary, (edition 1795);



and, more lately, in an edition of *Sheridan*, “corrected” and improved by Salmon;” and also in the octavo edition of *Perry’s* dictionary, published in 1805. Mr. Webster adopts it from *Entick*.

Here, in this short entry, six dictionaries are compared: John Walker’s, George Mason’s, John Entick’s, Thomas Sheridan’s, William Perry’s and Webster’s. Besides, this passage, which shows an aspect of the development of English dictionaries, may even be said that it reflects how intently Pickering did research on lexicography with a keen interest. This fact is not only seen in two examples above. He performed similar practice in 18 entries out of the 47 where dictionaries are referred to in my scope. The following is the table which indicates whose dictionaries Pickering compared in each of the 18 entries.

<b>jog</b> : Bailey and Grose	<b>muggy</b> : Pegge and Johnson
<b>to jeopardize</b> : Ash, Barclay and Johnson	<b>to narrate</b> : Johnson and Walker
<b>kedg</b> : Grose and Ray	<b>to notice</b> : Ash and Mason
<b>knoll</b> : Grose and Ray	<b>obnoxious</b> : Ash and Johnson
<b>leanto or lean-to</b> : Mason, Pegge, Johnson and Webster	<b>packet</b> : Webster and other English lexicographers
<b>to legislate</b> : Entick, Mason, Perry, Sheridan, Walker and Webster	<b>to peak or to peek</b> : Johnson and Webster
<b>liability</b> : Entick and Mason	<b>perk</b> : Johnson and Webster
<b>meadow</b> : Bailey, Johnson and Sheridan	<b>plenty for plentiful</b> : Ash and Johnson
	<b>poorly</b> : Ash, Pegge and Webster
	<b>punk</b> : Ash, Bailey and Webster

Table 2: Pickering’s Comparison of Dictionaries.

In the entries cited in this “Table 2”, we can clearly see a method of comparative lexicography. When Reinhard Hartmann and Gregory James’s *Dictionary of Lexicography* (1998) is referred to, this fact seems to back up the probability that Pickering’s *Vocabulary* is highly “philological”, whose notion I mentioned in the introductory section. However, the entry on **philology** in Hartmann and James’s dictionary, although it apparently focuses on words or grammatical constructions from different languages in different periods, reads thus: “A branch of linguistics concerned with the comparative-historical perspective in language study. The principles of philology have led to the development of historical lexicography and comparative lexicography”. If we base ourselves on Hartmann and James’s explanation here, there will be little problem if we position Pickering as one forerunner of researchers in comparative lexicography in America, his *Vocabulary* being probably the first philological work there, at the same time.

## 6 Conclusion

I, in Section 4, quoted a phrase from William Wimsatt’s review of James Sledd and Gwin Kolb’s book on Samuel Johnson’s *Dictionary*: “shaped somewhat like a doughnut, the hole being the *Dictionary* itself”. This phrase may also be applicable to research on Pickering’s *Vocabulary* until today. If this is the case, the substantial contents of the *Vocabulary* have been made light of for two centuries since its publication. Besides, we should be reminded of the fact that it is often the case with a historical dictionary that the contents of its prefixed material are greatly incompatible with the contents of its bodies, the former not giving a glimpse of the latter. In the case of the *Vocabulary*, we can hardly surmise its substantial contents, which is scholarly, scientific and descriptive, from Pickering’s “Essay” prefixed to it, whose contents is quite the opposite and strongly conservative and didactic.

Besides, it may be noted here, incidentally, that, as far as the contents of my scope is concerned, we can scarcely find didactic or prescriptive remarks in the *Vocabulary* as to the use of the language, despite the fact that it introduces, for instance, quite a few regional dialects with reference to dictionaries.

To summarize my analysis in this paper, Pickering was versed in the characteristics of wide range of English dictionaries, making the fullest possible use of them for his scientific research on the historical background of Americanisms. This will be regarded as highly notable in the 1810’s America.

It was half a century after Pickering’s *Vocabulary* that the first complete history of English lexicography, Henry Wheatley’s “Chronological notices of the dictionaries of the English language” (1865) appeared, but this work seems to still only tenuously go into the contents of each dictionary treated. As for Pickering, he did research on various dictionaries, closely perusing their contents, not only as a theoretician but also as a practitioner of compiling a dictionary. If we are to seek a predecessor of Pickering, it will be difficult to find any English lexicographer with the only exception of Samuel Johnson, who thoughtfully used various types of dictionaries, including the dictionaries of etymology, encyclopaedic dictionaries, as well as the general type of dictionary. In this sense, apart from the types of dictionaries they compiled, Pickering may be regarded as a successor to Johnson in respect of the way of using dictionaries.

## 7 References

- Andresen, Julie T. (1990). *Linguistics in America, 1769-1924: A Critical History*. London and New York: Routledge.
- Bailey, Richard W. (2009). National and regional dictionaries of English. In Cowie, A. P. (ed.) *The Oxford History of English Lexicography*, vol. 1. Oxford: Oxford University Press, pp. 279-301.
- Baugh, Albert C. and Thomas Cable (2002). *A History of the English Language* (5<sup>th</sup> edition). London and New York: Routledge.
- Cohen, Henning (1956). Drayton’s notes on Pickering’s list of Americanisms. In *American Speech* Vol. 31, No. 4 (Dec. 1956), pp. 264-270.
- Hartmann, Reinhard R. K., James, Gregory (1998). *Dictionary of Lexicography*. London and New York: Routledge.



- Johnson, Samuel (1755). *A Dictionary of the English Language* (1<sup>st</sup> edition) (2 vols.). Facsimile reprint, Tokyo: Yushodo, 1983.
- Krapp, George P. (1925). *The English Language in America* (2 vols.). New York: Appleton Century Crofts.
- McArthur, Tom (ed.) (1992). *The Oxford Companion to the English Language*. Oxford and New York: Oxford University Press.
- Mencken, Henry L. (1982). *The American Language: An Inquiry into the Development of English in the United States* (the fourth edition and the two supplements, abridged, with annotations and new material, by McDavid, Raven I. Jr.) (one-volume abridged edition). New York: Alfred A. Knopf. Inc.
- Pickering, John (1816). *A Vocabulary or Collection of Words and Phrases Which Have Been Supposed to be Peculiar to the United States of America*. Facsimile reprint, London: Routledge/Thoemmes Press, 1997.
- Read, Allen W. (1947). The collection for Pickering's 'Vocabulary'. In *American Speech* Vol. 22, No. 4 (Dec., 1947), pp. 271-286.
- (2002). *Milestones in the History of English in America* (edited by Bailey, Richard W.) (Publication of the American Dialect Society, No. 86) (Annual Supplement to *American Speech*). North Carolina: Duke University Press.
- Sledd, James H. and Gwin J. Kolb (1956). *Dr. Johnson's Dictionary: Essays in the Biography of a Book*. Chicago: University of Chicago Press.
- Wells, Ronald A. (1973). *Dictionaries and the Authoritarian Tradition: A Study in English Usage and Lexicography*. The Hague and Paris: Mouton & Co. N.V.
- Wheatley, Henry B. (1865). Chronological notices of the dictionaries of the English language. In *Transactions of the Philological Society*, 1865, pp. 218-293.
- Wimsatt, William K., Jr. (1956). Sledd, James H. and Gwin J. Kolb. *Dr. Johnson's Dictionary: Essays in the Biography of a Book*. University of Chicago Press, 1955. Pp. viii+255 (book review). In *Philological Quarterly*, Vol. 35, No.3 (July, 1956), pp. 308-310.



# Studying language change through indexed and interlinked dictionaries

Ore C.-E.<sup>1</sup>, Grønvik O.<sup>2</sup>

<sup>1</sup> University of Oslo, Norway

<sup>2</sup> University of Bergen, Norway

## Abstract

In this paper we present our study how to use the Meta Dictionary of the Norwegian Language Collections to measure lexical stability in standard dictionaries across a timespan. The Meta Dictionary uses the lexical item as its core unit, expressing each lexical unit in a separate Meta Dictionary entry. The success of this model rests on having access to electronic versions of major and generally accepted dictionaries from the different stages of the orthography of a language. With this documentation it is possible to see, for instance, how much and which parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk and Bokmål respectively, and whether this lexicon is present in its original orthography or not. This method for studies of the lexical development is comparable to remote sensing in archaeology and distant reading in literary studies. As an extended example of the application of the method we study a few issues related to the position of the pioneering lexicographers Ivar Aasen (1813-1896) and Hans Ross (1831-1912) in the description of Nynorsk, as shown in more recent lexicographical works, and in particular in two school dictionaries from 1954 and 1970 which border on being spellers.

**Keywords:** lexical item, lexicon, language change, dictionary, Meta Dictionary model, standard language, orthography, Norwegian

## 1 Lexical stability in standard dictionaries

How can one measure and document lexical and orthographic development in a written standard language? An obvious method is to start by comparing the selection of lexical items in a number of successive dictionaries. Although dictionaries do not document actual language usage, they represent what was thought essential vocabulary at the time of publication, in the form valid at the time. A systematic comparison of the lexical item inventory from a large number of dictionaries requires too much labour for manual execution, but becomes possible with the availability of systematized digital resources, that is, (retro) digitized dictionaries and a proper instrument for analysing them.

The instrument used in this study is the *Meta Dictionary* of the Norwegian Language Collections (Ore & Grønvik 2018). The *Meta Dictionary* is an electronic register of lexical items of Norwegian (Bokmål and Nynorsk), linking base forms with a wide range of usage examples and entries from 90 dictionaries. The *Meta Dictionary* also coordinates documentation of the vernacular with excerpts from literature in both the Norwegian Standard languages (Bokmål and Nynorsk). It has become a research tool, for instance for looking at change in standard orthographies through more than 150 years of documentation.

The *Meta Dictionary* format has also been adapted for the new editing and publication system of the *Dictionary of Old Norse Prose* (Ordbog over det norrøne Prosasprog, ONP), see for example Johannsson & Battista (2014). For dictionary linking see also Møller, Troelsgård and Sørensen (2019). The German *Wörterbuch Netz* is an example of a coordinated dictionary collection (Wörterbuchnetz 2020).

Our aim is to use existing dictionaries and other materials in electronic form to document the history of language standardisation in Norway, in order to create a digital historical record of standard Norwegian orthography (Bokmål and Nynorsk) and its documentation in dictionaries.

## 2 Measuring lexical stability

Lexical stability in dictionaries and spellers can be measured by looking at the numbers of lexical items occurring in a series of (intentionally or unintentionally) normative documents, such as dictionaries and spellers. Lexical stability in dictionaries and spellers do not give certain information about usage, but they show what authors and publishers have regarded as necessary and desirable.

Measuring lexical stability in running texts is no more complex if the text is lemmatized; that is, for every word form (token) a base form (type/lexical item) has to be identified. Many European languages, among which Norwegian, have undergone changes in orthography and inflectional morphology over the last 150 years. For these languages, a lemmatizer optimized for the current version of the written standard, does not work well for older texts. There are several ways to construct lemmatizers, statistical, deep learning and, more traditionally, from full form registers. In the latter method a diachronic lemmatizer will need a set of full form registers each representing the orthography after a major language reform.

At present the *Meta Dictionary* registers for each lexical item a list of dated base forms which correspond to the orthography of a given period. Lexical stability in preserved and searchable text can to some extent be explored and measured through being searched with this register of dated base forms. The base forms reflect the orthography of their



time. This method would therefore probably be useful in dating the searched text.

A filtering of a text on the basis of a base form register, without inflected forms for each base form, would yield less reliable results in identifying and establishing a register of lexical items in the searched text. Results would depend on the character of the language in question. A text from a heavily inflected language with a great deal of heterogeneity would present greater difficulties than a very regular language with few inflected forms and few deviations from base forms.

### 3 The instrument

The Meta Dictionary was originally designed and established in 1999 as a common index to the digitized source material for the Norwegian Dictionary project (NO2014). Each entry in the Meta Dictionary consists of a head and a body. The head is a list of base forms with information about POS, language, orthographic status and the time span for this status. The body of an entry consists of (hyper)links to separate source databases with examples of usage. See the Euralex 2018 paper by Ore & Grønvik (2018) and section 5 in this paper for a more detailed description of the sources.

The organizing principle is that each entry in the Meta Dictionary should correspond to one lexical item. Atkins and Rundell (2008: 163 f.) use the term ‘lexical item’ to denote everything that may deserve lexicographical treatment in a dictionary entry, including compounds, multiword expressions, symbols and abbreviations. We extend the concept to cover both diachronically and synchronically orthographic variants of base forms belonging to the same lexical item. Therefore, the Meta Dictionary method of comparing dictionaries at the level of lexical item is independent of orthographical variants of the headwords.

A lexical item that has been identified and described in a dictionary entry, and is generally accepted by language users, is likely to keep its place as a separate lexical item. In general, this holds true for Nynorsk. Two non-linguistic factors have nevertheless affected the issue of lexical identity for a few lexical items. They are (1) the policy of promoting orthographic approach between Nynorsk and Bokmål on the basis of vernacular word forms with a wide geographical and usage distribution (merging the verbs Nynorsk *ganga* + Bokmål *gå* to Nynorsk and Bokmål *gå* ‘walk’ is a case in point), (2) the more recent desire to use the dictionary entry head to inform language users of semantic alternatives (The Nynorsk nouns *lege* and *lækjar* ‘medical doctor’ have different forms and etymologies, and are different lexical items in linguistic terms, but can be found as alternative headword forms in dictionaries and spellers).

In the Meta Dictionary, pairs of type 2 are dealt with as separate lexical items, Transitions of type 1 are accepted. Derived forms, singly and in compounds, are treated as separate lexical items (a case in point are verbal adjectives, f.i. *-gjengen* (< Nynorsk *ganga* ‘walk’) and *-gått* (< Nynorsk and Bokmål *gå* ‘walk’).

The implication of using the lexical item through time, represented by the set of headword forms found for each particular lexical item in the standard orthography, is to aim for full homograph separation over time in the Meta Dictionary. The term ‘homograph’ is here used as in the editorial guidelines for *the Norwegian Dictionary (Norsk Ordbok)*, and is understood to mean ‘headword with identical orthographic form to another headword’ (Grønvik & Gundersen 2016: 78 ff.). The criteria for homograph separation in base form is that each homograph represents a different lexical item, identified by pronunciation, inflection, etymology and usage.

From a strictly synchronic view homography is (almost) identical to polysemy. From a diachronic point of view, the origin (etymology) is the discriminating feature for otherwise identical word forms. Distinguishing lexical items through a sequence of orthographic forms is not an exact science, and a complicating factor is that what at one point in time seems to be evidence that two word forms have different origins, does not preclude earlier common origins. The Norwegian adjective ‘feig’ has the two distinct meanings; (1) ‘cowardly’ from German ‘feige’ and (2) ‘close to death’ from Old Norse ‘feigr’, cf. English ‘fey’ with apparently the same Germanic root as the other two. In theory, the two different meanings of modern Norwegian ‘feig’ could be considered to be either homographs or two senses under the same lexical item. Most modern dictionaries of Norwegian chose the latter solution.

In our study the timespan is much more modest - only 150 years. Even so, full homograph separation can be difficult, as changing headword forms of different lexical items can change the pattern of homographs. Full homograph separation linked to materials therefore requires the sorting of materials according to context and meaning.

The Meta Dictionary was designed and funded to give a systematized access to the source materials for the NO2014 project. A large part of the source materials is found in the retro digitalized slip archive (3.2 million slips) comprising excerpts from a wide variety of sources. This digital slip archive was the basis for the construction of the Meta Dictionary. At this early stage the grouping of word forms followed the principles of lemma selection of the Norsk Ordbok, Heavily etymological, Norsk Ordbok grouped verbal derivatives (verbal nouns and adjectives) under the verb, without taking into account that a derived form might have developed into an independent lexical item. To some extent this grouping was used even for derived compound adjectives based on phrasal verbs. However, the empirical material we use is carefully linked to the Meta Dictionary, observing the principle of the lexical item with the exception that the material is not completely homograph separated. This latter fact should always be taken into consideration, but will have little or no influence on the results in the current study.

The success of the Meta Dictionary model rests on having access to electronic versions of major and generally accepted dictionaries from the different stages of the orthography of a language. The expression “generally accepted” implies ‘accepted for use in education and official documents; used as a standard in examination systems’.

Of the two standard languages of Norwegian, Bokmål and Nynorsk, Nynorsk is at present the better equipped with major electronically available dictionaries (Ore 2020). The documentation of Nynorsk standard orthography rests on the following resources, all of which are linked to the Meta Dictionary (see Table 1).

With this documentation of Nynorsk 1873 – 2012 interlinked, it is possible to see - for instance - (1) how much and which



parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk and Bokmål respectively, and (2) whether this lexicon is present in its original orthography, or not. It is also possible to see (3) what influence the vernacular can be said to have had on modern standard Norwegian (Bokmål and Nynorsk). This paper offers answers to these questions in relation to Nynorsk only. The data for Bokmål are not yet complete enough for a reliable analysis.

On the lexicographical level, this universe of dictionaries is also an excellent tool for studying the development in the focus of dictionary makers, that is, which lexical items are considered worthy of being described over time. The usefulness of the method is demonstrated and described in Ore (2020).

#### 4 Comparing lemma selection in dictionaries across a timespan

One can imagine the lexicon of a language as a vertical column consisting of fibres, where each fibre represents a lexical item. The column can be marked with years on the vertical axis. A horizontal cut at a given point will show the lexicon of the year in question. The ideal way of showing the lexicon of a given year is through a comprehensive dictionary using the orthography valid for that year, linked to earlier and later documentation of the same lexical item.

The model discussed here is based on the category system used in the Meta Dictionary. The Meta Dictionary uses the lexical item as its core unit, expressing each lexical item in a separate Meta Dictionary entry. A lexical item can have several headword forms in base form. The category schema of each headword form allow annotation by (1) language (Bokmål or Nynorsk), (2) part of speech (POS), (3) status within the orthography, and (4) start and end-date of a given status (Ore & Grønvik 2018).

The Meta Dictionary is not completely homograph separated - some entries comprise more than one lexical item. When comparing two dictionaries through shared lexical items in the Meta Dictionary we may get a slightly higher match than what is the true case. In the modern dictionary of Nynorsk, *Nynorskordboka* (2012), a little less than 9% of the headwords are marked as homographs. Therefore, there may be an uncertainty of around 5% in our results. This has no consequence for the tendencies shown. The most important aspect of this project at the present stage has been to create a practicable model with clear categories, which is true to the underlying linguistic and lexicographic description, is transparent and has general value (Ore 2016).

The model does not include the category "place". For the standard language Nynorsk this is unimportant. Nynorsk is used in Norway only and has never been a first written language outside Norway. For Bokmål the case is different. The development of Bokmål starts as a history of deviation from standard Danish around 1900. To get a starting platform for the Danish of Norway, a general dictionary of 19<sup>th</sup> century Danish should be included in the Meta Dictionary materials. Only with this starting point will it be possible to document systematically the divergence of Danish in Norway from the Danish of Denmark and into Dano-Norwegian, resulting in today's Bokmål. The standard language Nynorsk is based on the field work of the linguist and lexicographer Ivar Aasen in the mid 19<sup>th</sup> century, and became a written standard in parallel with Danish by an act of the Norwegian parliament in 1885.

#### 5 Dictionary sources

In the present study we focus on Nynorsk, for which we have better digital resources. The major orthographic reforms of Nynorsk occurred in 1873, 1901, 1917, 1938, 1959, 1986 and 2012. The orthographic reforms have influenced lemma selection in dictionaries in the sense that editors will have felt obliged to give information on changes. The selected dictionaries were published close in time to the reforms in order to capture such changes.

Aasen and Ross (1873 - 1895) are scholarly pioneer dictionaries designed to portray the Norwegian vernacular, primarily from the country dialects used by the majority of the population 1840-1890. Vocabulary perceived by their informants as foreign, was not included in either dictionary. Aasen's dictionary concentrated on the central, well documented vocabulary with derivations and some compounds. Ross includes more compounds and variants.

Skard 1903 (1901 orthography) is a bilingual school dictionary with Nynorsk headwords and Danish equivalents, with a focus on orthography and virtually no additional information. The Skard school dictionary (1. ed. 1901, expanded and corrected 1903) was the first series of school dictionaries covering Nynorsk.

The Norsk Ordbok Draft Manuscript (1917 orthography) was composed as a preliminary to Norsk Ordbok. Its lexical programme was to prepare for a scholarly dictionary covering the whole vocabulary of the Norwegian vernacular and written Nynorsk of all genres. The Draft manuscript is based on Aasen and Ross, some normative dictionaries (Schjøtt, Eskeland) from the beginning of the century, some large collections from specific dialects and a limited range of older, written sources of the vernacular from the Dano-Norwegian period 1550-1850. The Norsk Ordbok Draft Manuscript includes a fair amount of entries for imported vocabulary from Latin, Greek, French etc, but the specifically Norwegian vocabulary has priority.

Norsk Ordbok (1938 orthography) is at 330 000 lexical items almost three times the size of the Draft Manuscript. The special responsibility to present the original Norwegian materials is there, but in relation to lemma selection from written Nynorsk current general selection criteria were applied – i.e. no ejection of imported lexical items demonstrably in use in Nynorsk text.

Skard 1954 (1938 orthography) is the fifth edition of school dictionary originally published in 1922. It differs from Skard 1903 in size and scope, addressing the needs of the education system in general and of public administration.

Hellevik 1970 (1959 orthography) is the second series of school dictionaries to cover Nynorsk, and the first wholly monolingual series. Hellevik claims to give a "more complete and reliable image of actual usage in Nynorsk than any earlier (orthographic) dictionary", with explicit reference to the language collections ordered through the Meta Dictionary as a primary source.



Nynorskordboka (2005 orthography) was originally edited on the basis of the language collections now encompassed by the Meta Dictionary, in parallel with its sister volume, Bokmålsordboka. The two editorial teams drafted each their half of the alphabet and then swapped drafts, adapting each whole to the given standard language (Bokmål or Nynorsk), in order to avoid omissions and unnecessary differences. Nynorskordboka also had the rule that any word from the vernacular documented from three or more counties was to be included (Nynorskordboka 1986: VII).

Year of Orthography	Year of publication	Author and title	Type	Number of lexical items
1873	1873	Ivar Aasen: <i>Norsk Ordbog</i> .	General dictionary	38 711
1873	1895	Hans Ross: <i>Norsk Ordbog</i> .	General dictionary, presented as an addition to Aasen (1873)	49 220
1901	1903	Matias Skard: <i>Landsmaals-Ordlista</i>	School speller	12 000
1917	1991-1997	Draft manuscript of <i>Norsk Ordbok</i>	General dictionary	113 000
1938	1950 – 2016	<i>Norsk Ordbok</i> 1-12	Scholarly multivolume dictionary	330 000
1938	1954	Matias Skard; <i>Nynorsk ordbok for rettskriving og literaturlæsnad</i>	Speller for use in education and administration.	32 000
1959	1970	Alf Hellevik: <i>Nynorsk ordliste. Større utgåve</i>	Speller for use in education and administration.	29 000
1986	1986	<i>Nynorskordboka</i> 1. ed.	General dictionary	90 000
2012	2012	<i>Nynorskordboka</i> web edition	General dictionary	90 000

Table 1. Sources of Nynorsk Orthography 1873 – 2012.

## 6 Results and discussion

There are endless interesting byways to explore once dictionaries are properly interlinked and indexed at the lexical item level. In the rest of the paper we will give an extended example of how the Meta Dictionary method of comparing dictionaries can be used in practice.

We will study a few issues related to the position of the pioneering lexicographers Ivar Aasen (1813-1896) and Hans Ross (1831-1912) in the description of Nynorsk, as shown in more recent lexicographical works, and in particular in two school dictionaries which border on being spellers. The purpose is more to demonstrate what is possible to find out by using methods of sorting and grouping than to settle issues relating to the Nynorsk lexicon once and for all. Evidence from dictionaries and language collections cannot outweigh evidence from very large corpora. However, to move to that step, the base form history of each lexical item must be in place and a full form registry available for the whole.

This exploration of Nynorsk lexicography does not address issues arising from the rapprochement of Nynorsk and Bokmål in the course of the 20<sup>th</sup> century. Such comparisons must wait until the documentation of Danish as used in Norway is in place, together with materials showing the earliest deviations from the Danish of Denmark.

### 6.1 The dictionaries of Aasen and Ross and the later dictionaries in numbers and proportions

The dictionaries of Aasen and Ross, published in the second half of the 19<sup>th</sup> century, were the first description of the vernacular language in Norway. Table 2 gives an overview, showing to what extent the later dictionaries included lexical items from Aasen and Ross. Column 2 shows the scope of each dictionary by the number of lexical items given entries. Column 3 gives the percentage of the lexical items from Aasen and Ross with entries in the later dictionaries.

The size of the dictionaries differs. The school spellers (lines 2, 4 and 5) are smaller than the dictionaries of Aasen and Ross, while the scholarly dictionaries (lines 3, 6 and 7) are much larger. The numbers in the right column indicate the percentage of the lexical items in the dictionaries of Aasen and Ross which are brought forward. Skard (1903) contains 19 % of the lexical items found in dictionaries of Aasen and Ross. Hypothetically, it could contain at most 21 %. This means that Skard (1903) almost entirely consists of lexical items from Aasen and Ross. This is not surprising due to closeness in time.

Skard (1954) is almost 2.5 times the size of Skard (1903), and has a higher number of lexical items from the dictionaries of Aasen and Ross (17 410 lexical items). But the relative amount of lexical items from Aasen and Ross is smaller. Although more than 50% of the lexical items stem from Aasen and Ross, there is a clear tendency that the lexical items for the dictionaries of Aasen and Ross are less dominant, as so much other material is included. In Hellevik (1970) this tendency is much stronger. Here just 43% of the total (11 585 lexical items) stems from Aasen and Ross.

Two of the scholarly dictionaries, The Draft Manuscript of *Norsk Ordbok* and *Norsk Ordbok* itself, are much larger than the dictionaries of Aasen and Ross combined. They are also programmatically committed to include Aasen and Ross complete. Both cover 95% of the Aasen and Ross lexical items, i.e. all except some cross references.

The third dictionary, *Nynorskordboka* (1986), is a collegiate dictionary. The size and the focus on dialects allow for including a relatively large number of lexical items from Aasen and Ross.



	1	2	3	4	5
	Dictionary	Number of lexical items with entries	% of the total of lexical items in the dictionary stemming from Aasen and Ross	% lexical items in Aasen and Ross continued	% if all possible lexical items were continued
1	Aasen and Ross	64 334	-	-	-
2	Skard (1903)	13 390	92	19	21
3	N.O. Draft Manuscript	95 908	64	95	100
4	Skard (1954)	31 864	55	28	51
5	Hellevik (1970)	26 201	43	18	42
6	Nynorskordboka (1986)	87 145	26	34	100
7	Norsk Ordbok (2016)	300 117	20	94	100

Table 2. Number (percentage) of lexical items in the first description of Nynorsk reoccurring in later dictionaries. The right column indicate the percentage if all possible lexical items (limited by space) reoccurred.

## 6.2 How much and which parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk Dictionaries?

The first question in this chapter heading can easily be answered by inspecting the Meta Dictionary. There are 10 198 lexical items where a form of the headword has an entry both in the dictionaries of Aasen and Ross combined and in both of the two school dictionaries Skard (1954) and Hellevik (1970).

The second question in the chapter heading is more complex and can be reformulated as follows: To what extent do the lexical items in the dictionaries of Aasen and Ross that also occur in the post-war school dictionaries represent the central vocabulary of Nynorsk? The expression “central vocabulary” is found especially in second language teaching materials, but rarely defined. In this paper, “central vocabulary” is taken to mean the group of lexical items that are (deemed) essential to achieving mastery of a given language by virtue of their authenticity, meaning, relevance, frequency, stability in use across a timespan and their lexicogenetic potential. These indicators of centrality are based on a discussion in Fjeld and Vikør (2011:156 ff.).

The fact that an entry in an old dictionary reoccurs in a more recent dictionary does not in itself reflect the status of the word in the language at the time of selection for each dictionary. The Meta Dictionary entry does, however, link dictionary entries with usage examples from literature and from dialect collections, most of which were collected and published in the twentieth century, and with a majority of instances from the later period (post-1950). It is therefore possible to use the Meta Dictionary to look at the coverage of each lexical item in sources other than the dictionaries mentioned above.

When estimating the combined contribution of the dictionaries of Aasen and Ross to the lexical items of newer school dictionaries by means of the information found in the Meta Dictionary, it is relevant to look at the 10 198 joint lexical items in relation to following five categories:

- **Authenticity** here simply means independent proof that the word exists in the language outside the dictionaries. The Meta Dictionary shows that this requirement is met by all 10 198 lexical items – all have usage materials in addition to the dictionary entries, though not necessarily very many.
- **Frequency** cannot be measured directly through the Meta Dictionary, as it could in a lemmatized corpus. The number of registered instances (quotations and other lexical items of information) per entry gives an indication of centrality, since the range of sources is considerable (ca 5 000 literary sources, roughly 90 dictionaries, plus dialect archives and reference works). So looking at the number of instances behind each entry, compared to the Meta Dictionary as a whole, should give an indication of centrality. See table 3 below.
- **Stability across a timespan.** A high number of registered instances for an entry in the Meta Dictionary indicates stability across a timespan, since a high number from a short period of time is unlikely. But the best indicator of stability through time is the fact that almost all the 10 198 joint lexical items also occur in Nynorskordboka, edited 1974–1986, cf. table 2.
- **Word structure** is important in Norwegian, a Germanic language in which any simple (i.e. morphologically indivisible) lexical item can be used as a building block in complex lexical items (derivations and compounds). In general, simple lexical items tend to have more meanings and a higher lexicogenetic potential than lexical items with a complex structure. This tendency is strengthened if a simple lexical item also has high frequency in use. The Meta Dictionary headword forms are structure marked. It is therefore possible to see whether the joint lexical items have simple or complex headword base forms.



- **Spoken and written sources.** A lexical item documented both from spoken and from written sources will be more likely to hold a place in the central vocabulary of a language than a lexical item with only the one or the other, since lexical items with limited sources are more likely to be strongly marked in terms of style or subject field. In the case of the 10 198, they are all documented from the vernacular of the 19<sup>th</sup> century, through the dictionaries of Aasen and Ross. The majority of the additional sources registered in the Meta Dictionary come from literature, but it is possible that some of the 10 198 lexical items have speech sources only.

An analysis of lexical items as represented in the Meta Dictionary category system can give an indication on authenticity, frequency and word structure, and by implication say something about stability across a timespan. Estimating meaning and relevance of the lexical items common to Aasen and Ross and the school dictionaries, compared to the Meta Dictionary as a whole, would require a study of the materials behind each entry, which is beyond the scope of this paper.

Number of instances	Aasen & Ross + Skard (1954) and Hellevik (1970)	Percent (of 10 198)	Aasen & Ross + Skard (1954) and Hellevik (1970) restricted to preserved headword. forms	Percent (of 7 483)	Total for all entries in the Meta Dictionary - Nynorsk
1 (hapax legomenon)					264 788
1 – 9	37	0.4	22	0.3	263 065
10 – 99	5 147	50.5	3 590	48.0	91 768
100 – 999	4 856	47.6	3 742	50.0	9 427
>1000	158	1.5	129	1.7	241
In all	10 198	100	7 483	100	529 289

Table 3. Aasen and Ross lexical items grouped by number instances from other sources linked to the Meta Dictionary.

Almost half the lexical items of the entire Meta Dictionary (Nynorsk) are represented by only one instance of lexical information, while the group of lexical items with 1000 or more instances represent 0.04 per cent of the total. In other words: the distribution of instances (similar to tokens in a corpus) corresponds to Zipf's law. The distribution of the instances found in Aasen and Ross combined and in Hellevik (1970) and Skard (1954) have a different distribution. Table 3 groups the lexical items from Aasen and Ross found in Skard 1954 and Hellevik 1970 by the number of instances found in the Meta Dictionary.

The structure of the Meta Dictionary makes it possible to get a precise count of instances for groups of lexical items. The 10 198 lexical items from Aasen and Ross present in Skard 1954 and Hellevik 1970 (table 3, column 2) have in all 1 768 636 instances in the Meta Dictionary, an average per entry of 173. The median is 98. The group of lexical items from the dictionaries of Aasen and Ross present in their original orthographic form in Skard and Hellevik have 1 359 683 instances, an average of 182 instances behind each lexical item, and the median is 103. The Meta Dictionary as a whole has roughly 3.5 million instances from Nynorsk sources. If the hapax lexical items are disregarded, the average number of instances per Nynorsk entry is 9. The lexical items originating from the dictionaries of Aasen and Ross belong to the group of Meta Dictionary lexical items with the highest number of instances and are consequently among the best documented ones. This fact suggests that the 10 198 lexical items from Aasen and Ross and found in Skard (1954) and Hellevik (1970) belong to the central vocabulary of Nynorsk – and this is even more so the case for the lexical items from the dictionaries in Aasen and Ross with preserved orthographic form.

### 6.3 The orthographic form of the Aasen and Ross lexical items - in the original orthography, or in a newer orthographic form?

The Meta dictionary entry is indexed by headword base forms belonging to a standard orthography of Bokmål or Nynorsk, with a starting date and a final date. The latter is set to 31.12.9999 for base forms within the current orthography. This artificial final date is not shown in the web interface. Since there have been several revisions of the orthography, the entry head can look like this:

køyrar m (Nynorsk, 1873-) kjørar m (Nynorsk, 1959-2012)
--

Figure 1. The form *køyrar* ('driver' noun, masculine) is spelt as it was in Aasen 1873; the form *kjørar* was a permitted form from 1959 until 2012, but is no longer part of the Nynorsk orthography.



POS	(1) Aasen and Ross Lexical items in Skard (1954) and Hellevik (1970) (100 %)	(2) Subset of lexical items with present day orthography (73.5 %)	(3) Subset of simple (non-compounds) lexical items with present day orthography (60.8 %)
In all	10 259	7 516	6207
Adjective	1750	1213	858
Adverb	228	123	97
Conjunction /Subjunction	11	11	10
First part of compound	1	1	1
Interjection	6	2	2
Noun, no certain gender	8	3	3
Noun fem.	1576	797	632
Noun masc.	2667	2022	1683
Noun neut.	1561	1084	780
Numeral	26	13	12
Preposition	71	33	15
Pronoun	39	32	29
Verb	2631	2286	2182

Table 4. Lexical items common to Aasen and Ross, Skard (1954) and Hellevik (1970) distributed according to POS. (1) in all, (2) with preserved orthography from Aasen and Ross in the base form, (3) with preserved orthography and simple word structure. The total sum is smaller than the sum of lexical items with POS due to the fact that some nouns having more than one gender.

Several of the orthographic revisions were motivated by an expressed need for modernity, the implication being that Aasen's orthography is - and was - out of date. It is therefore interesting to see how many of the lexical items from the dictionaries of Aasen and Ross, included in Skard (1954) and Hellevik (1970), where the original orthographic base form is preserved in today's Nynorsk orthography. The numbers are as shown in table 4.

The Meta Dictionary has POS as a category. The POS distribution is included in table 4. What is striking in this table is the size of the verb group, compared to nouns and adjectives. Verbs have both derived adjectival forms (participles) and nominal forms (suffix-derived verbal nouns), which makes them lexicogenetically powerful.

The lexical items found in Aasen and Ross AND in Skard 1954 and Hellevik 1970 are so well represented in the Meta Dictionary that it seems reasonable to consider them part of the central vocabulary of Nynorsk. This assumption is supported by the fact that 73.5 % of the total have preserved their orthographic form. 60 % are have a simple word structure – they are not compounds, and if they are derivations they are most likely derived by suffix (f.i. *-leg* adj. ‘-ly’). Some orthographic changes have been minimal. In 1917, nouns with the feminine gender and ending in the vowel *-a* had the ending changed to *-e*. This change probably partly explains the fact that only half of the feminine nouns – the ones ending in a consonant – have preserved their original orthographic form. However, such detailed checking is outside the scope of this paper.

#### 6.4 What influence can the vernacular be said to have had on modern standard Nynorsk?

This discussion is limited to the presence of Aasen and Ross in the two school dictionaries Skard (1954) and Hellevik (1970), with a side glance at Nynorskordboka.

The dictionaries of Aasen and Ross represent 70 years of fieldwork in the Norwegian vernacular. This paper shows that a substantial part of smaller modern dictionaries consists of lexical items also found in the dictionaries of Aasen and Ross. A majority has preserved the original orthography, and a majority has a simple word structure, offering opportunities for easy re-use in new and complex word forms. It seems reasonable to conclude that the spoken language, both in standard form and in dialect form, has been included in the linguistic toolbox of modern Nynorsk.

#### 6.5 Modernising the selection of lemmas – how does Hellevik (1970) differ from Skard (1954)?

This essay in investigating dictionaries through sorting and grouping will round off by taking a closer look at the two school dictionaries, to see at what points they resemble each other and at what points they differ.

In the preface to his dictionary Hellevik (1970) claims to have a more modern lexicon to offer than other contemporary



school dictionaries. Below, in tables 5 and 6, the selection of lexical items unique to either dictionary is looked at and compared at some points. A full comparison of the two lists of lexical items would involve comparing categories which the Meta Dictionary does not cover, i.e. pronunciation, inflection and sense description. The comparison below is limited to POS, with observations on some derivational suffixes.

Table 3 shows that Skard (1954) has 55 % of its lexical items in common with the combined dictionaries of Aasen and Ross, while the Hellevik (1970) has 43 % in common with the combined dictionaries of Aasen and Ross. Relatively speaking and in absolute numbers Skard (1954) is closer to Aasen and Ross in its lemma selection than Hellevik (1970) is.

Dictionary	Lexical items	Lexical items from Aasen and Ross	Shared lexical items Aasen and Ross	Shared lexical items	Number of unique lexical items
Skard (1954)	31 864	17 421	10 198	18509	13 355
Hellevik (1970)	26 201	11 498	10 198	18509	7 692

Table 5. A comparison of contents between Skard (1954) and Hellevik (1970).

Table 5 shows that Skard (1954) and Hellevik (1970) have 18 509 lexical items in common, 58 % of Skard (1954), close to 67 % of Hellevik (1970). More than half of the shared lexical items, 10 196 are also found in the dictionaries of Aasen and Ross. The chief differences between them must therefore be found in the parts of the dictionaries that are unique to either. Hellevik has the smaller number of unique entries.

POS	Hellevik (1970)	Skard (1954)
Unique entries in all	7 692	13 355
Adjective	1 344	2 771
Adverb	280	137
Abbreviation	9	10
Compound first part	41	6
Interjection	13	7
Conjunction /subjunction	14	5
Derivational prefix	9	13
Preposition	61	32
Pronoun	9	7
Noun (no certain gender)	23	6
Noun feminine gender	790	3 392
Noun masculine gender	2 693	3 134
Noun neuter gender	2 005	2 208
Name (place, person)	265	611
Numeral	9	2
Symbol	4	
Verb	1 136	1 165

Table 6 The POS distribution in Hellevik (1970) and in Skard (1954) for the lexical items unique to either dictionary.

Table 6 shows the POS distribution in Hellevik (1970) and in Skard (1954). As some nouns have more than one gender, the sum of POS occurrences is slightly larger than the total of unique lexical items. What is notable in Skard (1954) is the high number of adjectives and of nouns with feminine gender. Unique to Skard (1954) is a large number of nouns which are verbal derivatives with the suffix *-ing*, feminine gender in Nynorsk. What is notable in Hellevik (1970) is the comparatively high number for the function POS – especially adverbs and prepositions – and the special entries for the



first part of compounds, since nesting is not used. A closer look shows that Hellevik (1970) has included a number of compound and phrasal adverbs and prepositions, much used in everyday language but largely unnoticed by grammarians. The high number of entries for first parts of compounds found in Hellevik (1970) may have a basis in a post war focus on systems of word formation – lexicogenesis – in the standardisation of written Norwegian (both Nynorsk and Bokmål).

## 7 Conclusion

Comparing the inventory of lexical items through a series of very different dictionaries becomes a history of lemma selection for 150 years. The dictionaries have different sources, and differ in size and planned usage. However, once a lexical item has been identified and described in a dictionary, it will not disappear as a lemma candidate. The inclusion or exclusion of a documented lexical item therefore expresses a choice on the part of the editor. The results of the comparison discussed in this paper will therefore show what direction lemma selection for Nynorsk dictionaries has taken over the years.

Plans for the future include developing a full form generator for the older orthographies – which will involve expansion of the existing schemas to include plural forms of verbs and the dative case of nouns, and, of course, adding materials on Bokmål in order to facilitate the use of the Meta Dictionary as a tool in tracing the development of the majority written standard of Norway.

## 8 References

- Aasen, I. (1873). *Norsk Ordbog med dansk forklaring*. Christiania: Mallings Boghandel, 1873.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bergenholtz, H., Cantell, I., Fjeld, R.V., Gundersen, D., Jónsson, J.H. & Svensén, B. (1997). *Nordisk leksikografisk ordbok*. Skrifter utgitt av Nordisk forening for leksikografi. Skrift nr. 4. Oslo: Universitetsforlaget.
- Bokmålsordboka (2019). Accessed at: <http://ordbok.uib.no> [30/05/2020]
- Draft manuscript of Norsk Ordbok (Grunnmanuskriptet–1940) (1997). Accessed at: <http://usd.uib.no/perl/search/search.cgi?appid=59&tabid=993> [30/05/2020]
- Eskeland, S. (1919). Framandordbok. med tyding og rettleiding um lesemaaten til 8-9 tusen av dei vanlegaste framandordi. Kristiania. Norli.
- Fjeld, R.V., & Vikør, L.S. (2011). *Ord og ordbøker*. Kristiansand: Høyskoleforlaget.
- Grønvik, O. (1980). Framandorda og norsk språkutvikling i nyare tid. I: *Sprog i Norden*, 1980, s. 39-60. Accessed at: <http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive> [30/05/2020]
- Grønvik, O. & Gundersen, H. (2016). *Redigeringshandbok for Norsk Ordbok 2014*. Accessed at: <http://no2014.uib.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf> [30/05/2020]
- Hellevik, A. (1970). Nynorsk ordliste. Større utgåve. Med fornorskings-tillegg og liste over forkortinger. Oslo: Det Norske Samlaget.
- Johannsson, E.T., & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users — Paper vs. Web-based Edition. In Abel, A., Vettori, C., Ralli, N. (eds) *Proceeding of the XVI EURALEX, The User in Focus; 15-19 July 2014, Bolzano*, Bolzano, Eurac Research, 2014, ISBN: 978-88-88906-97-3. Accessed at: <http://euralex.org/category/publications/euralex-2014/> [30/05/2020].
- Møller Svendsen, M.-M., Troelsgård, T. & Sørensen, N. H. (2019). *Salmesang og superordbok: ordbogslinkning i praksis* (Hymn Singing and a Super Dictionary: Dictionary Linking in Practise), presentation at NFL 2019, to be published in *Leksikografi 15*, abstract available at: <https://www.helsinki.fi/sv/konferenser/15-konferensen-om-lexikografi-i-norden/program-och-abstrakt> [30/05/2020]
- Norsk Ordbok (2016). Norsk ordbok – Ordbok over det norske folkemålet og det nynorske skriftmålet, 1-12. Oslo, Samlaget 1950–2016
- Nynorskordboka (1986). *Nynorskordboka – Definisjons og rettskrivingsordbok*. Oslo: Det Norske Samlaget.
- Nynorskordboka (2019). Accessed at: <http://ordbok.uib.no> [30/05/2020]
- Ore, C.-E. S. (2016). Gamle ordbøker og digitale utgaver. I *Nordiske Studier i Leksikografi 13 Rapport fra 13. Konferanse om Leksikografi i Norden København 19.-22. mai 2015*. Nordisk forening for leksikografi 2016 ISBN 978-87-992447-6-8. pp. 203-216
- Ore, C.E. S. (2020). Å ta Hans Ross på ordet. Ross' ordbok i relasjon til Aasens med Metaordboka som verktøy. In *Nordiske Studier i Leksikografi 15* (under publication)
- Ore, C.-E. S., Grønvik, O. (2018). Comparing Orthographies in Space and Time through Lexicographic Resources. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Znanstvena založba Filozofske fakultete 2018 ISBN 978-961-06-0097-8. pp. 159-172
- Ross, H. (1895). *Norsk ordbog: Tillæg til "Norsk ordbog" af Ivar Aasen*. In Ross, Hans: *Norsk ordbog*. Universitetsforlaget. Oslo. 1971.
- Schjøtt, S. (1909). *Dansk-norsk ordbog*. Kristiania. Aschehoug.
- Skard, M. (1903). *Landsmaals-ordlista godkjend til skulebruk. Norsk rettskrivningslære II*. Andre utgåve, gjennomsedd og auka. Kristiania. Aschehoug
- Skard, M. (1954). *Nynorsk ordbok for rettskriving og litteraturlæse*. 5. utgåve ved Vemund Skard. Oslo. Aschehoug.
- Wörterbuchnetz (2020). Accessed at: <http://www.woerterbuchnetz.de> [30/05/2020].









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicological Issues of Lexicographical Relevance**







# When neologisms don't reach the dictionary: occasionalisms in Spanish

Bueno Ruiz P.J.

Universitat Pompeu Fabra

## Abstract

In this research we have analysed a type of neologism that does not become part of the common speech and so has no chance of being incorporated into a general language dictionary: the *occasionalisms*. They have been defined by several authors as volatile language events difficult to detect and typified by their context-dependency and language creativity. With this research we intend, on the one hand, to contribute with the description of the theoretical concept of *neologism* by defining one of the groups that are obtained following the lexicographical criterion and, on the other hand, to play a part in the definition of occasionalisms by finding regularities and their tendencies in formation processes and in formation rules. The data we have used belongs to the *Neómetro* database and applying the methodology described below we have obtained an occasionalism corpus. After analysing the data, we can confirm that occasionalisms are not linguistic random acts and they are lexical units that fulfil an expressive function in a specific context.

**Keywords:** occasionalism; nonce-word; neology; lexicographical criterion

## 1 Introduction and Objectives

Our aim in this article is to analyse, within the framework of lexical innovation, a type of neologism that does not become established in speakers' use of and so has no chance of being incorporated into a general language dictionary: the occasionalism. In one of the first works in which the concept of *neologism* appeared, Matoré (1952) highlighted the distinction that had to be made between prototypical neologisms and other units that appear in social use. Algeo (1993) also focused on these less institutionalized words and established up to 43 categories within the units of a corpus that did not make it into normative dictionaries.

In this perspective of lexical innovation, the situation of occasionalisms is that of expressive rather than denominative units created by speakers, a highly active line of lexical research for French and English, but which has not received much attention for Spanish. Nevertheless, Spanish has a lot of material with which to study occasionalisms, since these neologisms appear in neology databases because they belong to the group of words that can be obtained following the lexicographical criteria, as we will see later with examples such as *neohisteria* ("Y entramos en un proceso de *neohisteria*. Y buscamos respuestas en el yoga, las terapias y las religiones alternativas")<sup>1</sup> or *cernudear* ("El deseo no se describe, observa Clavecía; los poemas, sí. Y sigue *cernudeando*")<sup>2</sup>.

In this article we will offer, first of all, a brief state of the art that will allow us to know at what stage the research on occasionalisms is and to establish the theoretical elements from which to develop our work. Next, we will present the data that make up our occasionalism corpus, since our main objective is precisely to contribute to the study of occasionalisms in Spanish. Finally, we will focus on the analysis of the data in order to isolate the main tendencies within these lexical creations that are usually considered transgressive inventions and for which no regularities have been described.

## 2 Neologisms and Occasionalisms

The lexicographical criterion proposed by Rey (1976) and systematically applied in the work of detecting neologisms has allowed research to advance considerably in the description of neology in the different languages of the world, given that the lexicographical criterion makes it possible to detect neologisms in an objective and largely automated process. Indirectly, this criterion has also been useful in neology from a theoretical point of view because it has made specialists aware of the need to distinguish the theoretical concept of neologism from the applied one, which constitutes one of the great unresolved issues in the studies of this field from its beginnings. Matoré (1952: 87) defined *neologism* as "*acception nouvelle introduite dans le vocabulaire d'une langue à une époque déterminée*", but he pointed to a distinction between what he considered *necessary creations* and *luxury creations*. In relation to the latter, Matoré stated that "*elle dépend en réalité du milieu et du moment ; pour des mondains qui désirent se distinguer, telle création de mot désignant un aspect nouveau et souvent fugace de la mode*" and added "*on ne sait jamais quand apparaît une création si elle est passagère ou non*" (1952 : 88). Other authors among those considered *fathers of neology* in the French context followed this line: Rey (1976) pointed out the existence of words that could be considered neologisms but that might be due to an artistic or ludic factor, and Guilbert (1975) also established this difference and called these units *néologismes de langue*.

In this section we will present the current state of knowledge, half a century after these first contributions to neology as a discipline. We will focus on the two aspects that are most relevant to our research: a review of the results offered by the lexicographical criterion and a summary of the elements used at present to distinguish occasionalisms from neologisms.

<sup>1</sup> *Fashion&Arts*, 19/02/2017.

<sup>2</sup> *La Vanguardia*, 3/03/1989.



## 2.1 The Lexicographic Criterion as a Starting Point

The lexicographic criterion applied to the practice of detecting neologisms indicates that the term *neologism* can be used to refer to any lexical unit which is detected in use (usually taken from press texts, but increasingly enriched with texts from other sources such as social and other media) and is not included in the *exclusion corpus*, i.e. the corpus taken as a reference and normally made up of general lexicographical works. But considering that dictionaries exclude from their list of lemmas many lexical unities for very different reasons, by the sole application of this criterion, a heterogeneous group of units which we call *lexicographical neologisms* is obtained, among which neologisms strictly speaking can be found.<sup>3</sup> In the case of Spanish, all the nodes of the two Spanish neology networks<sup>4</sup> apply the lexicographical criterion and have detected the following examples as the twenty lexicographical neologisms most frequently found in press in the last ten years: *smartphone*, *app*, *eurozona*, *blog*, *chavismo*, *hashtag*, *soberanista*, *bullying*, *copago*, *ciberataque*, *selfie*, *tablet*, *tuit*, *playoff*, *sobrecoste*, *mediocentro*, *reality*, *uribismo*, *megaproyecto* and *recapitalización*. Several of these twenty units are no longer neologisms because they have recently been included in the academic dictionary of Spanish, *Diccionario de la lengua Española* (from now on DLE): *blog*, *soberanista*, *tuit*, *reality*, and also *selfi* and *tableta* for *selfie* and *tablet*. Some of the remaining words are given another name in Spanish and therefore do not appear in the dictionary (*acoso escolar* for *bullying* or *aplicación* for *app*), others are *witness words* (such as *chavista*, *chavismo*, *urbismo* from Hugo Chávez and Álvaro Uribe, who were presidents of Venezuela and Colombia respectively), and many others are words that have been stabilized in use for years, so that the reason they are not in the dictionary is not their recentness but rather their semantic transparency and the predictability of their forms, as we can see in *recapitalización*, *megaproyecto*, *mediocentro*, *sobrecoste* or *copago*. But if we look closely at all the lexical units detected by applying the lexicographical criterion, and not only at the most frequent ones, we realize that other types of non-recent units appear: dialectalisms, specialized lexical units, colloquialisms, variants and occasionalisms, among others.

For this reason, since Cabré et al. (2004) the usefulness of continuing to apply this criterion in the detection of neologisms has been widely discussed and, aside from the absence of a better-known objective criterion, it has been determined that it is useful and that it presents a high degree of interest both for lexicography and for the study of lexical innovation; for lexicography, because it makes it possible to discover gaps in dictionaries and, for the study of lexical innovation, because it allows us to isolate different types of words, whether newly created or not, as well as to study lexical evolution mechanisms. Naturally, it also contributes to the exhaustive detection of neologisms themselves. We will now focus on occasionalisms, one of the types of unit that are detected, one which is difficult to dictionaries.<sup>5</sup>

## 2.2 Occasionalisms as a Non-evolutionary Unit

As we saw at the beginning of this section, from the start, the definition of neology has proved ineffective in distinguishing authentic, prototypical neologisms from other types of words that are documented in language use. A neologism is not just a “new word”, as any speaker might say, nor is it a “new word, meaning or turn of phrase in a language” as the normative dictionary of Spanish says. In addition to their recentness, neologisms must meet at least another requirement, which is, according to Hohenhaus (2007) and Schmid (2008), that they are to become established in speakers’ use of the language. With this second requirement, occasionalisms will not be considered neologisms, because they have a non-evolutionary character, as we will see below.

Lipka et al. (2004) and Hohenhaus (2007) place particular emphasis on the distinction between what the English literature has called *nonce-word* (also *nonce-formation* or just *nonce*), the first occurrence of a word in a language, and what is properly considered a neologism. Although in some studies the first appearance of a word is given the name of *hapax legomenon*, Hohenhaus’ distinction between these two words is relevant: *hapax* refers to the first time a word is registered in a given corpus and *nonce word* refers to the first time the speaker registers a word in his/her lexicon, that is, the first time a user “creates” the word.

The distinction between occasionalisms and neologisms is even more evident in the current literature in the French context, where the concept of *occasionalism* is described with greater clarity. Thus, Dal and Namer (2018) and Sablayrolles (2018), among others, define it as a unit that meets a communication need in a given and specific context. They consider that these units cannot be analysed outside the context in which they were created, since their sole purpose is to respond exclusively to an expressive need (whether ironic, playful, etc.) for a given phrase. In fact, Sablayrolles (2018) also questions whether these occasional creations should be considered neologisms simply because they are documented once in use.

The concept of the *nonce word* as the first stage of a word is defended by Štekauer (2002) and Hohenhaus (2007), so that, from its creation, this word could initiate a process in which the second stage is called *institutionalization* (the use of the word by the speaking community), and the last stage, *lexicalization*, at the end of which it could be included in dictionaries (Hohenhaus 2007). Schmid (2008) makes an adjustment to this which is especially interesting to us: for him the progression of a nonce-word can be developed at three levels, so that three parallel processes occur: *institutionalization* (socio-pragmatic level), *lexicalization* (structural level) and *hypostatization* (cognitive level).

What concerns us for our research on occasionalisms is *the first stage* seen from Schmid’s three perspectives. An occasionalism is a non-evolutionary unit and, from the moment of its creation, it shows no signs of following the process

<sup>3</sup> According to this criterion, different kind of units are collected, including occasionalisms, ephemeral neologisms (Cabré 1989), words with occasional use, neologisms themselves and terminological units among others.

<sup>4</sup> The network Antenas Neológicas (<https://www.upf.edu/web/antenas>) brings together since 2003 peninsular and American Spanish neology observatories, with nodes in Spain and in different Latin American countries; NEOROC (<https://www.upf.edu/web/neoroc>) brings together since 2004 neology observatories located in different places in Spain.

<sup>5</sup> To *dictionary*: to include a neologism in a dictionary.



that a neologism would follow, but it *can* be analysed from the three levels: its appearance in language use as a *one-time unit* with pragmatic value; its non-lexicalizable character because of its transgressive formation (it doesn't follow the traditional word-formation rules); and the absence of cognitive hypostatization, since its appearance does not initiate the construction of a concept: the unit is understood with the help of the context it depends on.

### 2.3 The Dictionarization of Neologisms and Occasionalisms

Lexicographic neologisms are then those neologisms that are documented in use and cannot be found in general normative dictionaries<sup>6</sup> whose mission is, in most Romance languages, to collect the most institutionalized words (term used under Schmid's perspective) in the common mental lexicon of speakers. But this does not mean that other neologisms are not lexicographically interesting, since many of them can appear in other types of dictionaries. In fact, between the general dictionary and a glossary of transgressive words there is a continuum of lexicographical works in which more neologisms can be found.

First of all, Spanish has a type of general dictionaries called *usage dictionaries* (*diccionarios de uso*) because they attempt to reflect real usage and not correct usage (although it cannot in fact be said that general dictionaries do *not* reflect the lexicon considered correct), including the *Diccionario de uso del español* by María Moliner (1991; 2016), the *Diccionario del español actual* by Manuel Seco et al. (2006) and the *Diccionario de uso del español actual* (*Diccionario Clave*) by SM editorial (2011). These dictionaries traditionally incorporate a greater number of neologisms, so that they sometimes represent their prelude to official dictionarization in the standard dictionary.

Secondly, Spanish also has dictionaries of neologisms whose list of lemmas is based on lexicographical criteria, so that they function as a complement to the standard dictionaries. In this case we must highlight the *Diccionario de voces de uso actual* by Manuel Alvar (1994) and the *Nuevo diccionario de voces de uso actual* by the same author (2001), together with the most recent work also published on paper *Neologismos del español actual* (Moliner 2013). Since dictionaries of neologisms have to make a sustained effort not to become outdated, many are published directly online so that they can be continually updated, such as the *Diccionario de neologismos online* (IULA 2007)<sup>7</sup> and the *Diccionario de neologismos* (NEOMA 2016),<sup>8</sup> or blogs specializing in neologism such as *Martes Neológico*,<sup>9</sup> produced by *Centro Virtual Cervantes*, and *Antenarío*,<sup>10</sup> which belongs to *Antenas Neológicas*<sup>11</sup> network.

All of these lexicographic resources focus, in fact, on a certain type of neologisms, those that have certain stability in use, thus playing the function of usage dictionaries before the dictionarization of a word.

Consequently, low-frequency neologisms are not included in the dictionaries we have seen so far, but they also have their own lexicographic place with transgressive word dictionaries – which are particularly frequent in French<sup>12</sup> with works such as the *Dictionnaire des mots sauvages* by Rheims (1989), the *Dictionnaire des mots rares et précieux* (2006), the *Dictionnaire des mots rares et savoureux* (Horvilleur 2013) or *Mes mots sauvages* (Brisac 2018).

The aim of these works is not to reflect the new lexicon of the language but to show the most audacious creations, and they contain very interesting material for the study of lexical innovation because the resources used in the creation of highly expressive units can be observed in them. We are not aware of the existence of current works of this kind of dictionaries for Spanish, but we do know of works that somehow include infrequent units<sup>13</sup> and, in any case, occasionalisms would find their lexicographical space in this last type of dictionary.

## 3 Data and Methodology

The data we have used for this research belong to the *Neómetro* database,<sup>14</sup> whose basis is data from BOBNEO,<sup>15</sup> to which a substantial amount of linguistic and documentary information has been added. At present this database has 6,134 lexicographical neologisms<sup>16</sup> in different stages of analysis, and for our research we have selected the 3,402 units with the maximum level of completeness, so that for each neologism we have the following information:

- Lemma, grammatical category, and variants (if any).
- Usage context with the source and the date.
- Linguistic analysis: type of neologism and analysis of its constituent elements.

<sup>6</sup> Some dictionaries have a section of institutionalized word that are being used but they have to be reconsidered every year and also, they are spotted apart from the regular dictionary pages, occasionalisms do not belong to this kind of lexical units.

<sup>7</sup> <http://obneo.iula.upf.edu/spes/>

<sup>8</sup> <https://www.um.es/neologismos/index.php/>

<sup>9</sup> <https://blogscvc.cervantes.es/martes-neologico/>

<sup>10</sup> <https://antenario.wordpress.com/>

<sup>11</sup> *Antenas Neológicas* is the Spanish neology network (<https://www.upf.edu/web/antenas>) which gathers neology observatories from Spain and America.

<sup>12</sup> For English we find a similar type of dictionary whose aim is also to present uncommon, transgressive or playful words, such as *The Dord, the Diglot and an Avocado or Two: The Hidden Lives and Strange Origins of Common and Not-So-Common Words* (Garg 2007) or *Where a Dobdob Meets a Dikdik: A Word Lover's Guide to the Weirdest, Wackiest and Wonkiest Lexical Gems* (Casselman 2010).

<sup>13</sup> For example, Muñoz Laso (2018) collects more or less institutionalized neologisms of French origin, and Ramos (2000) collects Spanish colloquialisms.

<sup>14</sup> *La medición de la neologicidad y la dictionariabilidad de los neologismos del español* (*Neómetro*), ref. FFI2016-79129-P (AEI/FEDER, UE), funded by the Spanish *Ministerio de Economía y Competitividad*.

<sup>15</sup> <http://obneo.iula.upf.edu/bobneo/index.php>

<sup>16</sup> The dictionaries that are part of the exclusion corpus are the *Diccionario de uso del español de América y España* (VOX) and DLE.



- Usage analysis: date of the first occurrence, types of text where it is documented, frequency of period (2015-2019), total frequency (1989-2019), record and thematic area.
- Documentary analysis: presence in usage dictionaries (*CLAVE*<sup>17</sup>, *Diccionario de uso del español*, *Diccionario del español actual* and *Diccionario de argentinismos, neologismos y barbarismos*),<sup>18</sup> in standard dictionaries for other languages (Oxford English Dictionary, *Lo Zingarelli* for Italian, *Le Grand Robert* for French and the *Gran Diccionari de la Llengua Catalana* for Catalan), in neologism dictionaries (*Diccionario de voces de uso actual*, *Nuevo diccionario de voces de uso actual*, *Diccionario de neologismos online* and *Neologismos del español actual*) and also the presence in corpora (CREA and CORPES XXI).

In order to analyse the phenomenon of occasionalisms in Spanish, from the initial corpus of 3,402 units we have selected those that present a minimum frequency (a single occurrence) and that, moreover, do not appear in the dictionaries we have mentioned. After applying this search criterion, we have obtained a second corpus of 682 units, which constitutes our corpus of analysis of occasionalisms, although in order to observe specific phenomena we have further filtered the selection ensuring that they were not present in Factiva<sup>19</sup> database or in Google's textual macrocorpus,<sup>20</sup> which has allowed us to verify that 199 of these neologisms are totally casual. In the following section we offer the results obtained after submitting the data to a morphological analysis in order to identify tendencies in their formation.

## 4 Results and Analysis

In this section we offer the results of our qualitative analysis of the occasionalism corpus, in which we looked for tendencies in word formation. The first observation is the great diversity of these units: some have been formed by semantic or syntactic processes, other occasionalisms are loanwords from different languages, but for the most part they are formed by derivational morphological mechanisms (*hipermalvinizar*, *corbatismo*, *acalambrante*, *alabatorio*, *dicharachería*), compounding (*ciberniño*, *cazaacadémicos*) and blending (*gorn*, *graficleta*). This diversity of processes reflects the diversity that characterizes neologisms in general, but there are preferences for certain processes that differ from general tendencies in neologism formation, as we will show in section 4.1 on tendencies in occasionalism formation processes. Leaving aside the processes and entering into the concrete level of word-formation rules, it can be observed that transgressions in occasionalism creation allow us to establish several groups that we will present in section 4.2, which deals with tendencies in occasionalism-formation rules.

### 4.1 Tendencies in Formation Processes

Several researchers interested in the study of playful resources in word formation have studied blending (Hohenhaus 2007; Renner 2013), which they have considered a resource that tends to form ephemeral neologisms. We did the same in a previous work (Bueno 2019) in which we observed a considerable proportion of occasionalisms in blends<sup>21</sup> and this is corroborated in the corpus analysed for this research with examples such as the following:

- (1) Igual que los leggins, los pantalones de corte pirata y los **pantafaldas** (que son la suma de ambas opciones) también entrarán en los armarios de los más atrevidos [*La Vanguardia*, 26/02/2017]<sup>22</sup>
- (2) La posición de Raquel M. Adsuar, **bolillotuber** con más de 11.000 suscriptores, como protagonista de una masterclass en una feria de artesanía ha levantado las quejas de la Asociación de Bolilleras de la Comunidad Valenciana. [*El País*, 27/03/2017]

In the first example, *pantafalda* instead of the usual name already found in dictionaries (*falda pantalón*), an acronym has been chosen with an incomplete lexeme (*pantalón*) and the other complete lexeme (*falda*); and in the second case, *bolillotuber*, a complete lexeme (*bolillo*) and a truncated one (*youtuber*) have also been kept for a new expression, with non-explicit denominative purpose.

Compounding is also often chosen for occasionalism creation. As in blending, an important part of the expressive force of these units lies in the union of two lexemes in a single word when this union is not very predictable. This tendency has also been observed for English occasionalism formation (Hohenhaus 2007) as well as in French (Dal & Namer 2018). In our corpus we find examples of compounding such as the following:

- (3) Su horror y compasión (de los muertos) le inducen a mirar hacia otro tiempo venidero, habitado por gente que surge de la desmemoria y en la que se integra Ed, el **poeta-lavaplatos**, el último en abandonar la isla mítica una vez los muros han sido de molidos y en aquel final de 1989 asume la conciencia de su libertad. [*La Vanguardia*, 01/04/2017]
- (4) Vivimos en una sociedad **smartphone-céntrica** y necesitamos transformarla en una sociedad **humanocéntrica**. [*La Vanguardia*, 13/01/2018]
- (5) En Twitter, Puigdemont ha asegurado literalmente que ni los **asustafuncionarios** ni los **cazaurnas** podrán parar un referéndum que él mismo ha prometido que será vinculante. [*Cadena Ser*, 28/06/2017]

We note that in occasionalism formation, both N+N, A+A and V+N structures are used, and that in all of them two lexemes with non-complementary and sometimes antithetical semantic content are joined, a situation that is consistently repeated

<sup>17</sup> <http://clave.smdiccionarios.com/app.php>

<sup>18</sup> <http://www.bibliotecadigital.gob.ar/items/show/1308>

<sup>19</sup> <https://global.factiva.com>

<sup>20</sup> Due to the documentary noise produced by the platform's search engine, we set a limit of up to 100 appearances and checked that the pages in question had no content and were without lexical or documentary interest.

<sup>21</sup> In Bueno (2019) we observed that, despite being one of the least productive formation processes for neologism formation, the hapax content in blending was considerably high, reaching up to 75% of the number of units collected.

<sup>22</sup> In all examples the date and source of the text are given; when the source is not Spanish, the country of origin is indicated.



in the other examples (such as *plaza-basura*, *cholloweb*, *poético-kitsch* or *calienta-pavas*). In the previous examples it can be observed, moreover, that there is a high degree of contextual dependency in occasionalisms because the fun factor is transferred to the immediate discursive environment: the adjective *smartphone-céntrico* plays with *humanocéntrico* and *asustafuncionarios* plays with *cazaurnas*, forming pairs in the same context.

The use of Greek and Latin forms is also one of the tendencies observed in occasionalism formation because they give the resulting unit an appearance of denominative formality which makes it sound cultured or specialized. When coupled with unpredictable lexematic bases, the transgressive effect is evident and is often used with a clearly ironic or critical purpose, as we can see in examples such as the following:

(6) Esto es el tarificador, **calculógrafo** le llamaban, donde se ponía la hora en que empezaba la conferencia y cuando terminaba para poder establecer la tarifa. [*Radio RNE*, 12/03/2015]

(7) Doy al play a otra joya de la **vladivideoteca**, en la que Montesinos se acerca a un espejo para apretarse la corbata al cuello y se hace un guiño a sí mismo, en realidad a la cámara que se esconde detrás de su imagen. [*El País Semanal*, 26/03/2017]

(8) Pichetto se adjudicó la mejor marca del **abucheómetro**. [*Página 12* (Argentina), 27/02/2016]

To finish this overview of formation resources with a higher presence in occasionalism formation, we will refer to the formation of words on the basis of a proper name. It would be usual in this case for the proper names used for this purpose to be non-local and markedly timeless, but the playful character is sometimes achieved by using names of people (sometimes first and last names, sometimes just the first name and more frequently, the last name) or also names of places, brands or institutions.

(9) ¿Es ahora un GP menos **alonsodependiente**? [*La Vanguardia*, 01/04/2017]

(10) Se daban las condiciones para que una coalición **azañista-conservadora**, macerada en Sevilla, le metiera un buen tajo a la autonomía catalana, con el presidente del Gobierno lavándose las manos. [*La Vanguardia*, 23/05/2010]

(11) El 51% de los **melenchonistas** dará su voto a Macron el domingo. [*El País*, 06/05/2017]

(12) Pero Fernández es solo provisional, y Susana Díaz y Pedro Sánchez se acusan de **filopodemismo** o de encubierto derechismo. [*El Periódico de Catalunya*, 16/04/2017]

With these examples we wanted to show that the use of proper names is frequent in different word formation processes: the compounds in the first two examples, *alonsodependiente* and *azañista-conservador*, are based on the name of the Formula 1 pilot Fernando Alonso, and the Spanish politician Manuel Azaña; in the third example, *melenchonista* has been formed by suffixation on the name of the French politician Jean-Luc Mélenchon; and in the fourth one, *filopodemismo*, the suffixation has been made with the name of the Spanish political party *Podemos*. *Leganordiano* (related to the *Lega Nord* political party in Italy), *cyberlouvre* (related to the French Museum of Louvre), *puigdemontista* (relating to the Catalan politician Carles Puigdemont) or the *vladivideoteca* mentioned above and many others confirm the tendency of current Spanish occasionalism formation to be based on proper names from different sectors of society.

## 4.2 Tendencies in Formation Rules

In the occasionalism corpus a large number of examples have been found of what Hohenhaus (2007) calls *attention-seeking device*, linguistic elements that manage to capture the recipient's attention. Occasionalisms are usually themselves *attention-seeking devices* in the way they are formed, which transgresses a rule either by the choice of the base, by the union of bases with antithetical features, by mixing registers or by other means. In this section we will focus on several word-formation rules, specifically suffixation and neoclassical compounding, which we have observed are particularly efficient in occasionalism formation.<sup>23</sup>

The suffix *-ez* is used to form abstract feminine nouns that indicate the quality of the adjective from which they derive, but the rule can be transgressed by attaching the suffix to nominal bases (especially in the case of proper nouns) in order to create a striking effect, as we can see in the following examples:

(13) Inventor de los términos "putrefacto" y "carnuzo" para definir a un tipo de señor inmovilista y estatal, dado a la caspa en la hombrera y a la **bigotez**, un tipo, pues, que despedía el hedor de una España muerta pero insepulta. [*El País*, 23/07/1995]

(14) La humillación a la que se sometió al PSOE, no por parte del PP precisamente, sino por parte d'ERC con la **rufianez** que fue una cosa bastante notable y también el señor de Bildu, que tampoco estuvo mal, así como la actitud de Podemos. [*Cadena Ser*, 31/10/2016]

(15) Yo sigo pensando que ha perdido **raphaelez**. [*El País*, 14/07/2012]

In the following examples we can see verbs that have been made with the suffix *-ear*. Although they all derive from a noun, which is a regular behaviour for this suffix, the result is striking because of the noun chosen in every case, which is unlikely to be transformed into an action:

(16) El dj alemán **deephouseará** en mandarine, compartiendo cabina con el remixero inglés Nic Fanciulli. [*Página 12* (Argentina), 10/12/2015]

(17) Según usted también debieron ser asesinados Bolívar, Santander y demás guerrillos que medio liberaron a este pobre país del yugo español, Napoleón debería haber sido **falsopositivado**, Robespierre había sido mejor motosierrarlo y está muy bien que hayan asesinado al guerrillo Jesús de Nazaret. [*El Espectador* (Colombia), 26/04/2016]

(18) Entre las cartas de Clos abundan las entradas gratis para ir al Fòrum, que hace las funciones de la cárcel en el Monopoly **tradicioneando** a la baja. [*La Vanguardia*, 25/12/2004]

<sup>23</sup> In our research we are focusing on knowing what the productive mechanisms of occasionalisms are in order to cover this issue in lexicographical works.



Next, we will show examples of occasionalisms created by the transgression of word-formation rules. In this case they are surprising because Greek and Latin roots are prototypically used for terminological unit formation. The neoclassical combining form *-metro*, for example, is used to name measuring devices, such as *pluviómetro* for rain or *termómetro* for temperature, where the base of this word is formal. For its part, *-itis* is used to name body part infections, such as *otitis* or *hepatitis*, where the base is also formal. The transgressive rupture that we observe in the following occasionalisms is found in the selection of the base, with the result somehow representing an instrument (*-metro* examples) or a metaphorical infection (*-itis* examples):

(19) El **masomenómetro** argentino trabajando a full. [*La Nación* (Argentina), 07/06/2011]

(20) El primer nivel de nuestro **fermentómetro** no es muy arriesgado: son productos fermentados que "forman parte de nuestra cultura", como afirma Samy Ali, chef de La Candela Restó. [*Icon*, 06/05/2017]

(21) Hay un interés recurrente por inocular el virus de la **consensualitis**, lamentó el representante de ERC en alusión al hecho de que sean PP y PSOE los que decidan sobre los territorios. [*La Vanguardia*, 09/02/2005]

(22) El pragmatismo y la determinación de la Fed se ha llevado buena parte del miedo reinante en el ambiente y aunque todo mundo sabe que las cosas no se arreglan de un plumazo, la **catastrofitis** se ha evaporado. [*La Vanguardia*, 20/09/2007]

We end the examples of tendencies in word-formation rules in occasionalisms with different cases that also involve neoclassical combining forms: *macro-* on the one hand and *ciber-* on the other:

(23) Casado y con tres hijas, sus firmes convicciones de izquierda e igualdad y su lealtad responden al perfil que el presidente socialista quiere para su **macroministro** de Trabajo y Asuntos Sociales. [*El Sur*, 18/04/2004]

(24) Si en lugar de centrarse en ir a por **macroinfluencers** o celebrities, que son muy caros, van a por microinfluencers que son más pequeños, van a llegar a menos gente pero van a tener más influencia en su público objetivo. [*La Vanguardia*, 23/10/2016]

(25) Una particular **ciberrepública** para trabajar, vivir y luchar. [*El País*, 05/05/2008]

(26) [...] artefactos de defensa e identificación, protección para guardianes, **ciberalambradas** que reciben los funcionarios, etc. [*El País*, 24/12/2004]

Occasionalisms formed with *macro-* are part of a larger group of occasionalisms, also frequently used, formed with a number of positive or negative intensifiers (*megamentimedio*, *minibarbi*, *microacoso*, *macrofortuna* or *nanocarrera*, among others); these are semantically transparent examples whose expressive value lies in their hyperbolic character. Occasionalisms formed with *ciber-*, on the other hand, have been on the increase in recent years, when this neoclassical combining form has demonstrated a plasticity that allows it any type of combination with any syntactic category and with different functions: several are denominative and have become institutionalized in speakers' usage (*ciberataque* or *ciberdelincuencia*), and some have even become part of the standard dictionary (*cibercafé* or *cibercultura*), while others, such as *ciberalambrada* and *ciberrepública*, clearly have a playful or critical function.

Besides the types of formation that we have chosen in order to illustrate that occasionalisms present regular tendencies, we find in our corpus other forms created by suffixation and neoclassical compounding, and by other word-formation processes that we will attempt to cover in future works.

## 5 Conclusion

The bibliography presents occasionalisms as volatile language events, difficult to detect and, because of their transgressive and context-dependent nature, linked to linguistic creativity and productivity. Our research confirms that the marked creativity they display is not at variance with their following certain recurring patterns; we have shown these by giving examples of the processes which are most productive of occasionalisms. Occasionalisms are not random linguistic acts.

This research has allowed us to take a step forward in understanding occasionalisms themselves. Our observation of the data corroborates the traits that the literature attributes to occasionalisms: lexical units that fulfil an expressive function in a specific context and do not follow the usual development of a neologism as they do not start the process of institutionalization after their first appearance and remain in the initial state of *nonce-word*. However, our observation also casts a doubt on whether all occasionalisms meet these features to the same extent, since, nowadays, with the existence of enormous dynamic word corpora, it is easy to find different casual occurrences of the same word, either because different speakers in different contexts can produce the same occasionalism or because an occasionalism can be reproduced by others (specially in social media) but without setting in motion the process of social institutionalization, grammatical lexicalization and cognitive hypostatization that would make it lose its occasionalism nature.

## 6 References

- Algeo, J. (1993). Desuetude among new English words, *International Journal of Lexicography* 6 (4), pp. 281-293.
- Alvar, M. (dir.) (1994). *Diccionario de voces de uso actual*. Madrid: Arco Libros.
- Alvar, M. (dir.) (2003). *Nuevo diccionario de voces de uso actual*. Madrid: Arco Libros.
- Delvaille, B. & Zylberstein, J. C. (2003). *Dictionnaire des mots rares et précieux*. Paris: 10.
- Brisac, G. (2018). *Mes mots sauvages*. Paris: Points.
- Buchanan, M. A. (1927). *A graded Spanish word book*. Toronto: University of Toronto Press.
- Bueno, P. J. (2019). *Entre el ocasionalismo y el neologismo: hacia la delimitación del concepto de neologismo y su institucionalización*, Unpublished PhD project, Barcelona: Universitat Pompeu Fabra.
- Cabré, M. T. (1989). *La neología efímera*. In J. Bastardas (1989). *Miscel·lània Joan Bastardas*. Barcelona: Abadía de Montserrat, 37-58.



- Cabré, M. T. (2015). *Bases para una teoría de los neologismos léxicos: primeras reflexiones*. In I. M. Alves, E. Simões Pereira (eds.), *Neologia das Línguas Românicas*. São Paulo: CAPES; Humanitas, pp. 79-107.
- Cabré, M. T., Domènech, O., Estopà R, Freixa, J. & Solé, E. (2004). "La lexicografia i la identificació automatitzada de neologia lèxica" en *De Lexicografia. Actes del I Symposium Internacional de Lexicografia*. Barcelona: IULA / Universitat Pompeu Fabra.
- Casselmann, B. (2010). *Where a dobdob meets a dikdik: a word lover's guide to the weirdest, wackiest, and wonkiest lexical gems*. Avon, Mass: Adams Media.
- Dal, G., Namer, F. (2018). Playful nonce-formations in French: Creativity and Productivity. In S. Ardnt-Lappe, A. Braun, C. Moulin & E. Winter-Froemel (eds.), *Linguistic Innovation, Morphological Productivity, and Ludicity*. De Gruyter.
- Garg, A. (2007). *The dord, the diglot, and an avocado or two: the hidden lives and strange origins of words*. New York, N.Y.: Plume Book.
- González, C. (2006). *Clave: diccionario de uso del español actual*. Madrid: SM.
- Grup Enciclopèdia Catalana (1998). *Gran diccionari de la llengua catalana*. Barcelona: Enciclopèdia Catalana.
- Guilbert, L. (1975). *La créativité lexicale*, Paris: Larousse.
- Haensch, G., Omeñaca, C. (2004). *Los diccionarios del español en el siglo XXI: problemas actuales de la lexicografía, los distintos tipos de diccionarios: una guía para el usuario, bibliografía de publicaciones sobre lexicografía*. Salamanca: Ediciones Universidad de Salamanca.
- Hohenhaus, P. (2007). How to do (even more) things with nonce words (other than naming). In J. Munat, *Lexical Creativity, Texts and Contexts*, Nottingham: John Benjamins, pp. 15-38.
- Horvilleur, A. (2013). *Dictionnaire des mots rares et savoureux*. Lyon: J. Andre.
- Lipka, L., Handl S. & Falkner, W. (2004). Lexicalization & Institutionalization. The State of the Art in 2004. *SKASE Journal of Theoretical Linguistics*, 1 (2), pp. 2-19.
- Matoré, G. (1952). "Le néologisme: naissance et diffusion", *Le français Moderne* 20 (2), pp. 87-92.
- Moliner, M. (1966). *Diccionario de uso del español*. Madrid: Gredos.
- Moliner, M. (2013). *Neologismos del español actual*. Madrid: Gredos.
- Muñoz Laso, F. (2018). *Los neologismos vienen de París*. Madrid: Liber Factory.
- Observatori de Neologia, IULA, DNOL, *Diccionario de neologismos on line*. Accessed at: <http://obneo.iula.upf.edu/spes/> [01/03/2020]
- Ramos, A., Serradilla, A. M. (2000). *Diccionario Akal del español coloquial: 1,492 expresiones y más*. Madrid: Akal Ediciones.
- Real Academia Española: *Diccionario de la lengua española*, 23.<sup>a</sup> ed., [v. 23.3 online] Accessed at: <https://dle.rae.es> [02/03/2020]
- Rey, A. (1976). Néologisme: un pseudo-concept? *Cahiers de lexicologie*, 28(1), pp. 2-17.
- Rey, A. (2013). *Le grand Robert : le dictionnaire le plus complet de la langue française*. Paris: Le Robert.
- Rheims, M. (1989). *Dictionnaire des mots sauvages: écrivains des XIXe et XXe siècles*. Paris: Larousse.
- Sablayrolles, J.F. (2018). Introduction, *Neologica* 12 (1), pp 1-12
- Schmid, H. (2008). "New words in the mind: concept-formation and entrenchment of neologisms", *Anglia* 126 (1), pp. 1-36.
- Seco, M., Ramos, G. & Andrés, O. (1999). *Diccionario del español actual*. Madrid: Aguilar. (Cited DEA).
- Segovia, L. (2018). *Diccionario de Argentinismos, Neologismos y Barbarismos*. Buenos Aires: Forgotten Books
- Simpson, J. A., Weiner, E. S. C. (1989). *The Oxford English Dictionary*. Oxford: Clarendon Press.
- Štekauer, P. (2002). On the theory of neologism and nonce-formations. *Australian Journal of Linguistics*. 22 (1), pp. 97-112.
- Universidad de Murcia, NEOMA, *Diccionario de neologismos del español actual* [online] Accessed at: <https://www.um.es/neologismos/index.php/> [01/03/2020]
- Vox Editores (2004). *Vox: diccionario de uso del español de América y España*. New York: McGraw-Hill.
- Zingarelli, N., Cannella, M. & Lazzarini, B. (2019). *Lo Zingarelli: vocabolario della lingua italiana*. Bologna: Zanichelli.

### Appendix 1: Occasionalisms, Translation and Context

Occasionalism	Translation	Context	Translation & Notes
neohisteria	neohysteria	"Y entramos en un proceso de neohisteria. Y buscamos respuestas en el yoga, las reparias y las religiones alternativas)".	And we start a process of <b>neohysteria</b> . And we look for answers in yoga and in alternative religions.
cernunedar	to cernudate	"El deseo no se describe, observa Clavecí; los poemas sí. Y sigue cernudeando".	Wish cannot be described -Clavecí states- poems can. And she keeps <b>cernudating</b> . [Related to the poet Cernuda]
hipermalvinizar	to hypermalvinize	Las opciones de la Argentina hoy no pueden oscilar entre "desmalvinizar" o *hipermalvinizar* la política exterior del país.	The options of Argentina today cannot range between "demalvinaze" or " <b>hypermalvinize</b> " the foring policy in the country.
corbatismo	tieism	Pero el *corbatismo*, en su impecable trayectoria tiene una mancha imperdonable.	But there is an unforgivable stain in the <b>tieism</b> faultless trend.



acalambrante	electric-shocking	El <i>*acalambrante*</i> mar es siempre caldo de tiburones que parece que sonríen, pero que en realidad no hacen otra cosa que enseñar los dientes.	The <b>electric-shocking</b> sea is always place for apparently smiling sharks which are really showing their teeth.
alabatorio	praisory	Hallaron un extenso y <i>*alabatorio*</i> reportaje a la candidata de la lista Celeste en la página web que dirige la hija de Gils Carbó.	They found a vast and <b>praisory</b> report on the candidate in the Celeste list in the web site run by Gils Carbó's daughter.
dicharachería	chattyness	Se habla de la protección del entorno y de secretos de familia, pero todo queda en <i>*dicharacherías*</i> .	We talk about environment protection and family secrets, but it's all <b>chattyness</b> .
ciberniño	cyberboy/girl	El mercado de los juguetes se dirige a <i>*ciberniños*</i> cada vez más pequeños.	The toy market is adressed to increasingly younger <b>cyberboys</b> .
cazaacadémicos	academic slayer	Buffy, la <i>*cazaacadémicos*</i> . ¿Twin Peaks?, ¿The wire? Frío, frío. La serie más analizada por los investigadores del mundo es Buffy la Cazavampiros	It's a gameplay between Buffy, the Vampire Slayer and "Buffy, the <b>Academic Slayer</b> ".
gorn	gorn	<i>*Gorn*</i> es la fusión del gore y el porno; fuckwash (revisión del brainwash o lavado de cerebro) es follarse a alguien hasta cambiar sus convicciones.	<b>Gorn</b> is the combination of gore and porn; fuckwash (a new version of brainwash) is fucking with someone to make them change their convictions.
graficleta	graffbyke	La <i>*graficleta*</i> es un proyecto de Joshua Kinberg inspirado en el célebre GraffitiWriter, en que las tecnologías digitales permiten convertir las bicis en un inédito medio de comunicación	The <b>graffbyke</b> is a project by Joshua Kinberg inspired in the famous GraffitiWriter.
pantafalda	trouskirt	Igual que los leggins, los pantalones de corte pirata y los <i>*pantafaldas*</i> (que son la suma de ambas opciones) también entrarán en los armarios de los más atrevidos.	Gameplay between <i>pantalón</i> (trousers) and <i>falda</i> (skirt).
bolillotuber	bobbintuber	La posición de Raquel M. Adsuar, <i>*bolillotuber*</i> con más de 11.000 suscriptores, como protagonista de una masterclass en una feria de artesanía ha levantado las quejas de la Asociación de Bolilleras de la Comunidad Valenciana.	Bolillo is a technique in which we use wooden thread bobbins to create some kind of confection. The youtuber teaches this technic and the gameplay is made with these two words <i>bolillo</i> and <i>youtuber</i> .
poeta-lavaplatos	dishwasher-poet	Ed, el <i>*poeta-lavaplatos*</i> , el último en abandonar la isla mítica una vez los muros han sido de molidos y en aquel final de 1989 asume la conciencia de su libertad.	Ed, the dishwasher-poet, the last to leave the island.
smartphone-céntrica	smartphone-centric	Vivimos en una sociedad <i>*smartphone-céntrica*</i> y necesitamos transformarla en una sociedad humanocéntrica.	We live in a smartphone-centric society and we need to turn it into a human-centric society.
asustafuncionarios	civil servant-buster	En Twitter, Puigdemont ha asegurado literalmente que ni los <i>*asustafuncionarios*</i> ni los cazaurnas podrán parar un referéndum que él mismo ha prometido que será vinculante.	In Twitter, Puigdemont has literally stated that neither the <b>civil servant-busters</b> nor the ballot box-hunters will be able to stop the referendum.
cazaurnas	ballot box-hunter	En Twitter, Puigdemont ha asegurado literalmente que ni los <i>*asustafuncionarios*</i> ni los cazaurnas podrán parar un referéndum que él mismo ha prometido que será vinculante.	In Twitter, Puigdemont has literally stated that neither the civil servant-busters nor the <b>ballot box-hunters</b> will be able to stop the referendum.
plaza-basura	rubbish-position	Se limitaron a protagonizar un encuentro de equipos mediocres y poco estructurados que en vez de aspirar a la Premier League compitieron buscando una <i>*plaza-basura*</i> para la próxima Liga de Campeones.	They were looking for a "trash-position" in the next Champion League instead of fighting for a good one.
cholloweb	bargain-web	Las <i>*chollowebs*</i> para ahorrar, intercambiar y compartir viven su edad de oro.	<b>Bargain-webs</b> to save, change and share money are living their best time.
poético-kitsch	poetic-kitsch	Un ambivalente manifiesto feminista cargado de esteticismo <i>*poético-kitsch*</i> y humor negro, la castración sin paliativos del único personaje masculino de la película fue más de lo que algunos estómagos pudieron soportar.	A polivalent feminist manifest full of <b>poetic-kitsch</b> skepticism and black humour.
caliente-pavas	chick-turner	Un ajedrez <i>*caliente-pavas*</i> que ha mezquinado información y al mismo tiempo ha dado de probar carne a las fieras constantemente.	A <b>chick-turner</b> chess that has skimped on information.
calculógrafo	calculographer	Esto es el tarificador, <i>*calculógrafo*</i> le llamaban, donde se ponía la hora en que empezaba la conferencia y cuando terminaba para poder establecer la tarifa.	This is the rating machine, or <b>calculographer</b> , where the start and finish time of the lecture appears to fix the tariff.
vlavidioteca	vlavideo library	Doy al play a otra joya de la <i>*vlavideoteca*</i> .	I press <i>play</i> on another gem in the <b>vlavideo library</b> .
abucheómetro	boometer	Pichetto se adjudicó la mejor marca del <i>*abucheómetro*</i> .	Pichetto got the best results in the <b>boometer</b> .



alonsodependiente	alonsoaddict	¿Es ahora un GP menos *alonsodependiente*?	Is a less <b>alonsoaddict</b> GP right now?
azañista-conservador	azañist-conservative	Se daban las condiciones para que una coalición *azañista-conservadora*, macerada en Sevilla, le metiera un buen tajo a la autonomía catalana, con el presidente del Gobierno lavándose las manos.	The perfect conditions arose to let an <b>azañist-conservative</b> coalition, marinated in Seville, struck the catalan autonomy with the President coping out.
melenchonista	melenchonist	El 51% de los *melenchonistas* dará su voto a Macron el domingo.	51% of <b>melenchonists</b> would give their vote to Macron on Sunday.
filopodemismo	philopodemism	Pero Fernández es solo provisional, y Susana Díaz y Pedro Sánchez se acusan de *filopodemismo* o de encubierto derechismo.	But Fernandez is just provisional, and Susana Díaz and Pedro Sánchez are accusing each other of <b>philopodemism</b> .
puigdemonista	puigdemonist	No hay dudas sobre el referéndum, pero se rechaza el "seguidismo" que desde el Parlament se practica con la CUP, que desdeña el modelo social del PDECat y volverá al discurso de la mafia, pero se declara *puigdemonista*.	They will retake the mafia discourse, but they declares themselves <b>puigdemonists</b> .
bigotez	moustachity	Para definir a un tipo de señor inmovilista y estatal, dado a la caspa en la hombrera y a la *bigotez, un tipo, pues, que despedía el hedor de una España muerta.	[...] to define a kind of state and immovelist gentleman, prone to dandruff, to shoulder pads and to <b>moustachity</b> , a guy who emits the stench of a dead Spain.
rufianez	rufianity	La humillación a la que se sometió al PSOE, no por parte del PP precisamente, sino por parte d'ERC con la *rufianez* que fue una cosa bastante notable y también el señor de Bildu, que tampoco estuvo mal, así como la actitud de Podemos.	PSOE was subjected to humillitation, not by PP but by ERC with remarkably <b>rufianity</b> .
raphaelez	raphaelity	Yo sigo pensando que ha perdido *raphaelez*.	I still think he has lost a lot of <b>raphaelity</b> . [Related to the Spanish singer Raphael]
deephousear	to deephouse	El dj alemán Matthias Tanzmann *deephouseará* en Mandarine, compartiendo cabina con el remixero inglés Nic Fanciulli.	German dj Matthias Tanzmann will <b>deephouse</b> in Mandarine sharing stand with English remixer Nic Fanciulli
falsopositivar	to fake-positive	Napoleón debería haber sido *falsopositivado*, Robespierre había sido mejor.	Napoleon should have been <b>fakepositived</b> , Robespierre would have been better.
tradicionear	to tradition	Entre las cartas de Clos abundan las entradas gratis para ir al Fòrum, que hace las funciones de la cárcel en el Monopoly *tradicioneando* a la baja.	Among Clos's letters there are plenty of free tickets to go to Fòrum, which makes Monopoly prision fuctions <b>traditioning</b> low.
masomenómetro	moreorlessometer	El *masomenometro* argentino trabajando a full.	The Argentinian <b>moreorlessometer</b> working "a tope".
fermentómetro	fermentometer	El primer nivel de nuestro *fermentómetro* no es muy arriesgado: son productos fermentados que "forman parte de nuestra cultura".	The first level of our <b>fermentometer</b> is not very risky: they are fermented products that are part of our culture.
consensualitis	consensualitis	Hay un interés recurrente por inocular el virus de la *consensualitis*, lamentó el representante de ERC en alusión al hecho de que sean PP y PSOE los que decidan sobre los territorios.	There is a frequent interest in inoculating <b>consensualitis</b> virus.
castastrofitis	catastrophitis	El pragmatismo y la determinación de la Fed se ha llevado buena parte del miedo reinante en el ambiente y aunque todo mundo sabe que las cosas no se arreglan de un plumazo, la *castastrofitis* se ha evaporado.	Even if everybody knows things cannot be fixed at a stroke, <b>catastrophitis</b> has gone.
macroministro	macrominister	Casado y con tres hijas, sus firmes convicciones de izquierda e igualdad y su lealtad responden al perfil que el presidente socialista quiere para su *macroministro* de Trabajo y Asuntos Sociales.	His character matches the profile the socialist President wants for his <b>macrominister</b> .
macroinfluencer	macroinfluencer	"Si en lugar de centrarse en ir a por *macroinfluencers* o celebrities, que son muy caros, van a por micro-influencers que son más pequeños, van a llegar a menos gente pero van a tener más influencia en su público objetivo", asegura Noelia Herrero, que ya trabaja en Digital Commerce de Desigual.	If they don't focus on <b>macroinfluencers</b> or celebrities, who are very expensive, but in micro-influencers who are smaller, they will get to fewer people but they will have greater influence on their objective audience.
ciberrepública	cyberepublic	Una particular *ciberrepública* para trabajar, vivir y luchar	A particular <b>cyberepublic</b> to work, live and fight for.
ciberalambrada	ciberwirefence	...artefactos de defensa e identificación, protección para guardianes, *ciberalambradas, que reciben los funcionarios...	Defence and identification mechanisms, protection for guardians and <b>ciberwirefences</b> the civil servants get.







# Arabic Loanwords in English: a Lexicographical Approach

Fournier P., Latrache R.

Sorbonne Paris Nord University, France

## Abstract

This article deals with Arabic loanwords in English from a lexicographical perspective. To create a representative corpus of Arabic loanwords in English, items are extracted from the *Oxford English Dictionary* database (henceforth *OED*) with an etymological advanced search. Among the criteria affecting the etymological tagging, the concept of two languages of origin is probably the most difficult one for lexicographers to deal with. This study presents some of the issues lexicographers are faced with in the dictionary-making process. Following that, Arabic loanwords are classified according to semantics, along with the date of their first attestation in the *OED* database. This quotation dating work that the *OED* systematically performs is not only an immense task, but also an essential one, as it enables researchers to determine the semantic spheres these corresponding loanwords are integrated into, as well as the cultural relationship between Arabic-speaking countries and English-speaking countries.

**Keywords:** Arabic loanwords; contemporary English; dictionary-based study

## 1 The Arabic Language

Arabic is considered as one of the major languages with a tremendous cultural impact in the world (Sapir 1921). According to Salloum & Peters (1996) 6500 Arabic loanwords are attested in the English language, though many of them have been introduced through the Spanish language (in Thawabteh 2011: 104). Indeed, Serjeantson (1935: 213-220) determines that Arabic loanwords can be either direct (i.e. with no intermediate language between Arabic and English) or indirect (with notably French or Spanish as transitory languages). She underlines the impact of Arabic on the English language (1935: 213): “It is from Arabic that English has borrowed the greatest number of Eastern loan-words, though it is true that a considerable proportion of them have not come to us direct”. The best explanation is linked to science (mathematics, astronomy...), business factors (especially during the 14<sup>th</sup> century in North Africa) and exploration as well. If older loanwords were rooted in science, the latter ones were representative of everyday life, in zoology and religion as Thawabteh’s description of semantic fields clearly demonstrates (2011). Darwish (2015: 106) offers a historical perspective of the Arabic loanwords in English: “Wilson (2001) notices that by the eighth century in North Africa, Arabic had ousted Latin as the dominant language; by the eleventh and twelfth centuries, Arabic civilisation had fully spread through Spain.” Thawabteh (2011:114) indicates that “[t]he development of Arab architecture, particularly in Granada, Seville, and Cordova was a catalyst for numerous borrowings.” The most prolific periods of borrowing were during the Middle Ages and the Renaissance when “English speakers came into contact with the prestigious intellectual centres of the Arab World” (Darwish 2015: 107).

The result was:

a flow of borrowings from Arabic into English, primarily in the fields of chemistry, medicine, philosophy, mathematics, astronomy, optics, physics, botany, literature, religion (chiefly Islam), music, warfare, shipping, trade, architecture, geography, government and sovereignty (Daher 2003 in Darwish 2015:107).

## 2 Corpus Building: Methods, Problems and Limits

Several factors can affect the creation of a representative database of Arabic loanwords in English. The corpus is elaborated thanks to the *OED* database using the advanced research tool. When mentioning “Arabic” as the language of origin and the main criterion of the advanced search, this first extraction attests 511 items.<sup>1</sup> Yet it appears that 102 words should be discarded because they do not truly correspond to the notion of “Arabic loanword.” First, some extracted words are morphologically composed of stems/roots of Arabic origin and English suffixes (examples: *Bohairic*, *Fatimite*, *Hanbalite*, *Saadian*). Such hybrid words are not representative of Arabic loanwords, but they show that these items can be truly integrated into the English lexicon, so that they can be affected by the word-formation rules of English. Approximate translations or rewriting of Arabic words (examples: *Hobson-Jobson*, *mockery*, *nugger*, *sheregrig*) are not kept in our analysis. Finally, items which are ultimately of Arabic origin, but whose transfer into English was made thanks to intermediate languages, are not kept either. The direct borrowing parameter is an ambiguous one and there may be controversy about it. Indeed, Serjeantson (1935) admits that Arabic loanwords can be either direct or indirect. Mossé (1943) assumes, concerning French loanwords in English, that whatever the ultimate origin of loanwords may be, the language which provides the new item should be considered as the source language. Well-known examples exist such as

<sup>1</sup> 511 items at the time of creating the database. The *OED* frequently adds more items that need further research.



*giraffe*, which is an Arabic word. However, it is considered to be a French loanword in English because it was transmitted to English through the French language. Therefore, researchers must be careful not to be mistaken by etymological information provided by dictionaries, which can include precise details concerning the ultimate origin of words and the source language/target language question. This analysis of Arabic loanwords is dictionary-based and the etymological tagging of the *OED* is adopted. However, some examples of non-selected words are listed below along with the etymological information provided by the *OED*. Some of the emblematic Arabic words are, interestingly enough, not tagged as Arabic loanwords because Arabic is not the transmission language.

The following examples are particularly relevant because they belong to the religious semantic domain of Islam, and since Arabic is the language of Islam and of its holy book, the Qur'an, the ultimate origin of these examples is necessarily Arabic. Hence as an example, the word *Islam* is not tagged as a direct borrowing from Arabic, but as being of multiple origins.

(1) '*Islam*'. (ID 99980)

Origin: Of multiple origins. Partly a borrowing from Turkish. Partly a borrowing from Arabic.

Etymons: Turkish *islām*; Arabic *islām*.

Frequency (in current use): 6

The word *Islam* had probably been transferred from Arabic into Turkish long before the first contact between the two languages, English and Arabic, was made. Thus, it is highly probable that English speakers may have had contacts with both Turkish and Arabic, which is the reason why the *OED* database could not determine exactly where the word had originated. The *OED* suggests some linguistic contact with the "strongly Arabized Ottoman Turkish language":

Contact between English speakers and Islam in the early modern period was chiefly through the institutions and peoples of the Ottoman Empire, which controlled south-eastern Europe, the Middle East, and North Africa and used the strongly Arabized Ottoman Turkish language.

Another example is related to the two major Muslim feasts; '*Eid-al-Adha*' and '*Eid-al-Fitr*', which are Arabic words, but which are tagged as follows by the *OED*:

(2) *Eid*. (ID 242685)

Origin: Of multiple origins. Partly a borrowing from Persian. Partly a borrowing from Arabic. Partly formed within English, by compounding.

Etymons: Persian *ʿīd*; Arabic *ʿīd*; *Eid-al-Adha* n., *Eid-al-Fitr* n.

Frequency (in current use): 4

(3) *Eid-al-Adha*. (ID 243484)

Origin: Of multiple origins. Partly a borrowing from Persian. Partly a borrowing from Arabic",

Etymons: Persian *ʿīd-i aẓḥā*; Arabic *ʿīd al-aḏḥā*

Frequency (in current use): 2

(4) *Eid-al-Fitr*. (ID 91137)

Origin: Of multiple origins. A borrowing from Arabic; modelled on a Persian lexical item.

Etymons: Arabic *ʿīd al-fīṭr*.

Frequency (in current use): 2

The above *OED* tagging is interesting. These words are semantically related, yet in the last item it appears that the origin is accounted for differently. The same argument evoked concerning Turkish can be used for the first two loanwords (*Eid* and *Eid-al-Adha*). First, '*Eid*' in Arabic means "a celebration of". Its listing as a borrowing from Persian suggests that it is a word from the pre-Islamic period. However, all the quotations used show that the word is related to Islam. Moreover, Persian uses other words to translate "celebration" and '*Eid*' is not one of them. It proves that Persian borrowed those two words from Arabic and that these loanwords may have been integrated into English through the exposure to the Persian language. However, the reference to the Persian origin is not consistent, as the last one is listed as "modelled on a Persian lexical item". *Eid-al-Fitr* is said to be an Arabic loanword, modelled on a Persian lexical item. This argument is problematic as it refers to two different processes. If this is an Arabic loanword, it means that Arabic is considered to be the source language. Yet as it is ultimately an Arabic word, how could it possibly be modelled on a Persian lexical item? The word *Ramadan* is another example of etymological tagging, showing that when lexicographers are confronted with deciding between several possible languages of origin, those languages are simply listed.

(5) *Ramadan*. (ID 157727)

Origin: Of multiple origins. "Of multiple origins. Partly a borrowing from Persian. Partly a borrowing from Turkish. Partly a borrowing from Arabic."

Etymons: Persian *ramāzān*, Turkish *ramāzān*; Arabic *ramāḍān*.

Frequency (in current use): 4

*Ramadan*, being the holy month for Muslims, can only be of Arabic origin. Its etymons in Persian and Turkish are probably mere pronunciation variations of the Arabic word.

The corpus of Arabic loanwords this study is based upon is composed of 409 items (see the whole corpus in Annex 1 along with the list of non-selected words in Annex 2). The Arabic loanwords will be analysed through different angles in the next sections: dating work, semantic domains and frequency. The last part deals with the *OED* quotations and the



impact data collectors may have on the influx of Arabic loanwords in English.

### 3 Influx of Arabic Loanwords Through Centuries

Initially when the dates of the first attestation of the Arabic loanwords in the *OED* were examined, it was apparent that the influx of Arabic loanwords per century was inconsistent. This method, which consists of determining the influx of loanwords thanks to the *OED* quotations, obviously has some limitations. These restrictions are linked to the *OED* data collectors (Gilliver 2015) (this point will be dealt with in Section 6). Table 1 classifies the number of Arabic loanwords integrated into English century by century.

14 <sup>th</sup>	15 <sup>th</sup>	16 <sup>th</sup>	17 <sup>th</sup>	18 <sup>th</sup>	19 <sup>th</sup>	20 <sup>th</sup>
1	6	29	73	76	162	62

Table 1: Number of Arabic loanwords integrated into English through centuries.

It turns out that the beginning of the phenomenon can be dated back to the 17<sup>th</sup> century with a peak in the 19<sup>th</sup> century with 162 loanwords out of 409 (i.e. 39.6%), which had been integrated into the English language. The correlation between these figures and the study of semantic domains in the next section may explain why the number of Arabic loanwords reached such a culmination in the 19<sup>th</sup> century.

### 4 Semantic Domains

A detailed survey of the semantic domains of the 409 Arabic loanwords in English along with the number of loanwords borrowed per century is presented in Table 2. Only semantic areas with at least 8 loanwords are integrated into the table. The other semantic domains are considered as minor ones. The corpus covers seven centuries, from 1393 to 1996, a period during which the *OED* database attested 113 items belonging to the religious semantic domain and 53 items belonging to the science semantic domain. Those two domains are unquestionably the most prolific ones. Science is the oldest semantic field Arabic loanwords are related to in the *OED*. The first direct Arabic loanword referenced by the *OED* is *Aldebaran*, which dates back to 1393, belonging to the semantic field of astronomy. Indeed, science brings together such domains as botany (example: *sebesten*), astronomy (example: *Al Nath*) or chemistry (example: *azoth*). If at first, only scientific Arabic words were borrowed, this phenomenon changed from the 16<sup>th</sup> century. Indeed, from then on, Arabic loanwords covered a large range of semantic areas such as religion, commerce, food and drinks and so on... Furthermore, it is worth mentioning that in the 17<sup>th</sup> century, the religious semantic domain began to prevail over the other semantic areas and that this field generated more loanwords than the totality of the loanwords from all the other categories.

This table raises essential issues. First, one can wonder why the borrowing process between Arabic and English, which was very limited before the 16<sup>th</sup> century, suddenly accelerated and covered various semantic fields. Then, Table 1 above shows that there was a massive influx of Arabic loanwords in the 19<sup>th</sup> century. How could these two phenomena be accounted for? The survey of *OED* quotations might be of interest in that respect.

	14 <sup>th</sup> c.	15 <sup>th</sup> c.	16 <sup>th</sup> c.	17 <sup>th</sup> c.	18 <sup>th</sup> c.	19 <sup>th</sup> c.	20 <sup>th</sup> c.	total
religion	/	/	5	30	20	48	10	113
science	1	4	8	6	7	25	2	53
clothing	/	/	2	1	7	14	1	25
food / drinks	/	/	1	3	2	9	9	24
titles	/	/	2	6	5	6	2	21
commerce	/	/	1	3	4	2	3	13
arts	/	/	/	2	4	6	/	12
weight	/	/	5	/	3	/	/	8
desert	/	/	/	1	1	3	3	8
sailing	/	/	/	3	/	3	2	8
geopolitics	/	/	/	/	/	/	8	8

Table 2: Number of Arabic loanwords classified by semantic domains along with centuries.

But before dealing with *OED* quotations, the next section investigates the frequency of Arabic loanwords in English.

### 5 Frequency

The *OED* frequency tool is used to investigate frequency in Arabic loanwords (see Annex 3 for the whole list of



loanwords sorted out with frequency numbers). It classifies words from 0 to 8, 0 corresponding to obsolete words and 8 to words very frequently used in English. For example, the most used Arabic loanwords in English are *coffee* and *Muslim*, which are classified 6 out of 8 on the frequency scale. The majority of Arabic loanwords are ranked 2 or 3 out of 8 and it corresponds to 303 items (i.e. 74.1% of the whole corpus). Some words in these categories are used in everyday life (examples: *hammam*, *falafel*, *harissa*), but there are also words, which are used in such restricted semantic areas that native English speakers are probably not familiar with them. The 4 out of 8 classification is subjectively considered to be the clear-cut separation between rare words and more attested ones. Indeed, such words as *giraffe* or *Sahara*, which are well-known Arabic loanwords, belong to this category. Table 3 gives examples of Arabic loanwords, from the less frequent items to the most attested ones.

-	- / +	+
<i>cabeer, caroteel / carotel, fana, gaiassa, ghazeeyeh, halawi, iggri / iggry, kuphar, nil, quaiss kitir, rottol, shereefa, tahalli, Takbir, Zarnich</i>	<i>adhan, alim, ardeb, arrack, bejel, burgoo, daman, doum, fennec, fitna, hadj, hakim, halal, halfa, hamza, hawala, hijab, ihram, jebel, jihadi, jinnee, jol, kantar, kazi, khamsin, khan, khatib, khor, khula, khutbah, kohl, lablab, lebbek, loofah, madhhab, madrasah, maghrib</i>	<i>diss, Iraqi, Koran, Qur'an, Saudi, Sufi, Sunni, sunt, Swahili, coffee, Muslim</i>

Table 3: Examples of Arabic loanwords based on frequency parameters.

## 6 Working with the OED Quotations

The last section deals with the *OED* quotations and the impact this tremendous work of data collecting may have on the perception of Arabic loanwords. A detailed analysis shows that a consequential number of quotations came from the same sources. For instance, 50 quotations (first and second citations) in the 19<sup>th</sup> century came solely from the works of two famous Orientalists, Edward W. Lane and Richard F. Burton. The *OED* does not give details about the suppliers of those quotations, but it is clear that they relied on the same sources; which is common according to Gilliver (2015: 51): “This pattern of a small group of readers producing the lion’s share of the quotations [...] has recurred throughout the Dictionary’s history.” If the loanwords massively integrate English through the translations of a few books, the meaning of words can be biased. One example would be the different translations of the *Arabian Nights* used as sources by the *OED*.

The most famous direct translations of the *Arabian Nights* from Arabic into English were done by the Orientalists Edward W. Lane (1801-1876), Richard F. Burton (1821-1890) and John Payne (1842-1916). Indeed, our research reveals that eleven Arabic loanwords in English were introduced through their different translations and editions of the *Arabian Nights* or *Thousand & One Nights*; eight items were introduced through Edward W. Lane’s works (as a first and / or second citation), while two citations (second and third) were from Richard F. Burton’s works. Only three items were used in third citations from John Payne’s works.

It is worth mentioning that it was Jean-Antoine Galland (1646-1715) who first introduced *Les Mille et une nuits* (known as *The Arabian Nights* in English) to Western countries by translating them directly from Arabic into French. According to Irwin Robert (1994:16), “Galland used a three-or four-volume manuscript, dating from the fourteenth or fifteenth century, as the basis for his translation”. The first volume of *Les Mille et une nuits* was published in 1704, the twelfth and final volume was published in 1717 (BNF Essentiels Littérature). The book was an immediate success, first in Europe, and then in America and in Australia. As a result, some parts of Galland’s work were translated into English soon after their publication. However, little is known about the translator or the exact date of the first English translation (1706? 1708?) (Knipp 1974: 52; Irving 1994:19).

In spite of the well-known great success of the *Arabian Nights*, our detailed study of the citations reveals that the largest number of Arabic loanwords in English has been included into the *OED* database prior to the publication of the *Arabian Nights*. Indeed, 40 items were used in quotations from the works of the same two Orientalists (who also translated the *Arabian Nights*). These two major books are *An Account of the Manners and Customs of the Modern Egyptians* (1836) by Edward W. Lane and *Personal Narrative of a Pilgrimage to El-Medinah and Meccah* (1855) by Richard F. Burton. Lane’s book introduced 30 items, including 19 first citations and 9 second citations, while Burton’s book introduced 10 items, including 4 first citations and 5 second citations. This phenomenon might explain why the influx of Arabic loanwords culminated in the 19<sup>th</sup> century. Furthermore, Burton’s work, which deals with religion, accounts for the massive number of religious Arabic words that entered the English language. The data collectors selected those translations from Arabic into English probably because they knew those words were direct translations from Arabic and the first evidence ever of these items in English.

The consequences of direct translation from Arabic into English are multiple. The most obvious one is the reduction of meaning of some words. The *OED* does not always offer a wide range of meanings for words such as *madrasa* for instance.

### (6) *Madrasa*. (ID 112073)

Origin: Probably of multiple origins. Probably partly a borrowing from Turkish. Probably partly a borrowing from Persian. Probably partly a borrowing from Urdu. Partly a borrowing from Arabic.



Etymons: Turkish *medrese*; Persian *madrassa*; Urdu *madrassa*, *madarsa*; Arabic *madrasa*.

Frequency: (in current use): 4

Definition:

1. In Muslim countries: a school of Islamic theology and law; (also more generally) a school (esp. a secondary school) or institution of higher Islamic education.
2. In other Muslim communities (esp. South Africa, in form *madressa* South African /mə'dresə/): a Muslim school, operating after normal school hours and teaching children subjects such as Islamic history, Islamic belief, and the reading, memorizing, and reciting of the Qur'an.

The two definitions provided by the *OED* give the word a religious dimension. However, *madrassa* does not necessarily mean a "religious school", it could also mean "a school" that provides any kind of non-religious education.

Thawatbeh (2011:111-112) also gives the example of *Jihad* which is defined as "a religious war of Muslims against unbelievers, inculcated as a duty by the Qur'an and traditions" whereas it also has other meanings such as "a struggle against one's self" or "stating the truth forcibly" or "refraining from [doing/saying] bad things." Thawatbeh (2011: 112) argues that this "limited view of *Jihad* seems to be ideologically motivated. The foreign text is imprinted with values specific to the target culture."

It is true that the selection of quotations can be subjective and therefore reflects the opinion of the suppliers rather than the different meanings of the words that can change and evolve according to the context, the period and the society in which they are used. The two examples of *Madrassa* and *Jihad* show that the current geopolitical context emphasizes one specific sense of those words. Gilliver (2015: 71) highlights the importance of quotations and their context when he writes:

I constantly find myself needing quotation evidence (...) recent evidence, for example, which demonstrates that a particular sense of a word (...) is still current, or an earlier example of a particular use of a word that can only be confirmed to be an example of that use by careful reading of the extended context.

Some words show that the *OED* sometimes provides several definitions reflecting the context in which a word is used. For instance, today, *fatwa* has a negative connotation as it is generally used in the sense of "a death sentence delivered by a Muslim authority". Not only does the *OED* give another meaning of the word *fatwa*, but it also explains the reasons behind its negative connotation:

(7) *Fatwa*. (ID 69642)

1. a. Islam. A formal, authoritative ruling on a point of Islamic law; a scholarly opinion given (typically in writing) by a mufti or other Muslim juridical authority in response to a question posed by an individual or a court of law.
- b. irregular. A declaration or decree by a Muslim authority calling for a person to be put to death, typically as a punishment for blasphemy or apostasy; a death sentence.

*Modern use in this sense appears to be influenced by the frequent (though inaccurate) glossing of fatwa as 'death sentence' in media reports of a fatwa issued by Ayatollah Khomeini in 1989, which called for the killing of Salman Rushdie, whose 1988 novel The Satanic Verses was considered by some to be blasphemous and insulting to Islam.*

2. An edict or statement issued by a religious authority belonging to a faith other than Islam. Also occasionally in colloquial or trivial use: a forcefully expressed opinion, judgement, or condemnation; a decree.

Many Arabic loanwords in English are exclusively defined according to religious criteria when broader definitions are attested in Arabic. However, the semantic areas those loanwords cover are restricted and conditioned by translators when integrating the English language. Indeed, especially in the 19<sup>th</sup> century, direct translations from famous Arabic books, whose topics mainly deal with religious matters, necessarily restrict the scope of the meanings of the original Arabic words in English. The topics of translated books are therefore determining parameters in the survey of semantic properties of loanwords. It appears that this translation phenomenon from literary sources can account for the great majority of Arabic loanwords in English.

## 7 Conclusions and Future Research

The complexity of a dictionary-making process is reflected in this case study of Arabic loanwords in English. Determining the origins of words can be challenging for lexicographers because it appears that the etymological tagging of the *OED* can raise questions. The *OED* database provides interesting information related to semantics and frequency. It has also been possible to measure and date the influx of Arabic loanwords in English through the quotation system. The diversity of quotations in terms of periods and sources is also another challenging, yet determining, aspect that must be taken into consideration. Writing dictionaries is a multidimensional process that should be considered through an interdisciplinary lens. This preliminary research on Arabic loanwords in English needs further investigation on several issues. It would be interesting, from a geopolitical perspective, to work on the views English-speaking countries may have on Arabic-speaking ones through the integration of loanwords in institutionalized references as well as on the literary sources the *OED* quotations are extracted from. The last section on the *OED* quotations has only tackled a few points related to translation and semantics, but it would be interesting, especially when dealing with 20<sup>th</sup> century loanwords, to try to connect the integration of Arabic loanwords with geopolitical events that have taken place in the world. In that the massive influx of Arabic loanwords in the 19<sup>th</sup> century can legitimately be assumed to be the result of literary translation; it still has to be determined whether more recent loanwords have been integrated into the English language through the same prism.



## 8 References

- BNF Essentiels Littérature. *Les Mille et Une Nuits*. Galland (1704). Accessed at: <https://gallica.bnf.fr/essentiels/galland/mille-nuits> [25/05/2020].
- Darwish, H. (2015). Arabic Loan Words in English Language. In *IOSR Journal of Humanities and Social Science*, 20(7), pp. 105-109.
- Gilliver, P. (2015). The Quotation Collectors: A Conspectus of Readers for the *Oxford English Dictionary*. In *Dictionaries*, 36, pp. 47-71.
- Irwin, R. (1994). *The Arabian Nights: A Companion*. London: Allen Lane.
- Knipp, C. (1974). The "Arabian Nights" in England: Galland's Translation and Its Successors. In *Journal of Arabic Literature*, 5, pp. 44-54.
- Mossé, F. (1943). On the Chronology of French Loan-Words in English. In *English Studies*, 25, pp. 33-40.
- Salloum, H. & Peters, J. (1996). Arabic Contributions to the English Vocabulary: English Words of Arabic Origin; Etymology and History. Librairie du Liban Publ.
- Sapir, E. (1921). *Language. An Introduction to the Study of Speech*. New York: Harcourt, Brace.
- Serjeantson, M.S. (1935). *A History of Foreign Words in English*. Routledge & Kegan Paul.
- Simpson J. & Weiner, E. (1989). *Oxford English Dictionary. Online Edition*. Oxford: Clarendon Press. [20/04/2020]
- Thawabteh, M. (2011). The Other Side of the Coin of Lexical Borrowing from Arabic into English. In *Transcultural*, 1, pp. 103-122.

## Appendix 1: Corpus of Arabic Loanwords in English (409 items)

aba, abaya, abjad, Adeni, adhan, agal, Ahmadiyya, Alawi, Al Borak, Aldebaran, Al-Hajj, Alhamdulillah, alim, Allah, Allahu Akbar, al Qaeda, Altair, Amal, ameer, Ansar, Ansayri, ardeb, argan, argel, argileh, ariel, arrack, askar, athanor, atlas, azoth, Baath, baba ganoush, bahar / barr(e)<sup>2</sup>, Bedu, bejel, ben, bint, bismillah, burgoo, Caaba, cabeer, cabob, cadi, cafila, camise / camiss, caratch, caroteel / carotel, coffee, coffle, cossid, cowle, dabuh, dahabeeyah / dahabiah, daman, darbuka, deloul, Deneb, dibs, dieb, dirhem / dirham, diss, douar / dowar, doum, Druse / Druze, durra / dhurra, Eid-al-Fitr, El Nath, emir, ezan, faki, fakir, fana, Fatah, Fatiha(h), faufe(l), fedayeen, feddan, falafel, fellagha, fellah, fennec, fils, fingan / finjan, fiqh, fitna, fluce / floose, Fomalhaut, fonduk, frasilah, freekeh, futah, gaiassa, galabiya, gandoura, Garshuni, Gelalaeen, ghaffir, ghawazee, ghazeyeh, Ghazi, gholam, ghoul, Ghuzz, gibli, gimbril, giraffe, girba, gourbi, grab, gufa, gundi, habara, haboob, Hadith, hadj, hadji / hajji, haik / haick, hakeem / hakim, hakim, halal, halala, halawi, halfa, Hamas, hammada, hammal / hummaul, hammam / hummaum, hamza, Hanif / Haneef, haram, haram<sup>3</sup>, harissa, harka, hashish / hasheesh, hawala, henna, hijab, Howeitat, hummum, huzoor, hygeen / hajeen, iggri / iggry, ihram, ijtihaad, ilb, imam, imshi, inshallah, intifada, Iraqi, ithel, izar, jann, jebel, jelab, jerid / jereed, jerm, jihad, jihadi, jinn, jinnee, jol, jubbah, Kababish, Kabyle, kadish, Kaffir, kali, kantar, kat, kazi, kef / keif / kief, keffiyeh, khamsin, khan, kharaj / kharatch, khatib, Khilafat, khimar, khor, khula, khutbah, kibbeh, kiblah, kibrit, kissar, Kiswa, Kitab, kitar, kohl, Koran, Koreish, kourbash / koorbash, kuphar, Kuwaiti, lablab, leban, lebbek, litham, loof, loofah, madhhab, madrasah, mafeesh, maghrib, Maghribi, mahaila, mahalla, Mahdi, mahmal, mahr, Makhzan, malem, malesh, Maliki, Mameluke, Mamur, mancala, mandarah, mandil, maqam, marid, mashallah, mastaba, maulana, maund, mauz, mawla, mawlid, medina, mellah, Mendoub, merissa, Metawileh, Metran, mhor, mihrab, millime, minbar, Miraj, mithqal, mohair, Muallaqat, mudhif, mudirieh, muezzin, muhajir, Muhammad, Muharram, Mukhabarat, mukhtar, mulai, murid, Murji'ah, murshid, Muslim, Muslimah, Mussulmin, Mussulmin, mut'a, mutawwa, nabk, naib, Naqib, Nasara, naskhi, Nasrani, nil, niqab, Nusayri, omda, Othman, oud, pastilla, qasida, Qatari, quaiss kitir, quies, Qur'an, qursh, raad, rafiq, rayah, rebab, reg, resalgar, riad, riba, riqq, robin, roc, rottol, Roumia, Rwala, sabkha, Sahara, Sahrawi, sakia, salaam, Salafi, Salafiyya, salat, saluki, samn, santir / santour, sarsar, Saudi, sebesten, seif, semsem, Senussi, Senussia, serab, serir, seyal, shadda, shadoof, shahada, shahid, Shaitan, shamal, Shammar, Sharia, shawarma, sheikh, Sheikha, Sherarat, Sherari, shereef, shereefa, sherifi, sherwal, Shiah, shisha, shott, shufti / shufti, shura, Siddi, sief, simoom, sim-sim, Subbhanallah, sudd, Sufi, sulham, sultany, Sunna, Sunni, sunt, sura, Swahili, syce / sais, tabbouleh, tabl, taboot, tahalli, tahina, taj, tajine, takaful, Takbir, talak, talha, tarada, tarboosh, tarfa, tariqa, tawaf, tazia, tecbir, tell, terjiman, tezkere / teskere, tibbin, tobe, torba, tuba, tumbak, tumbaki, ulema, umma, umrah, urs, wadi / wady, Wafd, Wahabi / Wahabee / Wahhabi / Wahabee, wakf / waqf, wali, wazir, weli / wely, worral, Yahudi, yashmak, Yemeni, yimkin; Yunani, zaatar, Zaidi, Zar, zarf / zurf, zariba / zareba, zarnich, zawiya, Zendik, zibib, ziczac, zikr, ziraleet

## Appendix 2: Non-Selected Words (102 items)

afion, afrit, Alawite,alconde, Alcoran, alfaqui, Algebar, Alhaji, alif, alkanet, almeh, askari, Bohairic, bulbul, burkundaz / burkundauze, cabaan / caban, cazimi, chermoula, dewan, dghaisa, dinar, dubba / dubber, Eid, Eid-al-Adha, Fatimite, fatwa, fedai, felucca, fistic, gerfaunt, halva, Hanafite, Hanbalite, hardun, harem, Hezbollah, Hobson-Jobson, hookah, Hubshee, iddat, Islam, Kadarite, kalia / kalioun, karrozzin, keiri, kharif, Kurd, lampuki, Lyhianic, liwa, Madan, maid, majlis, mangal, manzil, Marinid, mashrabiyya, masjid, mattamore, mazar, mimation, minaret, mocker, mocked up, Moslemah, mudir, mufti, mujahid, mujahidin, mullah, muqaddam, Mussulman, Mutazilite, nabi, nugger, nunation, oojah, otto, ouguiya, Ramadan, razzia, redif, reis, rial, Rifai, Roumi, rubai, rufiyaa, Saadian, Sabian, sahib, Sahidic, Salafist, sandal, sayyid, Shafiite, sheregrig, spitchered, tabasheer, tamasha, tamboura, vizierate

<sup>2</sup> The items which are separated by slashes correspond to the different spellings.

<sup>3</sup> *Haram* has two distinct entries in the dictionary.



### Appendix 3: Frequency of Arabic Loanwords in English

No data = ameer, Ansar, cadi, dabuh, faufe(l), freekeh, mandil, mauz, Mussulmin, Othman, resalgar, robin, sief, sultany, terjiman, worral, zikr (17 items)

1/8 = cabeer, caroteel / carotel, fana, gaiassa, ghazeeyeh, halawi, iggri / iggry, kuphar, nil, quaiss kitir, rottol, shereefa, tahalli, Takbir, Zarnich (15 items)

2/8 = abjad, agal, Al Borak, Aldebaran, Alhamdulillah, Allahu Akbar, Altair, Amal, Ansayri, argan, argel, argileh, askar, athanor, azoth, baba ganoush, bismillah, Caaba, cabob, cafila, camise / camiss, caratch, cossid, cowle, dahabeeyah / dahabiah, darbuka, deloul, Deneb, dibs, dieb, Eid-al-Fitr, El Nath, ezan, faki, Fatah, fellagha, fingan / finjan, fluce / floose, Fomalhaut, fonduk, frasilah, futah, galabiya, gandoura, Garshuni, Gelalaeen, ghaffir, ghawazee, gholam, gibli, gimbri, girba, gourbi, gufa, gundi, habara, haboob, haik / haick, hakeem / hakim, halala, Hamas, hammada, hammal / hummaul, harka, Howeitat, hummum, huzoor, hygeen / hajeen, ilb, imshi, inshallah, ithel, izar, jann, jelab, jerid / jereed, jerm, jubbah, kadish, keffiyeh, khimar, kibbeh, kiblah, kibrit, kissar, Kiswa, kitar, kourbash / koorbash, leban, litham, loof, mafeesh, mahaila, mahmal, malem, malesh, mancala, mandarah, marid, mashallah, Mendoub, Merissa, Metawileh, Metran, mhorr, Muallaqat, mudhif, mudirieh, Murji'ah, Muslimah, mutawa, nabk, Nasara, naskhi, Nasrani, niqab, omda, qursh, raad, rafiq, riad, riqq, Roumia, sakia, samn, santir / santour, sarsar, sebesten, semsem, Senussia, serab, serir, shadda, shadoof, shamal, shawarma, Sherarat, Sherari, sherifi, sherwal, shisha, shufti / shufti, sim-sim, Subbhanallah, sulham, taboot, takaful, talha, tarada, tarfa, tecbir, tezkere / teskere, tibbin, torba, tumbak, tumbaki, umrah, Yahudi, yimkin, zaatar, zarf / zurf, Zendik, zibib, ziczac, ziraleet (165 items)

3/8 = abaya, Adeni, adhan, Ahmadiyya, alim, ardeb, ariel, arrack, bahar / barr(e), Bedu, bejel, burgoo, coffle, daman, douar / dowar, doum, durra / dhurra, Fatiha(h), falafel, fennec, fitna, Ghuzz, grab, hadj, hadji / hajji, hakim, halal, halfa, hammam / hummaum, hamza, Hanif / Haneef, haram, haram, harissa, hawala, hijab, ihram, jebel, jihadi, jinnee, jol, Kababish, kantar, kazi, kef / keif / kief, khamsin, khan, kharaj / kharatch, khatib, khor, khula, khutbah, kohl, Koreish, lablab, lebbek, loofah, madhhab, madrasah, maghrib, Maghribi, mahalla, mahr, Makhzan, Mamur, maqam, mastaba, maund, mawla, mawlid, medina, mellah, millime, minbar, Miraj, mithqal, muhajir, Muhammad, Mukhabarat, mukhtar, mulai, murid, murshid, Muslimin, mut'a, naib, Naqib, Nusayri, oud, pastilla, qasida, Qatari, quies, rayah, rebab, riba, Rwala, sabkha, Sahrawi, salaam, Salafi, Salafiyya, salat, saluki, seif, Senussi, seyal, shahada, shahid, Shaitan, Shammar, Sheikha, shereef, Shiah, shott, shura, Siddi, simoom, sudd, sura, tabbouleh, tabl, tahina, taj, tajine, talak, tarboosh, tariqa, tawaf, tazia, tell, tuba, urs, weli / wely, yashmak, Yunani, zariba / zareba, zawiya (138 items)

4/8 = aba, Alawi, Al-Hajj, Allah, al Qaeda, atlas, Baath, ben, bint, dirhem / dirham, Druse / Druze, emir, fakir, fedayeen, feddan, fellah, fils, fiqh, Ghazi, ghoul, giraffe, Hadith, hashish / hasheesh, henna, ijtihaad, imam, intifada, jihad, jinn, Kabyle, Kaffir, kali, kat, Khilafat, Kitab, Kuwaiti, Mahdi, Maliki, Mameluke, maulana, mihrab, mohair, muezzin, Muharram, reg, roc, Sahara, Sharia, sheikh, Sunna, syce / sais, tobe, ulema, umma, wadi / wady, Wafd, Wahabi / Wahabee / Wahhabi / Wahabee, wakf / waqf, wali, wazir, Yemeni, Zaidi, Zar (63 items)

5/8 = diss, Iraqi, Koran, Qur'an, Saudi, Sufi, Sunni, sunt, Swahili (9 items)

6/8 = coffee, Muslim (2 items)







# Loanblends in the speech of Greek heritage speakers: a corpus-based lexicological approach

Gavriilidou Z.<sup>1</sup>, Mitits L.<sup>2</sup>

<sup>1</sup> Democritus University of Thrace, Greece

<sup>2</sup> Democritus University of Thrace, Greece

## Abstract

Found in situations of language contact between Greek and English, Greek heritage speakers living in the US, Canada, Australia, etc. produce loanblends, which combine an English stem e.g. *fence* and a Greek affix e.g. *-i*, as in *fénsi* ‘fence’. These loanblends are very frequent contact-induced formations that have become part of the Heritage Speakers’ everyday language usage. This study analyses fifty (50) such loanblends found in the Greek Heritage Language Corpus, which contains data from Greek Heritage Speakers living in Chicago, US, tests the borrowability scale constraint and the unmarked gender hypothesis for loanwords, and discusses the lexicographic protocol for the compilation of an online dictionary of loanblends of Greek Heritage Speakers.

**Keywords:** loanwords, loanblends, Greek Heritage Language, borrowability scale, gender assignment, unmarked gender

## 1 Introduction

Heritage language speakers are individuals “who have been exposed to a particular language in childhood but did not learn it to full capacity because another language became dominant.” (Polinsky & Kagan 2007). The term Heritage Language, on the other hand, is used for languages of diasporic communities, especially ones with a history of migration, which are spoken by simultaneous or sequential early bilinguals, the heritage speakers (HSs), who are typically the children of immigrants and are usually bilingual in the dominant language of the host country and the heritage language to varied degrees. HSs grow up acquiring the language of their parents’ country of origin at home until they start school, at which time they begin to acquire the language of the host country. Gradually, they become dominant and more fluent in the majority language, limiting the use of the heritage language to the interaction with family and friends from the same ethnolinguistic background (Benmamoun et al. 2013; Karatsareas 2018). The incomplete acquisition of the heritage language, possible subsequent attrition, and interference from the majority language gradually lead to the formation of new heritage grammars and vocabularies characterized by innovations (Karatsareas 2018). This phenomenon is reinforced by code-switching (CS), the phenomenon of alternating between two or more languages in conversations, in a clause, a discourse segment, or on the word-internal level (intra-word CS) (Mager et al. 2019).

This paper reports the results of the project entitled *Varieties of Greek as Heritage Language* (HEGREEK, MIS 5006199) which aimed at profiling Greek heritage speakers (GHSs) living in the US and Russia as well as at collecting data for the compilation of the open-access online Greek Heritage Language Corpus (GHLC). It focuses on loanblends used by GHSs from the US, extracted from the GHLC, and offers quantitative data about gender assignment, grammatical category frequency and adaptation strategies used. It finally elaborates on the principles of a lexicographic protocol for the compilation of an online dictionary of loanblends which could include data of various pairs of languages in contact.

The paper starts with the literature review focusing on loanblends found in the speech of Greek heritage communities, the borrowability scale constraint, morphological adaptation in borrowings, gender assignment, the unmarked gender hypothesis and the classification of loanwords in semantic fields. It then describes the methods of the study: the data about the sample, the methodology adopted, the principles of data analysis, the results yielded, and the discussion of the main findings. The next part offers the lexicographic protocol for the compilation of an online dictionary for the loanblends of GHSs. Finally, the conclusion summarizes the main findings, provides cues for further investigation and addresses the limitations of the study.

## 2 Loanblends used by GHSs

Loanblends are borrowings that combine bound morphemes from two languages as in *fénsi* ‘fence’, where there is a combination of the English stem *fence* and the Greek inflectional affix *-i*, in *matrmátzi* ‘mattress’ which combines the German stem *Matratze* and the Greek inflectional affix *-i* or in *runeando* ‘running’ which combines the English stem *run* and a Spanish affix *-eando*. Sometimes the basis can be a collocation as in *biloziri* (below zero+*i*) ‘below zero’ or a part of a compound as in *ófi* ‘day-off’. Considering Corbin’s (1987) tripartite categorization of words, we claim that loanblends



are [-constructed, + structured] formations since no construction rule can be applied synchronically in Greek.

Following Haugen's (1950) typology on borrowings, loanblends are types of borrowings in which only part of the phonemic shape of the word has been imported, while a native portion has been substituted for the rest. Actually, contrary to loanwords which show only morphemic importation, loanblends show morphemic substitution as well as importation. In literature, these formations are treated either as types of code-switching (Gardner-Chloros 2009), cases of loanwords (Karatsareas 2019, Alvanoudi 2019) or word-internal language mixing governed by the Free Morpheme constraint which predicts that a switch may not occur between a bound morpheme and a lexical form unless the latter has been phonologically integrated into the language of the bound morpheme (Poplack 1980, Alexiadou & Lohndal 2018). Alvanoudi (2019) uses the term *derivational blends* to refer to such constructions, even though they are rarely constructed through derivation.

Seen from a functional perspective, loanblends like the ones studied in this paper are used to fill vocabulary gaps of heritage speakers who find it easier to use stems from the majority language, in which they are generally more proficient, and affixes from their HL, when they produce speech in the heritage language. In situations of contact between English and Greek, since English does not mark grammatical gender, an obligatory feature in Greek, the above combination becomes an efficient vocabulary compensation strategy for overcoming lexical gaps in Greek by assigning grammatical gender to English words through the addition of a Greek affix. The loanblend and the equivalent native word with the same meaning form couples of words (e.g. *bóksi-koutí* 'box', *tséci-tsek* 'check', *káro-aftocínito* 'car', *blóci-ikodómikó tetrágono* 'block', *bascéta-kaláthi* 'basket') that co-exist with a different distribution in communication, since native speakers never use loanblends, while heritage speakers mainly use them but may also more rarely use native words.

Greek or Cypriot-Greek loanblends have been previously studied from a sociolinguistic (Gardner-Chloros 2009, Alvanoudi 2019, Karatsareas 2019) or a morphosyntactic perspective (Alexiadou 2011, 2017, Matejka-Hanser 2011). Gardner-Chloros (2009: 49), investigating Greek Cypriots in London, considers such formations as English words, mainly nouns, that were adopted and morphologically/phonologically adapted to the Greek Cypriot Dialect, either for referring to new concepts connected to the British culture (e.g. *φισιάτικο* 'fish and chips shop'), or for replacing native words for the sake of facility, as happens in the case of the word *marcéta* 'market'. Alvanoudi (2019) analyses 31 derivational blends (as she calls these formations) used by immigrants in Cairns, Queensland (Australia) and maintains that they are, phonologically and morphologically integrated into Greek, core borrowings given that they duplicate elements that Greek already possesses. She also argues that "such loanwords are perceived by friends and relatives in Greece as indexes of otherness, that is their Greek Australian identity" (Alvanoudi 2019:42). Karatsareas (2019: 154), on the other hand, studying Cypriot Greek as a heritage and community language in London, claims that "this type of lexical borrowing is labelled Grenglish and is associated, especially among second- and third-generation speakers, with low socioeconomic status and low level of education". Alexiadou (2017) discusses examples found in Fotopoulou (2004) and Gardner-Chloros (2009) and claims that "the borrowed nouns have become active members of the speakers' vocabulary, because they are assigned one of the Greek declension classes, as determined by the overall sentence context" (Alexiadou 2011:46) and that in cases where a combination of a root from one language with a functional morphology from another is not allowed this happens because "the language mode of the speaker suggests that the functional morphology should come from the language with overt default realization or because morpho-phonological reasons rule out the particular mixing in question" (Alexiadou 2017:13). Finally, Matejka-Hanser (2011: 88) studies 12 loanblends from Greek spoken by Greek-Americans of Chicago and observes that "in most cases the Standard Greek variants are morphologically more complicated and phonologically more difficult (for non-natives) than the loanwords. This fact might point towards language economy as motivation for the borrowing." No previous studies have investigated so far loanblends from a lexicological-lexicographic point of view.

### 3 The Borrowability scale constraint

Previous literature laid special emphasis on the investigation of linguistic properties that facilitate or even promote borrowing (Matras 1998, Matras & Sakel 2007, Haspelmath 2008, Matras 2011). The authors concluded that there are borrowability scales or hierarchies that can be interpreted in four different ways:

- (i) **Temporal:** A language borrows elements on the left before it borrows elements further to the right.
- (ii) **Implicational:** A language that contains borrowed elements on the right also contains borrowed elements further to the left.
- (iii) **Quantitative:** A language borrows more elements belonging to the types on the left than elements belonging to the types further to the right.
- (iv) **Probabilistic:** Elements belonging to the types on the left are more likely to be borrowed than elements further to the right. (Haspelmath 2008: 6)

The borrowability scale is one of the most important types of constraints for borrowing, predicting which morpheme type or part of speech is borrowed more easily. In particular, crosslinguistic data show that lexical items are more easily borrowed than grammatical items, unbound morphemes are more easily borrowed than bound morphemes, content words are more easily borrowed than function words, nouns are more easily borrowed than verbs or adjectives (Thomason & Kaufman 1988, Van Hout & Muysken 1994, Field 2002, Myers-Scotton 2002). There are no previous studies focusing on



loanwords in Greek which investigate whether the borrowability scale constraint is valid in the case of Greek language data.

#### 4 Morphological adaptation of borrowings

When words are borrowed from other languages, these words are phonologically and morphologically adapted according to the sound and morphology of the recipient language. Moreover, when a borrowing enters a certain word class in a recipient language, it should acquire all features of that word class (or the features of respective subclasses, if they are distinguished). This means that, for example, in the case of Greek, a new member of the nominal category should be able to express case, number and gender or a new member of the verbal category should mark person, number, tense, modality, aspect, etc. Languages use different adaptation strategies to assign loanwords to specific word classes and conform them with the morphological system of the recipient language. However, the degree of adaptation may vary, depending on the time of the introduction of the borrowing into the receiving language, the possible multilingualism of recipient language speakers or their stance towards the donor language (Haspelmath 2009).

The strategies adopted for the morphological adaptation of loanwords are complex, language-dependent and include, in general terms, the following (Haspelmath 2009, Matras 2009, Pakerys 2016):

- (i) **Zero morphological adaptation:** In some cases, borrowings are not adapted in the recipient language, resulting in indeclinable words in cases of inflected languages like Greek, e.g. *tsek* ‘check’, *snítsel* ‘schnitzel’, *reportáz* ‘reportage’, *kolxóz* ‘kolkhoz’.
- (ii) **Addition of inflectional affixes / assignment to an inflection class:** e.g. *gázi* ‘gas’.
- (iii) **Addition of derivational suffixes or class markers:** e.g. *provokáro* ‘provoke’, *flertáro* ‘flirt’, *buniá* ‘bunch’.
- (iv) **Truncation of a derivational suffix:** e.g. *tenístas* (\**tenisístas*) ‘tennis player’.

Anastassiadis (1994) investigated morphological adaptation in Greek loanwords and classified borrowings in two classes: +adapted e.g. *imresionismós* ‘imressionism’, and -adapted: e.g. *traktér* ‘tractor’. No previous studies have investigated so far in detail morphological adaptation in Greek loanblends.

#### 5 Gender assignment in borrowings

Gender is an inherent feature of the nominal category and it can be predicted from semantic information stored in the lexical entry or from morphophonological characteristics (Anastassiadis & Mitsiaki 2012). Gender assignment, on the other hand, is one of the most common procedures for word morphological adaptation. According to Haspelmath (2009: 42), “languages with gender and inflection classes need to assign each word to a gender and inflection class, so that it can occur in syntactic patterns which require gender agreement or certain inflected forms”. To achieve that, each language develops systematic mechanisms for gender assignment that can be tested and verified by studying the frequency of prototypical cases, gender assignment in loanwords, neologisms or pseudowords and data from language development.

Greek has a three-gender system, classifying nouns in masculine, feminine and neuter according to the word ending vowel (inflectional ending), which “reflects a fusion of the grammatical categories of case (nominative, genitive, accusative, vocative) and number (singular, plural)” (Anastassiadis & Mitsiaki 2012: 190). Based on frequency, developmental or semantic criteria, Anastassiadis (1994), Kavoukopoulos (1996) and Anastassiadis & Chila (2003) maintain that neuter is the default gender in Greek. This claim is further supported by empirical data provided by Tsimpli (2011) and Tsimpli & Hulk (2013) which showed: a) that neuter is used during language acquisition and is also the learner default gender and b) that neuter is the default gender “on the grounds of syntactic distribution in contexts where gender agreement is inert” (Tsimpli & Hulk 2013: 138).

Anastassiadis & Chila (2003) consider the semantic feature of animacy (-/+animate) and the morphological criterion of ending vowels as the defining factors for prototypicality in gender assignment. The authors consider as prototypically masculine nouns all masculine animate nouns ending in -s e.g. *patéras* ‘father’, and non-prototypical the non-declinable masculine animate nouns e.g. *komándo* ‘commando’ and inanimate nouns e.g. *kompjúter* ‘computer’ or -animate nouns ending in -s e.g. *uranós* ‘sky’. Prototypically feminine nouns are feminine animate nouns (e.g. *jajá* ‘grandmother’, *nífi* ‘bride’, *nixú* ‘manicurist’) or feminine inanimate nouns ending in -a, i and u (e.g. *enérjia* ‘energy’, *alají* ‘change’) and non-prototypical all feminine nouns referring to masculine entities, e.g. *frurá* ‘guard’ or those that are indeclinable e.g. *béibisítér* ‘baby sitter’. Finally, prototypically neuter are all inanimate nouns, all neuter nouns ending in -o, -i and -a and all indeclinable nouns. Non prototypical neuters are inanimate neuters ending in -n or -s (e.g. *méros* ‘place’, *mélon* ‘future’) or animate indeclinable neuters e.g. *garsón* ‘waiter’. Table 1 (taken from Anastassiadis & Chila (2003: 34)) presents the prototypical characteristics for each gender:



Grammatical gender	Masculine	Feminine	Neuter
Natural gender	male	female	Ø or male/female
Ending vowel	-s	-a	-o
		-i	-i
		-u	-a
			Non declinable

Table 1: Prototypical Standard Modern Greek gender system

Poplack et al (1982:11) elaborate on the factors responsible for gender assignment in borrowed nouns. They claim that these factors include:

- (i) The physiological sex of (animate) referent, in other words the natural gender divided in masculine and feminine;
- (ii) The phonological gender, depending on the qualities of word endings (e.g. in Greek, nouns ending in -a are prototypically feminine; for a detailed account of gender prototypicality in Greek see Anastassiadis-Symeonidis & Markopoulou-Chila (2003));
- (iii) The analogical gender, which relates the gender assigned to the borrowing with the gender of a semantically equivalent word or a hyperonym in the recipient language (e.g. *járða*<sub>[fem]</sub> / *avlí*<sub>[fem]</sub> ‘yard’, *argó*<sub>[fem]</sub> / *the language*<sub>[fem]</sub> *argó* ‘slang’);
- (iv) Homophony, in other words the gender assigned to words having a homophone suffix (e.g. *gazolíni* ‘gasoline’, *grosaría* ‘grocery’);
- (v) Suffixal analogy.

Anastassiadis (1994: 94) on her part, builds on Poplack et al. (1982) and proposes five general rules for gender assignment to borrowings, considering two basic criteria, [-/+ animate] and [-/+ adapted]:

- 1<sup>st</sup> rule:** A [+animate] noun in the donor language will be included in the equivalent gender in the recipient language irrespectively of its degree of adaptation, e.g. *metr*<sub>[masc]</sub> ‘master chef’, *mazoréta*<sub>[fem]</sub> ‘cheerleader’;
- 2<sup>nd</sup> rule:** A [-animate] [-/+adapted] noun will be assigned in recipient language the gender that this element has in donor language under certain conditions (marked gender in L1 and L2, sociolinguistic parameters, etc.), e.g. *kuáf*<sub>[fem]</sub> (Fr. la coiffe) ‘coiffe’, *agráfa*<sub>[fem]</sub> (Fr. L’agraffe) ‘buckle’ (interlinguistic analogy);
- 3<sup>rd</sup> rule:** If the conditions of the 2nd rule are not fulfilled, the [-animate] [-/+adapted] noun will be assigned in neuter gender, e.g. *test*<sub>[neut]</sub> ‘test’, *tsekáp*<sub>[neut]</sub> ‘checkup’, *Ji*<sub>[neut]</sub> ‘mistletoe’;
- 4<sup>th</sup> rule:** The [-animate] [+adapted] nouns comply with the gender of the items of the inflectional class in which they are included, e.g. *bufés*<sub>[masc]</sub> ‘buffet’ (morphological analogy);
- 5<sup>th</sup> rule:** The [-animate] [-adapted] nouns can be assigned the gender of a quasi-synonym or hyperonym in the recipient language.

Gender instability in inanimate adapted loanwords (e.g. *kolié*<sub>[neut]</sub>/*koliés*<sub>[masc]</sub>, *sinemá*<sub>[neut]</sub>/*sinemás*<sub>[masc]</sub>, *stiló*<sub>[neut]</sub>/*stilós*<sub>[masc]</sub>) is the result of a two-stage morphological adaptation: at the first stage, loanwords enter Greek as neuter indeclinable nouns, while at the second they are fully adapted to the morphological system (Anastassiadis 1994).

According to Poplack et al. (1982), the unmarked or default gender is attributed to borrowings. With the exception of Anastassiadis (1994), no previous research has investigated gender assignment mechanisms or the unmarked gender hypothesis in Greek borrowings in general or in Greek loanblends.

## 6 Classification of loanblends in Semantic fields

Tadmor (2009) maintains that the semantic field to which a word belongs affects the probability for that word to be borrowed. In other words, certain semantic fields are better candidates for borrowing than others. For instance, semantic fields like ‘Religion and belief’, ‘Social and political relations’, ‘Clothing’ or ‘The house’ correspond to domains which have been affected by intercultural influences (Tadmor 2009: 64). These fields are more prone to borrowing. On the other hand, semantic fields like ‘Sense perception’ or ‘Spatial relations’ are least amenable to borrowing since practically every language is expected to have indigenous words for such concepts.

In order to compile a comparable sample with crosslinguistic data on lexical borrowing, Haspelmath & Tadmor (2009) in their study *Loanwords in the languages around the world* compiled a fixed list of 1460 lexical meanings assigned in the following 24 semantic fields: ‘The Physical world’, ‘Kinship’, ‘Animals’, ‘The body’, ‘Food and drink’, ‘Clothing and grooming’, ‘The house’, ‘Agriculture and vegetation’, ‘Basic actions and technology’, ‘Motion’, ‘Possession’, ‘Spatial relations’, ‘Quantity’, ‘Time’, ‘Sense perception’, ‘Emotions and values’, ‘Cognition’, ‘Speech and language’, ‘Social and political relations’, ‘Warfare and hunting’, ‘Law’, ‘Religion and belief’, ‘Modern world’, ‘Miscellaneous function’.



words’.

Gavriilidou (2018) investigated the distribution of Russian borrowings in Greek into the above-mentioned semantic fields. With the exception of that study, no other research up to date has been conducted on how borrowings, in general, or loanblends, in particular, are classified into different semantic fields.

## 7 Aims and hypotheses

Taking into consideration the gaps in previous literature, as shown in the literature review, in this paper we investigate:

- i) whether the borrowability scale (hierarchy) constraint is supported by our data. In line with the literature on borrowability scales and hierarchies (Matras 2007, Haspelmath 2008), we expect that nominal loanblends from our corpus will exceed in numbers the verbal ones;
- ii) how strategies of morphological adaptation of loanblends are attested in our sample. Based on the literature on morphological adaptation of borrowings, we investigate whether the four different adaptation strategies proposed in Pakerys (2016) concern loanblends. Given that loanblends are combinations of an English stem with a Greek affix, we expect to find three different types in our sample: a) a combination of an English stem and an inflectional affix, b) a combination of an English stem and a derivational affix, c) a combination of an English stem and a class marker;
- iii) how our data are distributed in grammatical genders and whether the unmarked gender hypothesis for borrowings is validated by the data. Taking into consideration Anastassiadis (1994), we expect to find more neuter nouns. We also expect that our data belong to prototypical inflectional classes of each gender.
- iv) which gender assignment factors operate with loanblends found in our corpus;
- v) how data are distributed in the fixed list of semantic fields of Haspelmath & Tadmor (2009).

## 8 Methods

### 8.1 Data

Fifty (50) loanblends were extracted from the Greek Heritage Language Corpus (GHLC) and more precisely from the Chicago sub-corpus. GHLC is a speech corpus developed at Democritus University of Thrace, Greece within the frame of the project *Varieties of Greek as Heritage Language* (HEGREEK, MIS 5006199) and is available at <http://synmorphose.gr/index.php/el/projects-gr/ghlv-gr/corpus-gr>. It is one of the very few corpora containing heritage language data.

It consists of 1st, 2nd, and 3rd generation Greek Heritage Language Speakers’ oral productions, elicited from the interviews of 37 GHLSs from Russia (Moscow and Saint Petersburg) with Russian as their dominant language, and 32 GHLSs from the US (Chicago) with L1 English. In particular, the GHLC includes approximately 130,000 words (20,000 from Moscow, 25,000 from Saint Petersburg, and 85,000 from Chicago) and approximately 90 hours of recordings (30h from Moscow, 30h from Saint Petersburg, and 30h from Chicago). It contains: (a) audio recordings, (b) transcriptions of the recordings with metadata.

Considering issues raised in previous literature (Matras 2007) about the comparison of frequency-based hierarchies drawn from conversational data like the ones in GHLC, we chose not to study loanblends in terms of token or type frequency but in absolute numbers. This is the reason our study is based on the above mentioned 50 loanblends extracted from the corpus.

Data were extracted from the corpus and ordered according to linguistic information, specifically gender (masculine, feminine, neuter), grammatical category (noun vs. verb), declination code (we used the codes adopted in the Dictionary of Standard Modern Greek), mode of construction (stem+addition of inflectional affix vs. stem+addition of class marker), ending vowel, semantic information, gender assignment procedure (see 5 above), meaning in Greek and English.

### 8.2 Results and discussion

Data analysis provided answers to the working hypotheses set in 6, based on a detailed literature review. In this section of the paper, we present our findings and discuss them with respect to the results found in previous studies.

#### *The Borrowability scale constraint*

Out of 50 loanblends, only three (3) were verbs (1,5%) and the rest forty-seven 47 (98,5%) were nouns. This finding confirms our hypothesis that nominal loanblends of our corpus would exceed in number the verbal ones and offers a strong argument about the borrowability scale constraint (Matras 2007, Haspelmath 2008), which predicts that nouns are borrowed before verbs (temporal interpretation), are more likely to be borrowed than verbs (probabilistic interpretation), are more frequently borrowed than verbs (quantitative interpretation), and that their borrowing is a precondition for the borrowing of verbs (implicational interpretation) Haspelmath (2008). In our data no adjectival loanblends have been attested.

According to (Van Hout and Muysken 1994: 42), nouns exceed verbs in number because of the referential role they play



in comparison with verbs or adjectives and given that “one of the primary motivations for lexical borrowing is to extend the referential potential of a language”.

### *Strategies of morphological adaptation*

According to this criterion, two strategies of morphological adaptation and consequently two types of loanblends were attested in our sample:

- i) those constructed by the addition of an inflectional affix to a borrowed lexical root as in *tráci* ‘truck’, *bóksi* ‘box’, *karpéta* ‘carpet’, *rífti* ‘roof’, *sáina* ‘sign’, *járða* ‘yard’ (Class 1), and
- ii) those created by the addition to a borrowed root of a derivational suffix-like ending as in *farmaðóros* ‘farmer’, *grosaría* ‘grocery’, *musikános* ‘musician’, *ruffjános* ‘roof-maker’, *mováro* ‘move’, *frizjázó* ‘freeze’ (Class 2). These nouns have a complex structure without a compositional meaning, in the sense that only the borrowed lexical root contributes to meaning formation, while the suffix-like ending is a pseudo-suffix without any semantic instruction, functioning exclusively as a class marker or paradigmatic integrator (Corbin 1987, 1991). It is important to note here that class markers copy the form and the intrinsic properties of a derivational suffix and their selection is not arbitrary.

From the 50 items of our corpus, 38 (76%) belonged to class 1 and only 12 (24%) to class 2, suggesting that morphological adaptation of loanblends demonstrates a strong preference for the inflectional and not the derivational procedure. However, all verbal loanblends belonged to class 2 and were constructed with the class marker was *-áro* (e.g. *mováro* ‘move’). Standard Modern Greek has only two verb inflectional affixes: *-o* [o] (basic class) and *-ó* [o] (reduced class). Borrowed verbs enter Greek morphological system exclusively with the addition of a class marker (mainly *-áro* and marginally *-iázo*). As put by Anastasiadis & Masoura (2012), class marking regulates both diachronically and synchronically the Modern Greek verbal system.

As expected, no cases of zero morphological adaptation or truncation of a derivational suffix were attested in our sample. However, no cases of addition of a derivational suffix were found either, contrary to what was initially predicted. This probably happens because the role of the suffix used in loanblends is to assign the loanwords into a grammatical category and referential class or, in other words, to permit the borrowing to conform morphologically to a certain word-class and function as a member of a certain lexico-morphological subgroup and not to convey a specific semantic information. On the other hand, derivational suffixes are semantically transparent for gender since they provide semantic information (combined with formal indications) which is not needed in the case of loanblends. This finding needs to be validated with more data.

### *The Unmarked gender hypothesis for borrowings*

Fifty-two percent (52%) of our data are neuter, 26% feminine and 16% masculine. Taking into consideration the fact that the unmarked or default gender of a language is attributed to borrowings (Poplack et al 1982), the high frequency of neuter in loanblends provides evidence for supporting that neuter is the default/unmarked gender in Greek.

Furthermore, the study of our sample showed that gender assignment in loanblends of Greek heritage speakers seems to operate according to whether the loanblend is animate or inanimate. In case of animate referents, the natural gender (masculine vs. feminine) is marked, e.g. *bósis* ‘boss’ vs. *bosína* ‘female boss’, *musikános* ‘musician’ vs. *musikána* ‘female musician’. In case of inanimate nouns, gender is attributed to the loanblends:

- i) in analogy with the gender of a host language semantic equivalent (analogical gender), e.g. *kéci* ‘cake’/kéik<sub>[neut]</sub> ‘cake’, *kontráto/simvóleo*<sub>[neut]</sub> ‘contract’, *káro/aftocínito*<sub>[neut]</sub> ‘car’,
- ii) in phonological analogy, e.g. *bíli* ‘bill’, *sáina* ‘sign’,
- iii) in suffixal analogy, e.g. *gasolini* ‘gasoline’, *grosaría* ‘grocery’, *markéta* ‘market’ where the suffixes *-ini*, *-aría*, *-éta* are feminine in Greek,
- iv) in combination of (i) and (iii) e.g. *karpéto* ‘carpet’, *marcéta* ‘market’. In this last word both the suffix *-éta* and the gender of the host language semantic equivalent *ayorá* ‘market’ condition the feminine gender.

The frequency of cases of each gender-assigning factor attested in our sample is presented in Table 2 below.



Type of adaptation	Frequency of cases %
Natural gender	21,5
Analogical gender	40
Phonological analogy	21,5
Suffixal analogy	8,5
Combination	8,5
<b>TOTAL</b>	<b>100</b>

Table 2: Factors of gender assignment in Greek loanblends

From Table 2 it becomes obvious that the most frequent factor for gender assignment in Greek loanblends is the analogical gender of a semantic equivalent in the recipient language. This finding is in line with Poplack et al. (1982:24) who found that the analogical gender has a “large and pervasive effect” in gender assignment in Puerto-Rican Spanish and French and has to be verified with psycholinguistic experiments investigating data from other borrowing categories as well.

The prototypicality-based analysis of our data revealed that, without exceptions, all loanblends fell within the prototypical Standard Modern Greek gender system as described in Anastassiadis & Chila (2003) (see table 1). More specifically, all masculine nouns were animate in -s (*bósis* ‘boss’, *séfis* ‘chef’, *farmaðóros* ‘farmer’), and all inanimate nouns were neuter ending in -i, e.g. *fláti* ‘flat’, *tikéto* ‘ticket’, *xadóci* ‘hot dog’ (18 cases) or in -o, e.g. *káro* ‘car’, *karpéto* ‘carpet’ (5 cases). As far as feminine loanblends are concerned, feminine inanimate nouns ended in -a, e.g. *fríza* ‘freezer’, *stófa* ‘stove’, *basíkla* ‘bicycle’ (10 cases) and only in one case in -eta, (H,η) (*γκαζολίνη* ‘gasoline’), while there were two cases of feminine animate blends (*musikána* ‘female musician’, *bosína* ‘female boss’). This finding provides strong support to Anastassiadis & Chila’s (2003) prototypicality principle in gender assignment and their model of masculine, feminine and neuter prototypical gender specification in Greek and suggests that the prototypicality principle operated also in borrowing and specifically in loanblends.

Finally, no cases of gender instability between neuter, on the one hand, and masculine and feminine, on the other, were found in our sample, indicating that neuter loanblends do not undergo morphological pressure towards masculine or feminine gender and consequently there is no gender change in progress in this restricted subset of vocabulary. This finding is a supplementary argument for claiming that neuter is prototypically the default gender in borrowings. In other words, the prototypicality and unmarkedness of neuter in loanblends ensures gender stability.

#### *Distribution of loanblends in semantic fields*

Given that loanblends found in our sample are commonly used in everyday communication between GHSs, we deemed necessary to investigate the most frequent semantic fields in which the sample is classified in order to test whether the borrowing rate by semantic field hypothesis by Tadmor (2009) can be validated by our data and check whether Greek loanblends used by GHSs fall within the semantic fields more affected by borrowing. To do so, we adopted the Haspelmath & Tadmor (2009) 24-item semantic fields classification scheme and classified our sample semantically in the following of the 24 categories. The results are presented in Table 3.



Semantic Field	Borrowing rate in the present sample %	Borrowing rate in World Loanword Database % (Tadmor 2009: 63)
Modern World	61	42,5
The house	22	37,2
Food and drink	15	29,3
Agriculture and vegetation	2	30
TOTAL	100	-

Table 3: Loanblend borrowing by semantic field

As shown in Table 3, the distribution of loanblends over semantic fields is analogical with data found in World Loanword Database to a high degree. Furthermore, the semantic fields of “Modern World”, “The house”, “Food and Drink”, and “Agriculture and Vegetation” to which our sample belongs are included in the list of the most affected by borrowing semantic fields in the Loanword Typology project. More specifically, a comparison between the hierarchy based on the contribution of each semantic field to the total number of loanblends in our corpus with the hierarchy of semantic fields found by Tadmor (2009) showed that the two hierarchies correlate.

These data verify our initial observation that loanblends used by GHSs refer to everyday objects, places or food and this is strongly supported by the sociolinguistic instances in which loanblends are used. Furthermore, they provide supplementary support for the claim that, cross-linguistically, certain semantic fields are more likely to be borrowed.

## 9 The mini-dictionary of loanblends used by GHSs

The detailed lexicological analysis of loanblends held so far provided data for the compilation of a mini-dictionary for loanblends. The mini-dictionary of loanblends used by GHSs is a multilingual online dictionary, addressed both to:

- a) the Greek-speaking community, whether it be heritage speakers around the world or native speakers, and
- b) the academia who wishes to study loanblends used by GHS as innovations in the vocabulary of heritage speakers.

It complements the Greek Heritage Language Corpus (GHLC) and is available at <http://synmorphose.gr/index.php/el/#>. The metalanguage of the dictionary is Greek. For the moment, the dictionary macrostructure includes data from bilingual English-Greek heritage speakers extracted from the Chicago-sub corpus of GHLC, but this initial wordlist will be complemented with the inclusion of more data: a) extracted from Russian-Greek heritage speakers’ oral productions included in the Russian sub-corpus of GHLC, or b) from manually collecting all examples presented in previous research investigating such formations (see relevant literature in 2).

The components included in each entry are the following:

1. **Headword**, in the form of nominative singular, in the case of nouns, and in 1st person singular in the present tense of indicative mode, in the case of verbs.
2. **Pronunciation**, both in I.P.A. transcription and as a wag file to facilitate access to blind people.
3. **Grammatical information**, and more specifically the grammatical category (noun or verb), the gender (masculine, feminine and neuter), the inflectional paradigm in which the lemma is classified according to Inflectional Category codes used in the Dictionary of Standard Modern Greek and gender assignment procedure (analogical gender, phonological gender, suffixal analogy, combination).
4. **Etymological information**: the etymological component includes information about the construction procedure of each loanblend.
5. **Semantic fields**: each loanblend is classified according to the classification scheme of Haspelmath & Tadmor (2009) (see 6).
6. **Meaning**: Each entry provides the equivalent word in the recipient language (in our case Greek) and the loanblend translation in the donor language (English, German, Russian, etc.)

The interface is user-friendly, with the alphabetical list displayed above the entry list, with the added option to sort entries according to different criteria. The search function offers the possibility to search by headword, definition, keyword or synonyms with additional search modes: “Begins with”, “Contains”, “Exact term”, “Sounds like”. The functionality of the mini dictionary of loanblends will be further developed so as to link each entry with the exact point in GHLC where the loanblend-lemma is found. Other future plans include the enrichment of the mini-dictionary macrostructure with loanblends used by heritage speakers of other languages, in order to transform it into a useful, for heritage languages research, database with cross-linguistic data.



## 10 Concluding remarks

Given that, from a lexicological point of view, the nature of language contact is a complex phenomenon, this study offers some insights into the complexity of borrowing attested in the speech of Greek HSs with English as their dominant language. The investigation of 50 loanblends, mainly nouns, created and used by Greek HSs from the Greek Community of Chicago:

- (i) provided arguments about the borrowability scale constraint,
- (ii) highlighted two main modes of construction of loanblends, one more frequent operating with the addition of an inflection affix and a marginal one operating with the addition of a class marker to an English stem,
- (iii) offered strong support to the claim that neuter is the default gender in Greek,
- (iv) showed that analogical gender is the most frequent strategy employed for gender assignment in loanblends used by Greek heritage speakers
- (v) provided cues for supporting that, cross-linguistically, certain semantic fields are more likely to be borrowed.

From a lexicographic perspective, making dictionaries like the one described in this paper goes beyond pure lexicographical work. It is an attempt of preservation and documentation of Greek as heritage language and of the culture of heritage speakers.

Finally, this area of study is of increasing interest, since collecting cross-linguistic data about contact-induced borrowing in cases of heritage speakers is important for understanding universals in HSs' neological lexical creations and vocabulary acquisition and use.

## 11 References

- Alexiadou, A. (2011). Remarks on the morpho-syntax of code switching. In *Proceedings of the 9th International Conference on Greek Linguistics*, University of Chicago, pp. 44-54.
- Alexiadou, A. (2017). Building verbs in language mixing varieties. *Zeitschrift für Sprachwissenschaft*, 36(1), pp. 165-192.
- Alexiadou, A. & Lohndal, T. (2018). Units of language mixing: a cross-linguistic perspective. *Frontiers in psychology*, 9, pp. 1-15.
- Alvanoudi, A. (2019). *Modern Greek in Diaspora: An Australian Perspective*. Springer.
- Anastassiadis, A. (1994). *Neologic borrowing in Greek*, Thessaloniki [In Greek].
- Anastassiadis, A. & D. Chila-Markopoulou. (2003). Synchronic and diachronic tendencies in Greek gender: A theoretical approach. In A. Anastassiadis, A. Ralli & D. Chila-Markopoulou (eds.), *The Gender*. Athens: Patakis, 13-56 [In Greek].
- Anastassiadis-Symeonidis, A. & Masoura, E. (2012). Word ending-part and phonological memory: a theoretical approach. *Irregularity in morphology (and beyond)*, pp. 127-140.
- Anastassiadis-Symeonidis, A. & Mitsiaki, M. (2012). Linguistic self-regulation: The case of Greek grammatical gender change in progress. In *Current Issues in Morphological Theory*, pp. 189-216. John Benjamins.
- Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Benmamoun, E., Montrul, S. & Polinsky, M. (2013). Heritage Languages and Their Speakers: Opportunities and Challenges for Linguistics. *Theoretical Linguistics*, 39 (3-4), pp. 129-181.
- Corbin, D. (1987). *Morphologie dérivationnelle et structuration du lexique*, 2 vol. Tübingen: Max Niemeyer. (2nd edition 1991. Villeneuve d'Ascq. Presses Universitaires de Lille).
- Corbin, D. (1991). Introduction-La formation des mots: structures et interprétations, in: *Lexique*, 10, pp. 7-30.
- Field, F. W. (2002) *Linguistic borrowing in bilingual contexts*. Amsterdam: Benjamins.
- Fotopoulou, G. (2004). *Code switching in the case of 2nd generation Greek-German bilinguals: An empirical study*. Stuttgart: University of Stuttgart MA thesis.
- Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.
- Gavriliadou, Z. (2018). Russian Borrowings in Greek and Their Presence in Two Greek Dictionaries. In *The XVIII EURALEX International Congress Proceedings*, pp. 297-307.
- Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. *Loanwords in the world's languages: A comparative handbook*, pp. 35-54.
- Haspelmath, M. & Tadmor, U. (2009). *Loanwords in the world's languages*. The Hague: De Gruyter Mouton.
- Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In T. Stolz, D. Bakker & P.-R. Salas (eds.), *Aspects of language contact: New theoretical, methodological and empirical findings with special focus on Romancisation processes*, Berlin: Mouton de Gruyter. pp. 43-62.
- Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, (26), pp. 210-231.
- Karatsareas, P. (2018). Attitudes towards Cypriot Greek and Standard Modern Greek in London's Greek-Cypriot Community. *International Journal of Bilingualism*, 22(4), pp. 412-428.
- Karatsareas, P. (2019). The Morphology of Silliot: Paradigmatic Defectiveness, Paradigmatic Levelling and Affix Pleonasm. In *The Morphology of Asia Minor Greek*, pp. 148-180. Brill.
- Kavoukopoulos, F. (1996). Nouns, adjectives, and verbs: statistics and other remarks. *Questions of the Modern Greek Language: A Didactic Approach*, pp. 7-16.
- Mager, M., Çetinoğlu, Ö. & Kann, K. (2019). Subword-Level Language Identification for Intra-Word Code-Switching.



- arXiv preprint arXiv:1904.01989.
- Matejka-Hanser, L. (2011). Greek American Greek: Lexical Borrowing in the Speech of Greek Americans. *Wiener Linguistische Gazette*, 750, pp. 84-99.
- Matras, Y. (1998). Utterance modifiers and universals of grammatical borrowing. *Linguistics* 36 (2), pp. 281-331.
- Matras, Y. (2007). The borrowability of structural categories. *Grammatical borrowing in cross-linguistic perspective*, 31-73.
- Matras, Y. (2009). *Language contact*. Cambridge University Press.
- Matras, Y. (2011). Universals of structural borrowing. *Linguistic universals and language variation*, 204-233.
- Matras, Y. & Sakel, J. (2007). Investigating the mechanisms of pattern replication in language convergence. *Studies in Language*. (31) pp.829-865. 10.1075/sl.31.4.05mat.
- Myers-Scotton, C. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press on Demand.
- Pakerys, J. (2016). On the derivational adaptation of borrowings. *Baltistica LI*, (2), pp. 239-269.
- Polinsky, M. & Kagan, O. (2007). Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass* 1(5), pp. 368-395.
- Poplack, S., Pousada, A. & Sankoff, D. (1982). Competing influences on gender assignment: Variable process, stable outcome. *Lingua*, 57(1), pp. 1-28.
- Tadmor, U. (2009.) Loanwords in the world's languages: Findings and results. In Haspelmath & Tadmor (eds.), *Loanwords in the world's languages*. The Hague: De Gruyter Mouton, pp. 55-75.
- Thomason, S. G. & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.
- Tsimpli, I. & Hulk, A. (2013). Grammatical gender and the notion of default: Insights from language acquisition. *Lingua*. (137), pp. 128-144.
- Tsimpli, I. M. (2011). External interfaces and the notion of 'default'. *Linguistic Approaches to Bilingualism*, 1(1), pp. 101-103.
- Van Hout, R. & Muysken, P. (1994). Modeling lexical borrowability. *Language variation and change*, 6(1), pp. 39-62.

### Acknowledgements

This study is part of the project entitled Varieties of Greek as Heritage Language (HEGREEK MIS 5006199). It was held in the frame of the National Strategic Reference Frame (Ε.Σ.Π.Α) and was co-funded by resources of the European Union (European Social Fund) and national resources.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Reports on Lexicographical and  
Lexicological Projects**







# Inventory of New Romanian Lexemes and Meanings Attested on the Internet

Barbu A.M., Lupu I., Stoica-Dinu O., Teleoacă D.L., Toroipan T.

*Institute of Linguistics “Iorgu Iordan – Al. Rosetti”, Romanian Academy, Bucharest, Romania*

## Abstract

This article presents a project that monitors the new lexemes and meanings attested on the Internet for the Romanian language and records them in a descriptive dictionary. This project tries to capture the dynamics of the language in the smallest details (e.g. the loan adaptation process) and to update the lexicographic inventory. The Internet is the best source for this purpose. The article begins with the definition of the term *new word* in the sense of this project, and with the characterization of a descriptive dictionary compared to a normative one. Then the method of selecting new words is described which is of random type, i.e. the words are selected by lexicographers from everyday life, without going through a predetermined volume of texts and are registered provided they have 10 attestations from different sources on the Internet. The article also provides a description of some technical aspects and the structure of the dictionary entries. Some solutions to the problems encountered, the first results and how to continue the project are also discussed.

**Keywords:** new words; Internet; descriptive dictionary; lemma variants; random selection

## 1 Introduction

The main goal of this paper is to present an overview of the project *Inventar de cuvinte și sensuri noi atestate în mediul online* ‘Inventory of new Romanian lexemes and meanings attested on the Internet’, henceforth ICSO, carried out at the Institute of Linguistics in Bucharest. The idea of building the ICSO was born from a previous project, coordinated by a software company, that had as objective the digitization of the Romanian lexicon in literal and phonetic form and with syllable separation (Diaconescu 2015). The digitally processed lexicon was extracted from three electronic explanatory dictionaries: DEX, MDA and DN. At one point, this lexicon was confronted with the words found in a pretty large electronic corpus of books and newspapers, belonging to the coordinating software company, and it was discovered that a relatively big number of common words in the corpus were not found in the digitized lexicon. This result showed the extent to which dictionaries had to be updated. Therefore, the team of the Institute of Linguistics participating in that project decided to inventory and gloss words that are not registered in the main general dictionaries of the Romanian language (which constitute an exclusion corpus – described below), extending the research from a particular corpus to the Romanian language used in the online environment, that is, generally, on the Internet. The preparatory work of ICSO started in 2017 and in 2019 the first volume of about 1400 entries was ready for publication (Barbu et al. 2020). ICSO is currently continuing with the second volume, with slightly changed lexicographic rules (more description-oriented) and as a database.

The project described here aims to monitor the Romanian language especially in the lexical aspect. It tries to update and keep up with the lexical explosion in the Romanian language produced by globalization, advanced technology and increasing access to information through digitization.

This lexical explosion is nowhere better seen than on the Internet, which gathers large volumes of highly diverse texts. “Legal, religious, literary, scientific, journalistic, and other texts will all be found there, just as they would in their non-electronic form”, says David Crystal (2001: 31), who devotes an extensive study to the particularities of the language used on Web. Actually, using Internet as a corpus for linguistic (including lexicographic) research has already been set by linguists such as Fujii & Ishikawa (2000), Kilgariff (2001), Grefenstette (2002) or Fuertes-Olivera (2012), among others. In addition to the benefits already mentioned in the literature, using the Internet as a lexicographic source presents some important features, from our point of view, such as:

- increased access to familiar language and internet slang on forums, comments, social media, etc.;
- access to the multitude of categories of commercial products from online stores, that deserve to be defined for the general public;
- documentation of the adaptation stages of some loanwords, by recording their graphical or morphological variants (e.g. smartphone also circulates in Romanian as smartfone and smartfon, with the plural smartphone-uri, smartfon-uri and smartfoan-e);
- greater opening of specialized terms to the general public.

The main purpose of this project is to build a descriptive new-lexemes inventory, which addresses the following aspects:

- to provide primary material for the explanatory dictionaries which decide what entered and what did not enter the language;

- to offer lexicographic support to the general public for as many terms as it can access on the Internet and not found in the published explanatory dictionaries;
- to provide a record of the words circulating at a given time in the language, even for a very short period, in support of linguists and other specialists from a distant future. Note that some new terms can reflect various language external



events. For instance, the actual coronavirus crisis has already coined or (re)vitalized in Romanian at least twenty candidates (e.g. *coronavirus*, *coronacriză* ‘coronavirus crisis’, *a carantina* ‘to carantinate’, *coronabonduri* ‘corona bonds’, *izoleată* ‘isolation stretcher’, etc.), who are likely to die sooner or later after this crisis is gone, but they could testify over the years for this social event;

- to offer a larger lexical resource for natural language processing (NLP), in a variety of tasks, such as automatic extraction of new words, marketing tasks and sentiment analysis, etc.

In order to make the design of ICSO and the working methodology clearer, in the next section we will define what we mean by new words (or new lexemes) and by what features a descriptive dictionary as ICSO differs from a normative one. The paper continues with a section describing the working method and the entry structure. The following section presents some aspects regarding the inventoried lexemes and the problems encountered in the construction of ICSO. The final section is dedicated to conclusions and further work.

## 2 Normative versus Descriptive Dictionary

In our opinion, the normative (or prescriptive) dictionary that includes new words is the result of a complex and rigorous *selection* process.

First of all, a *new word* is assimilated to the concept of *neologism*, defined for the first time by Zgusta (1971: 179): “neologism is a term which can refer to any new lexical unit, the novelty of which is still felt”. Over time, this definition has been enriched with a series of requirements that constitute many criteria for the selection of words considered as neological. Thus, Cabré (1993: 445) mentions four parameters to determine the neological character of a lexical unit:

- a. diachrony – a unit is neological if it has appeared in a recent period;
- b. lexicography – if it does not appear in dictionaries;
- c. systematic instability – if it shows signs of formal instability (morphological, graphical, phonetic) or semantic;
- d. psychology – if speakers perceive it as a new unit.

Recent literature has paid more and more attention to the selection criteria for neologisms to be introduced in general dictionaries, sometimes following different lexicographic traditions (O’Donovan and O’Neill 2008, Adelstein and Freixa 2013, Sánchez Manzanares 2013, and others). By far the most adopted criteria are the frequency and the stability of new words in the language. Frequency refers to the dispersion in use (although some specialized terms are accepted), and stability refers to the period of time during which the new words are used, excluding (possible) ephemeral units. Another criterion is that of the neological feeling, which refers to the perception of the speakers about a word’s novelty. This criterion excludes analyzable words (with a transparent meaning) from some dictionaries, but requires the introduction of less frequent words, in others (cf. Freixa & Torner 2019). However, the criterion with the highest prescriptive load is the denominative necessity. This refers to the exclusion of new words, often loanwords, which already have a correspondent in the target language.

Despite all these selection criteria, there are no purely normative dictionaries today. Words in nonstandard registers have special use marks (nonstandard, offending, regionalism etc.) that can be seen as prescriptive advices. However these words have entered the dictionary. If they were missing, one would not know whether a word does not exist in the general dictionary because it does not belong to the standard language or because it was accidentally omitted or because the general dictionary is not yet updated. As Rafel (2007: 20) claims, it has even been considered that the real limits between one type of dictionary and the other are not entirely precise.<sup>1</sup> However, a descriptive dictionary still differs by several essential features from one that is not purely prescriptive.

As stated in the literature, “a descriptive dictionary is one that attempts to describe how a word is used, while a prescriptive dictionary is one that prescribes how a word should be used (Naparsteck 2005: 28). In other words, a descriptive dictionary “aims to give a real and complete definition of each lexical item, without any restrictions based on prescriptive criteria.” (Rafel & Soler 2016: 443). However, the descriptive character is understood differently in the specialized literature. One meaning is the one that refers to the way definitions are elaborated. For example, descriptive dictionaries could be called those that use, as definitions, detailed semantic descriptions and paraphrased meaning (in the usual sense of monolingual dictionaries) (cf. Gouws & Prinsloo 2005: 48-49). Another meaning refers to how words are defined based on their use. For instance, Collins COBILD formulates definitions in this way, e.g. the verb *condemn*: “If you *condemn* something, you say that it is very bad and unacceptable”. Besides, this dictionary gives the syntactic pattern(s) in use for each word.<sup>2</sup> It should be added that descriptive dictionaries rely heavily on data collected from very large corpora. This is the case with dictionaries such as the Oxford English Dictionary and Collins COBILD.

In our opinion, a descriptive dictionary, dedicated to neologisms, closely monitors the language, especially pursuing lexical creativity and the way in which speakers use the available linguistic means. In these conditions, the descriptive dictionary is much less or not at all selective. It should not apply any of the above criteria and, thus, nonce words could find their place in such a dictionary. Furthermore, the systematic instability, which Cabré (1993: 445) spoke of, should also be reflected here, given that only a normative approach can establish a standard (morphological, graphical or phonetic) variant of several in use. The use marks (nonstandard, offending, regionalism etc.) have no prescriptive role in this type of dictionary, but a purely descriptive one.

ICSO, as a *descriptive* dictionary, does not apply criteria that are very selective. According to ICSO rules, one main selection criterion is applied, that is, new words are those units that are documented in use, on the Internet, and that are

<sup>1</sup> “Sin embargo, a pesar de esta oposición conceptual, las relaciones entre la actividad lexicográfica de carácter descriptivo y la de carácter normativo son bastante complejas; incluso se ha considerado que los límites reales entre un tipo de diccionario y el otro no son del todo precisos.” (Rafel 2007: 20).

<sup>2</sup> This type of information is also found in other descriptive dictionaries, such as DDLC (Rafel & Soler 2016).



not listed in the lexicographical corpus of exclusion constituted by 7 general dictionaries: DCR, DEX, DEXI, DN, MDA, MDN and NODEX. Note that all these dictionaries are in electronic form and allow a relatively easy search. However, in general, words that do not have at least 10 occurrences in different sources on the Internet are excluded. This condition has been established because on the Web there are numerous sites obtained through automatic translations, which contain words that do not actually belong to the Romanian language or to the Romanian natives. No other selection criteria are applied. Thus, we also introduce words with analyzable structure, which may seem trivial, because this is the Romanian lexicographic tradition and because it is useful for NLP tools. In addition, as Langemets et al. (2019: 9) note, the fact of defining and exemplifying such words is useful for L2 learners. It is worth mentioning that we introduce in the dictionary even the new words and meanings that have been criticized in language cultivation shows, because they are frequent.

### 3 The Building Procedure and the Entry Structure

Because our institute does not have an IT department that facilitates the (semi-)automatic search of new lexemes and due to the lack of fundamental electronic language resources (exclusion lists, large corpora, reliable IT tools, etc.) the new lexemes / meanings selection is done manually, by the lexicographers involved in the project. This method differs from (semi-)automatic methods (Klosa & Lungen 2018, Kerremans et al. 2012, Falk et al. 2014) in that it does not restrict the search to previously fixed sources to which regular crawling is applied, but it addresses any source on the Internet that uses Romanian and which, preferably, does not represent (automatic) translations from other languages. Online sources can be mass media, blogs, company sites, stores, forums, portals, books, social media etc. In this project, we select single words, multi-word expressions, abbreviations, relevant proper names that serve as derivation bases (e.g. Facebook, Instagram, Nobel, etc.), new elements of word formation (e.g. e- “electronic”, robo- “robotic”, etc.) and, also, new meanings of older words, provided they gather at least 10 occurrences from different sources and are not found in the exclusion corpus. Each of these constitutes a separate entry in our dictionary, regardless of the existence of morphological or semantic links between them.

Even if the selection is done manually, it is not done in the traditional way, as described by Kilgarriff et al. (2015: 196), that is, by ‘reading and marking’: “Lexicographers read texts which are likely to contain neologisms – newspapers, magazines, recent novels – and mark up candidate new words, or new terms, or new meanings of existing words”. About this method the cited authors claim that it is a low-recall approach, because lexicographers cannot read everything, so there are many neologisms that will be missed. Unlike this method, with low-recall, we instead use a random selection method. Each lexicographer writes down every word seems to be a candidate new word (or meaning), regardless of the communication context in which it is discovered (Web, television, individual talks, personal lectures etc.). If the candidate word is not in the exclusion corpus, we look for further attestation on the Internet, where we find other contexts – of which we choose examples for ICSO – which, in turn, may contain other new words. This way, one can get clusters of new words related by a certain domain or social layer or topic etc. For instance, from a Romanian news story about setting up a skateboard park in a certain locality, a cluster of elements specific to this sport was obtained, e.g. ollie (box), trick, freestyle (board), (grinding) rail, bank, quater pipe, etc. It is worth mentioning that in the list randomly built by lexicographers only about 10% of the candidate words already existed in the exclusion corpus. This gives a hint about the high-precision of the lexicographers’ selection, favoured also by the fact that the Romanian explanatory dictionaries do not have, until ICSO project, a consistent and sustained updating program. Furthermore, the lexicographer is not required to read large amounts of text from predetermined sources, hoping to find new words, but they are detected in everyday life according to personal interests. Metaphorically speaking, new lexemes come to the lexicographer provided that he/she pays constant attention to them. Finally, each member of the project team has his/her own list of candidates. Team members’ lists may partially overlap, but so far, the number of overlapping words has been relatively small (around 10-15 words in multiple lists). One criticism that could be made of this method is that it does not involve a systematic search and that the precision and recall of the method cannot be accurately calculated. This is true, but we believe that this method is the fastest, least demanding and most productive in terms of the number of new lexemes detected from a source as wide as the Internet. This method could be compared to others, based on the number of new lexemes inventoried in a year with an equivalent workforce, but we do not know such evaluations of other methods.

The inventory is built using the Professional Lexicography Software TshwaneLex ([tshwanedje.com](http://tshwanedje.com)). This lexicographic editor has many facilities but we only mention the XML editor, the possibility of RTF or HTML export and the so-called WYSIWYG (“what you see is what you get”) view. These facilities help to easily get the dictionary in machine-readable and printed formats. We build the inventory in XML format and in the form of an ODBC database, remotely accessed by the lexicographers (see Figure 1). This format helps lexicographers to see, any time, the whole work and to better collaborate. Moreover, the risk that more lexicographers select and work the same entries independently is eliminated.

As can be seen in Figure 1, the workspace in TshwaneLex is divided into 3 main areas. The area on the left contains the list of all (completed or not) entries in the database. The middle area has a part at the top where the XML elements suitable for the written entry are chosen, according to the hierarchical structure defined in the DTD (Document Type Definition). The selected XML elements are filled with content at the bottom of the middle area. The entries in printed form are displayed in the left area. Entries preceded by a small padlock show that they are “locked” in the database and cannot be opened as long as another user is working on them, so as not to create conflicts. It should be noted that lexicographers do not have exclusivity in writing an entry, the editor or other colleagues may intervene in entries made by others. Of course, the database allows access to remote lexicographers, being very suitable for homework specific to this historical period and also allows access to several users at the same time.



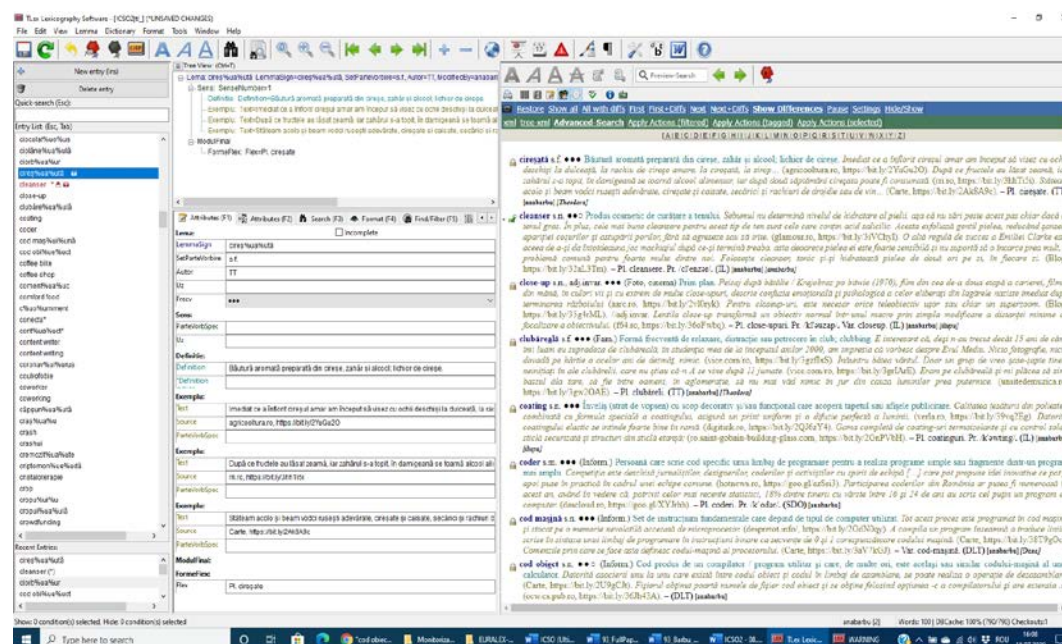


Figure 1: TshwaneLex Framework.

ICSO has the usual entry structure of a general dictionary, see Figure 2. On the lemma, the stressed vowel is indicated, if the lexeme is pronounced as in Romanian, otherwise the pronunciation is indicated in a special field. After the lemma, there is a series of parts of speech associated with it. If necessary, the part of speech is resumed in each sense. For the parts of speech obtained by conversion, no separate senses are described, but only specific examples are given, preceded by the respective part of speech (see *adv.* in Figure 2). The following field indicates the frequency of the lemma on the Internet. This is done with a very rough approximation, because the real dimension of the Internet is completely unknown and an exact count of valid occurrences is impossible. We use 3 symbols: ●○○ (i.e. low frequency) – if the lemma has been found (with Google) on less than four Internet pages; ●●○ (i.e. medium frequency) – if the number of pages containing the lemma is placed between four and ten pages, and ●●● (i.e. high frequency) – if the lemma is present on more than 10 pages. The number of pages is, of course, indicative, because the linguist must eliminate the results in which the lemma appears, for example, in company names, as search tags, in non-Romanian texts, etc. Despite this very rough numerical approximation, we consider that, lexicographically, this frequency information is useful.

The next field, after the frequency, is dedicated to usage information that may refer to selective restrictions, domain, stylistic register, etc. This is followed by the definition of the lemma, expressed by paraphrase and / or synonyms. We try as much as possible not to use, in the definition, words from the same lexical family as the lemma, in order to make the definition as clear and independent as possible. After the definition, a few examples are given that reflect the use of the lemma (and its variants, if any). Each example is accompanied by its online source, consisting of two fields: the web domain and an abbreviated link obtained with the public applications [goo.gl](http://goo.gl) or [bit.ly](http://bit.ly). The web domain can provide a good hint about the level of education expected from the language of the text. For example, if this domain belongs to a television station or public institution, one expects the language to be more elevated than if it belongs to a comment on a post or to a forum. The abbreviated link allows the reader to go directly to the site from which the example was taken, to see the expanded context. Another way to get to that site is to search Google for the exact text of the example. It should be noted that choosing the examples is perhaps the most laborious task. This choice must take into account several criteria such as: a) the most credible source of the example (newspapers published in Romania are preferred and sources with Romanian translations are avoided);<sup>3</sup> b) the most appropriate illustration of the meaning (possibly with its explanation); c) the length and clarity of the example. Article titles are also accepted as examples and it should be noted that sometimes a new lexeme appears only in the title (probably due to its brevity), but not in the article.

ICSO entries end with a module containing 4 properly marked fields:

1. Immediately after the symbol “—” the standard inflectional forms are given (if any), for nouns: the plural, and for adjectives: feminine singular, masculine and feminine plural. If one of these is not actually used, the symbol ^ precedes the unattested form. For verbs, the present indicative form of the first or third person and the participle form are specified.
2. After the mark “Pr.”, the pronunciation of foreign words or an accent variant (see Figure 2) is given. IPA symbols are used for pronunciation, but only those specific to the Romanian language, because we assume that an ordinary Romanian speaker is not sensitive and would not know how to pronounce the sounds he is not used to.
3. After the mark “Var.”, the circulating variants of the lemma are indicated. Often, these variations reflect loanword adaptation trends, such as reduction of double consonants, e.g. *contactles* in Figure 2. Variants are also listed, with reference to the basic form of the lemma.

<sup>3</sup> The Romanian language is also spoken in the Republic of Moldova, but there are obvious differences between the language varieties spoken in the two states.



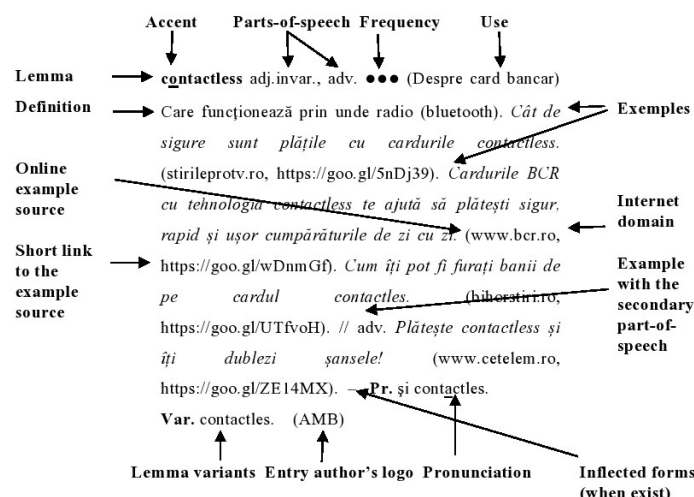


Figure 2: The Entry Structure.

4. After the “Abr.” mark, a possible abbreviation of the lemma is given, for example RV for *realitate virtuală* ‘virtual reality’. Abbreviations are described as standalone entries that have the same definition as the corresponding abbreviated lexeme and examples focusing on the abbreviation uses.

The last information in the entry refers to its author through an individual logo, e.g. (AMB) in Figure 2. This is an innovation in the lexicographic practice. For communication between lexicographers – both those involved in the project and those who use their results – this innovation has already proved its worth. Other justifications are given below, in section 4.

It should be noted that ICSO does not provide information on the first attestation and on the etymology of the new lexemes. Although, for the very old words, the first attestation is very important, sometimes attracting ample studies to establish it, nowadays, the first attestation is not so relevant. In fact, it is very difficult to establish it in such a large volume of texts. An approximate dating of a new lexeme may be made, in the future, depending on its appearance in a dictionary such as ICSO. A regular and sustained series of a dictionary of new words can be a good benchmark for the first attestation of a lexeme or, more precisely, for the moment of its entry into the language. As for etymology, we have omitted this information so as not to slow down the work. Establishing an etymology is not as trivial as it seems and requires specialized research that is time consuming.

The lemmas that already exist in the exclusion corpus are marked with the symbol \*. They are entered in our dictionary if they have new meanings or the definitions in the other dictionaries are (no longer) appropriate or if they present any other element of novelty (e.g. changes in spelling, selective restrictions, etc.).

#### 4 First Results and Encountered Problems

During two years (including one year for the project design and lexicographic rules setting), about 1400 entries have been built by 5 lexicographers working part time (1,5 full time).

Regarding the lexical creativity reflected by the new lexemes, we do not intend to make a detailed description here, but only mention a few aspects. It is worth mentioning that most of the inventoried lexemes, at this stage, represent lexical creations and extensions of lexical families in Romanian, and the loanwords in original form and their naturalized derivatives cover only 20% (despite the general perception of the overwhelmingly invasion of anglicisms, for instance). Another aspect concerns a whole series of words that, most likely, will not last in time, but which reflect a notable social phenomenon: the strong influence of highly publicized people on society. Lexical families of some proper names belonging to persons very frequently seen on TV were registered. For example, the name of a controversial businessman, Becali, created the family made of adj. *becal-ic* / *becal-ist* ‘specific / loyal to Becali’, vb. *becal-iza* (infinitive) / *becal-izat* (past participle) ‘become as Becali’. Of course, the person’s name Becali is not entered as a separate entry, as we consider that the reference in the definition to the person with that name is sufficient.

Graphic adaptation of loanwords can be traced with the help of the registered variants of different lexemes. They reflect the following types of adaptations, among others:

- phonetic writing, specific to the Romanian language: *brandui* ‘to brand’ > *brendui*, which is written as it is pronounced; *flash* > *fleş*; *foosball* > *fusbal*; *rider* > *raider*; *vlogger* > *vlogăr*; etc.
- deleting letters that are not pronounced (double consonants, mute vocals etc.): *contactless* > *contactles*; *couponing* > *cuponing*; *fratello* > *fratelo*; *pattern* > *patern*; etc.
- writing in ordinary Romanian letters: *cyberterorism* > *ciberterorism*; *flash* > *flaş*; *photoshopare* ‘photoshopping’ > *fotoșopare*; etc.
- changing the English plural ending to the Romanian one: *dreadlocks* > *dreadlock-uri*, etc.

In fact, the vast majority of variants reflect hesitations in using the hyphen to mark compounds or derived words. For instance, for compounds, especially loanwords, there are almost complete series of variants: *flash-mob* / *flash mob* /



*flashmob* / *fleşmob*; *microjob* / *micro-job* / *micro job*; *off-grid* / *off grid* / *offgrid*; etc. But there are also hesitations regarding Romanian compounds: *cardiotoracic* / *cardio-toracic* ‘cardiothoracic’; *colorectal* / *colo-rectal* ‘colorectal’; *euroentuziast* / *euro-entuziast* ‘one who has full confidence in the values of the European Union.’; etc. The same phenomenon is found in words derived in Romanian with prefixes: *anti-avort* / *antiavort* ‘anti-abortion’; *co-inventator* / *coinventator* ‘invention partner’; *interreligios* / *inter-religios* ‘interreligious’; or suffixes: *rohmerian* / *rohmer-ian* ‘in Rohmer’s style’.

The graphic variation can also come from other sources, such as writing abbreviations as they are pronounced: *CAP-ist* / *ceapist* ‘worker in an agricultural cooperative’; *PSD-ist* / *pesedist* ‘member in the Social-Democrat Party’; or lowercase / uppercase writing: *new-age-ist* / *New-Age-ist* ‘adept of New Age philosophy’, *secret Santa* / *Secret Santa*, *youtuber* / *YouTube* ‘person who regularly makes and posts videos on youtube.com’.

The recording of all these variants, useful for a deeper understanding of language trends, also creates some problems. The main problem we encountered in building ICSO is related to the fact that our institute is seen as the main author of normative academic workings. Thus, we faced the same problem reported by Joaquim Rafel (see also Langemets et al. 2019: 12):

Uno de los problemas que plantea la elaboración de un diccionario descriptivo por una academia de la lengua es que este diccionario contiene palabras o acepciones que no son reconocidas por la normativa vigente, a pesar de encontrarse documentadas en los textos; por una parte este diccionario puede ser considerado más científico que el normativo por cuanto intenta dar cuenta de una manera sistemática de la realidad de la lengua a partir de datos empíricos, pero por otra parte puede ser visto como un peligro para el uso lingüístico considerado correcto. (Rafel 2007: 21)

Actually, in general, the public demands linguistic norms even for things that cannot be normed. It is so eager for normative works. Therefore, given this expectation, a descriptive dictionary can create confusion, because it contains many variants under which a lexeme circulates and many terms belonging to the „colourful” language used in forums, comments, etc. To draw attention to the fact that not all forms belong to the literary language, we have adopted the following solutions (in addition to the warning note in the generally ignored introduction).

Firstly, we have called this work “inventory”, not “dictionary”. The Romanian term “inventory”, mainly used in dialectology, suggests a simple enumeration and differs from the usual titles of academic dictionaries.

Secondly, we have paid more attention to the stylistic-use information, such as depreciative, ironic, affectionate, etc. We could not mark the unrecommended entries, as it has been done in Rafel (2007: 21) for instance, because ICSO entries have not been yet subject to prescriptive analysis.

Finally, we have adopted an innovative solution by ending each ICSO entry with its author's logo, in order to highlight the fact that the entry content is under the responsibility of a person, not an authority. For the general public, the personalization of the entry could diminish the normative perception, and for the lexicographer colleagues it helps to a better communication later. The author logo also concerns the loanwords pronunciation which, in the absence of specialized studies, reflects the choice of the entry author.

Another encountered issue is related to the volatility of the examples on the Internet. This is indeed a risk, mitigated by the fact that if a lexeme is certified at least 10 times, it is likely to remain attested even if the example in ICSO disappears. In addition, when selecting examples, archival sources are preferred and volatile texts, like those found on sites of second-hand sales, are avoided.

## 5 Conclusions and Further Work

ICSO is a descriptive dictionary that monitors new lexemes in Romanian, in order to capture the dynamics of the language in the smallest details (e.g. the loan adaptation process). The main purpose is to provide primary material for the systematic updating of normative dictionaries. A secondary goal is to provide the general public with explanatory definitions for the explosion of terms on the Internet.

In order to draw the public's attention to the fact that this is not a normative dictionary, we have adopted small lexicographical innovations such as the title “Inventory” and the indication of the entry author through a logo. In order to ensure an increased working speed, required by the urgent need to update general dictionaries, we have adopted the random selection method which excludes the regular browsing of predetermined sources. We also waived the information regarding the first attestation and the etymology of the registered lexemes. The use of the Internet as a lexical source allows us access to a field almost ignored by normative dictionaries, that of familiar language in social media, which actually reflects everyday language.

The project aims at the sustained elaboration of a volume of at least 1500 words every 2 years, taking into account the inherent delays due to publication. The publication of the dictionary on the site of our institute is considered, possibly in a dynamic way, so that the public can benefit as soon as possible from the definitions of the new lexemes. Another aspect worth considering is the use of the crowdsourcing method for word gathering. With the launch of ICSO2 site, a facility can be created to allow the general public to suggest new words to be introduced in the dictionary. Furthermore, we hope that in the near future we can use tools for automatic search of new lexemes that have sufficiently good results.

## 6 References

- Adelstein, A. & Freixa, J. (2013). Criterios para la actualización lexicográfica a partir de datos de observatorios de neología. Unpublished presentation at Congreso Internacional El Diccionario: neología, lenguaje de especialidad, computación, Ciudad de México (Mexico), 28-30th October 2013. Accessed at: <https://repositori.upf.edu/handle/10230/34891> [04/10/2020].



- Barbu, A. M., Croitor, B., Niculescu-Gorpin, A. G., Radu, I. C. & Vasileanu, M. (2020). *Inventar de cuvinte și sensuri noi atestate în mediul online (ICSO 1)*, vol. 1. Editura Academiei.
- Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona: Editorial Antártida/Empúries.
- Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.
- DCR – Dimitrescu, F., Ciolan, Al. & Lupu, C. (2013). *Dicționar de cuvinte recente*. Editura Logos.
- DEX – *Dicționarul explicativ al limbii române* (1998-2016). Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti”: Editura Univers Enciclopedic Gold.
- DEXI – Dima E. (ed.) (2007). *Dicționar explicativ ilustrat al limbii române*. Editurile ART și GUNIVAS.
- Diaconescu, S. Ș. (ed.) (2015). *Fonetica limbii române*, 4 vol. SOFTWIN: CreateSpace Independent Publishing Platform.
- DN – Marcu, F. & Maneca, C. (1986). *Dicționar de neologisme*. Editura Academiei.
- Falk, I., Bernhard, D. & Gérard, C. (2014). From Non-Word to New Word: Automatically Identifying Neologisms in French Newspapers. In *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, May 2014, Reykjavik, Iceland, pp. 4337–4344.
- Freixa, J. & Torner, S. (2019). Beyond Frequency: On the Dictionarization of New Words in Spanish. In *Kernerman Dictionary News* 27, July 2019, p. 6. (Presentation at the Globalex Workshop on Lexicography and Neologism, Bloomington, Indiana, US, 8th May 2019.) Accessed at: [www.academia.edu/39070136/](http://www.academia.edu/39070136/) [04/10/2020].
- Fuertes-Olivera, P. A. (2012). Lexicography and the Internet as a (Re-)source. In *Lexicographica* 28(1), pp. 49-70.
- Fujii, A. & Ishikawa, T. (2000). Utilizing the world wide web as an encyclopedia: Extracting term descriptions from semi-structured text. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL-2000, 3-6 October 2000*, Hong Kong, pp. 488-495.
- Gouws, R. H. & Prinsloo, D. J. (2005). *Principles and Practice of South African Lexicography*. SUN Press.
- Grefenstette, G. (2002). The WWW as a Resource for Lexicography. In M.-H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Grenoble, France: EURALEX, pp. 199-215.
- Kerremans, D., Stegmayr, S. & Schmid, H.-J. (2012). The NeoCrawler: Identifying and Retrieving Neologisms from the Internet and Monitoring Ongoing Change. In K. Allan, & J. A. Robinson, (eds.) *Current Methods in Historical Semantics*. De Gruyter Mouton, pp. 73-59.
- Kilgariff, A. (2001). Web as corpus. In *Proceedings of Corpus Linguistics 2001*. Lancaster, UK (March). Reprinted in G. Sampson & D. McCarthy (eds.), *Corpus Linguistics. Readings in a Widening Discipline*. 2004. London and New York: Continuum, pp. 471–473.
- Kilgariff, A., Herman, O., Bušta, J., Kovář, V. & Jakubíček, M. (2015). DIACRAN: a framework for diachronic analysis. In *Corpus Linguistics 2015. Abstract Book*. Lancaster: UCREL, pp. 195-197.
- Klosa, A. & Lungen, H. (2018). New German Words: Detection and Description. In J. Čibej, et al. (eds.) *Proceedings of the XVIII EURALEX International Congress*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 559-569.
- Langemets, M., Kallas, J., Norak, K. & Hein, I. (2019). New Estonian Words and Senses: Detection and Description. In *Kernerman Dictionary News* 27, July 2019, p. 8. Accessed at: [globalex.link/wp-content/uploads/2019/05/gwln2019\\_langemets-kallas-norak-hein.pdf](http://globalex.link/wp-content/uploads/2019/05/gwln2019_langemets-kallas-norak-hein.pdf) [04/10/2020]
- MDA – *Micul Dicționar Academic* (2010). Institutul de Lingvistică „Iorgu Iordan – Al. Rosetti”: Editura Univers Enciclopedic.
- MDN – Marcu F. (2000). *Marele Dicționar de Neologisme*. Editura Saeculum.
- Naparstek, M. (2005). *Honesty in the Use of Words*. Rochester, New York: Lake Affect Publishers.
- NODEX – *Noul Dicționar Explicativ al Limbii Române* (2002). Litera Internațional: Editura Litera Internațional.
- O'Donovan, R. & O'Neill, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. In E. Bernal, J. DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress*, Barcelona, 15-19 July 2008. Barcelona: IULA-UPF, pp. 571-579.
- Rafel i Fontanals, J. & Soler i Bou, J. (2016). A Descriptive Dictionary of Contemporary Catalan: The DDLC Project. In E. Corino, C. Marelló, C. Onesti (eds.) *Proceedings of the XII EURALEX International Congress*, vol. I, Turin, Italy, 6–9 September 2006. pp. 443-455.
- Rafel i Fontanals, J. (2007). Prescripció y descripción en la actividad académica: el Diccionari descriptiu de la llengua catalana. In M. Campos et al. (eds.) *Reflexiones sobre el diccionario, Actas del I Congreso Internacional de Lexicografía Hispánica*, Coruña: Setembre 2004. Universidade da Coruña, pp. 9-33.
- Sánchez Manzanares, C. (2013). Valor neológico y criterios lexicográficos para la sanción y censura de neologismos en el diccionario general. In *Sintagma* 25, pp. 111-125.
- Zgusta, L. (1971). *Manual of Lexicography*. (Janua Linguarum Series Maior 39). Prague/The Hague: Academia/Mouton.







# LBC-Dictionary: a Multilingual Cultural Heritage Dictionary. Data Collection and Data Preparation

Farina A.<sup>1</sup>, Flinz C.<sup>2</sup>

<sup>1</sup> University of Florence, Italy

<sup>2</sup> University of Milan, Italy

## Abstract

An increasing number of a wide variety of texts on Italian cultural heritage are available today, both online and on paper. However, there are no specific tools (dictionaries, reference materials on technical translations) that can train and support specialists involved in cultural tourism. Mainly focusing on Florence and its cultural heritage, the LBC project (Farina 2016) will try to fill this gap by providing tools for those who have to write/translate for dissemination in various languages: in a first step by building monolingual corpora (English, French, German, Italian, Russian, Spanish) that the user can freely search; in a second step by developing a plurilingual LSP internet dictionary on cultural heritage which uses the above-mentioned corpora as a primary source. The aim of this paper is to give an insight in the lexicographical process of the LBC-Dictionary, concentrating in particular on data collection and data preparation, which, as is usual for dynamic dictionaries, are open-ended and ever ongoing (Klosa 2013). In particular, we will illustrate the main characteristics of the French and German LBC Corpora and reflect on the provisional French and German entry list, also illustrating the procedure adopted, an alternation of corpus-driven and corpus-based steps (Tognini-Bonelli 2001), for their extraction.

**Keywords:** Corpora; cultural heritage; internet dictionary

## 1 Introduction<sup>1</sup>

An increasing number and a wide variety of texts on Italian cultural heritage are available today, both online and in print, from tourist guidebooks to museum web sites, from art catalogues to critical essays. Provided in different languages, these works attempt to satisfy an international public increasingly in need of information on Italian cultural heritage. However, at present across Europe, there are no specific tools (dictionaries, reference materials on technical translations) that are able to convey such knowledge in an appropriate way (Billero/Nicolas Martinez 2017: 203), or specialised institutions that can train and support specialist translators and other specialists involved in cultural tourism (tourist guides, tourist information centres, museum staff, etc.).

The LBC project (Farina 2016), which involves experts from different disciplines (among others lexicography, corpus linguistics etc.) and universities (Florence, Bologna, Lisbon, Milan, Paris, Pisa etc.), tries to fill this gap mainly by focusing on Florence and its cultural heritage and providing tools for those who have to write/translate for its dissemination in the various languages. Our principal aim is to create monolingual dictionaries of Italian Heritage in all the languages involved in our project, which could be used as plurilingual tools thanks to translation links created among them.

In a first step, we have built monolingual comparable corpora (English, French, German, Italian, Russian, Spanish, see Figure 1)<sup>2</sup> that could also serve the principal target user of our dictionaries (persons who must write or translate texts about Tuscan Cultural Heritage). We decided not to set limits of time and place, but to use each text by referring to the cultural heritage of the city of Florence<sup>3</sup> in each language featured in the project:

...the city of Florence as it has appeared in the actual use of language over the centuries and in the discourses of the people who have described it in the seven languages featured in the project, thinking that in this way the cultural basis might emerge and lead us to design lexicographical articles which shed light on the cultural and historical connotations of the words actually used to describe it. (Farina 2015: 125)

<sup>1</sup> The present contribution was conceived jointly by the two authors and discussed in detail in its individual parts, in particular Annick Farina the French section, and Carolina Flinz the German one.

<sup>2</sup> In a next step, the corpus platform will also contain parallel corpora (see Zotti 2017).

<sup>3</sup> Its extension to other cities and to the entire region of Tuscany is in the planning stage.



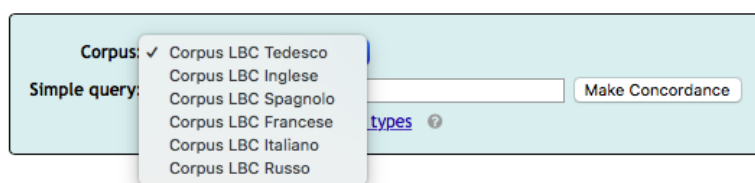


Figure 1: Screenshot of the available corpora (04.2.2020).

The corpora make use of open-source *no-sketch engine*<sup>4</sup> software, which provides users with several functionalities (including the search for a word in its context, the sorting of references according to various criteria, the filtering of texts according to text types and authors) and thus enables intra- and interlingual applications, from both a synchronic and diachronic perspective (see Ballestracci/Buffagni/Flinz in prep.).

As a second step, these corpora will be used as the primary source (Engelberg/Lemnitzer 2009: 235-237) for the plurilingual LSP internet dictionary on cultural heritage, which is currently in the planning stage (Farina/Billero 2018): with an interplay of automated procedures and manual selection/interpretation (Geyken/Lemnitzer 2016: 208), the data will be extracted from the above-mentioned corpora to construct the provisional entry list.

The aim of this paper is to provide an insight into the lexicographical process of the LBC-Dictionary, by concentrating in particular on data collection (1) and data preparation (2), which, as usual for dynamic dictionaries, are open-ended and ever ongoing (Klosa 2013). We will:

- (1) concentrate in particular on the LBC French and German Corpora, illustrating their main characteristics: size, text types (among others popular, technical, literary texts), time period involved (from the Renaissance to the present), authors etc.;
- (2) focus on the provisional French and German entry list, also illustrating the procedure adopted, an alternation of corpus-driven and corpus-based steps (Tognini-Bonelli 2001) for their extraction. The main characteristics of the list will be presented in order to reflect on their items.

We will conclude by proposing perspectives, for example the extraction of the concordances related to headword lists.

## 2 Data collection

In the data collection phase, the sources for the dictionary base have been compiled. Corpora are the primary sources<sup>5</sup> of many contemporary dictionaries (Klosa 2020: 11), and their use in the lexicographical process, mostly with a quantitative-qualitative approach, has opened up a variety of new possibilities (Lemnitzer/Zinsmeister 2015: 170) that were previously unthinkable with traditional collections of documents, and opportunities impossible with any other type of source, since they are accessible regardless of location and provide an authentic picture of the language depicted (Geyken/Lemnitzer 2016: 203).

Even if we were advised to use existing corpora for a variety of reasons (i.e. they realise the criteria of size and representativeness, see Lemnitzer/Zinsmeister 2015: 137), in our project we could only partially follow this procedure because there are no existing LSP-Corpora on art and it was impossible to create virtual LSP-Corpora from existing ones. So, we decided to adopt a combined procedure of using both *ad hoc* created monolingual LSP-Corpora and existing reference corpora<sup>6</sup>. As secondary sources we used existing monolingual and bilingual dictionaries (among others TLFi, Duden online, Zanichelli 2009) and, as tertiary sources, manuals and grammars.

### 2.1 LBC French and German Corpora

The creation of corpora is associated with methodological problems that, however, can be solved with careful planning (see Flinz 2019; Hunston 2008; Lemnitzer/Zinsmeister 2015). For example, among others: 1. The choice of the type of corpus must be carefully considered, since not all corpus types are suitable for all lexicographical purposes; 2. The requirements of the corpus, as for example its size (see Kupietz/Schmidt 2015: 302f), must be taken into consideration, because the larger a corpus is, the higher the probability of finding rare constructions or obtaining good results from statistical analyses will be (see Geyken 2007: 37); 3. The origin and quality of the texts, which should not be chosen

<sup>4</sup> While our researches depend on *public found* (*public Universities*) we publish all tools in Open Access: corpora can be freely searched by its above-intended users. See the LBC-Platform, <http://corpora.lessicobeniculturali.net> (04.02.2020).

<sup>5</sup> For the division in primary, secondary, tertiary sources see Wiegand 1998: 140.

<sup>6</sup> See §2.1



arbitrarily<sup>7</sup>; 4. The documentation of primary sources to ensure the value of a corpus.

In our project the basic jointly determined criterion for selecting works and authors was their importance for Florentine Renaissance art and culture, considering both translations and original texts: The German and French LBC corpora in fact - like the other LBC corpora - consist of original language texts as well as texts translated from the other languages of the project (Italian, French and English). In their nature as monitor corpora (Lemnitzer/Zinsmeister 2015: 140) they can be constantly expanded, so what we present in this paper is only a snapshot of the actual situation, but when we decided to extract the provisional entry list, we fixed a minimum of 1,000,000 words (Table 1):

	French LBC-Corpus	German LBC-Corpus
Tokens	3,818,747	1,183,484

Table 1: Size of the French and German LBC-Corpora

Each LBC-Corpus contains texts that belong mainly to two macro categories: technical and literary texts<sup>8</sup> (see Figure 2).

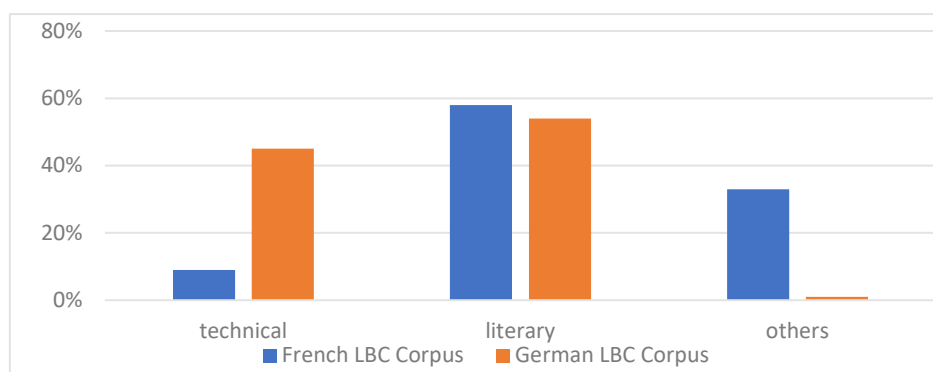


Figure 2: Diagram of the French and German LBC-Corpora

As we can see from the diagram, the-most important difference is the higher number of technical texts in the German Corpus and of the category 'others' in the French Corpus. In this latter group we can find mainly informative texts. The aim of the research group is now to integrate texts into the corpus of any language involved, especially in the above-mentioned macro-categories, in order to increase their comparability (see Billero/Farina/Nicolas in prep.).

Among the literary texts, we can find different texts genres, such as biographies (for example, travel diaries and travelogues), fictional narrative works (i.e. short stories, novels and other literary narrative texts with Florence or Tuscany as their main theme or locale) and essays. The technical texts comprise LSP-texts focusing especially on art and on architecture. The texts are mostly original ones. Our purpose was in fact to guarantee a certain variety for all text types, while giving preference to those that were particularly representative of Italian culture and art and its international dissemination and reception<sup>9</sup>. Concerning the involved authors<sup>10</sup> and the diachronic variation, we decided to offer a certain variety: 74 authors<sup>11</sup> for French and 16 authors<sup>12</sup> for German. The time laps covered by the texts (date of writing

<sup>7</sup> See also Farina/Billero 2018 for the semi-automatic evaluation of text translations.

<sup>8</sup> For a detailed description of the corpora see Farina in prep. for the French LBC Corpus and Ballestracci/Buffagni/Flinz in prep. for the German LBC Corpus.

<sup>9</sup> As for example: Giorgio Vasari's *Vite* (1550, 1568) - a work that is fundamental for the art and culture of the Renaissance, and that contributed to spread the myth of the Italian Renaissance in most European countries; non-Italian authors, who played a major role in spreading the Italian Renaissance culture in foreign countries (John Ruskin, Jacob Burkhardt); and famous authors or writers who travelled to Italy and written about it (among others Dumas and Stendhal for the French Corpus and Johann Wolfgang Goethe, Karl Philipp Moritz for the German one).

<sup>10</sup> We have mentioned only the authors belonging to the literary and technical field.

<sup>11</sup> Allais, Auzias, Bard, Bazin, Beaugrand, Bertheroy, Broses, Camus, Cellini, Chateaubriand, Colet, Colin, Colombari, Creuzé de Lesser, da Vinci, De la Borie, De Navenne, Delacroix, Dufay, Dumas, Erdan, Faudre, Favre le Bret, Félibien, Fernandez, Feuillet, France, Fréville, Gaboury, Gautier, Giono, Goncourt, Goupil, Grandgeorge, Grimaldi, Jaucourt, Klaczko, La Sizeranne, Labourdette, Lafenestre, Lang, Le Routard, Lescure, Libri, Machiavelli, Mallarmé, Mallet, Maurel, Méry, Meyer, Michel, Michel-Ange, Montaigne, Moran, Musset, Nobecourt, Palustre, Pasquin, Perrot, Pommier, Powell, Prieur, Renan, Revel, Rosov, Schmitz, Staël-Holstein, Stendhal, Taillasson, Taine, Vasari, Viollet Le Duc, Wyzewa.



in both languages or date of translation) is from the 16<sup>th</sup> to the 21<sup>st</sup> century.

### 3 Data preparation

Central steps in the data preparation phase are extracting the provisional dictionary entry list and modelling the lexicographical data into a database structure. In our paper we concentrate on the first aspect by illustrating the alternation of the corpus-driven and corpus-based procedures (Tognini-Bonelli 2001) used for extracting the list.

#### 3.1 LBC French and German entry list

We could not create the provisional lemma list on existing lemma lists of other dictionaries, since no lexicographic resources of this type exist, so instead we used a combination of different corpora as our primary source:

- *ad hoc* created monolingual LSP-Corpora (see table 1);

- reference corpora of the involved languages: for French we used *L'Est Républicain* (15,000,000 tokens) and in the Sketch Engine integrated *French Web 2017* (frTenTen17) with 6,845,630,573 tokens; for German we chose das *Deutsche Referenzkorpus DeReKo* (2017-I, Release of 08.03.2017) and the in the *Sketch Engine* integrated *German Web 2013* (deTenTen13) with 19,808,173,163 tokens.

We first extracted automatically and manually different types of word lists for each language (see Table 2):

	List-Name	Corpora	Measure	Automatic/Manual	N.
French	K-LBC (fr)	LBC-Korpus (fr) / frTenTen17	keyness Score	automatic	2000 single units 2000 multiple units
	L-LBC (fr)	LBC-Korpus (fr)	absolute frequency	automatic	25,337
	K-L-RIF (fr)	L'Est Républicain	absolute and relative frequency	automatic	145,644
	G-LEX (fr)	Dictionaries		manual	1806
German	K-LBC (de)	LBC-Korpus (de) / deTenTen13	keyness Score	automatic	2000 single units 2000 multiple units
	L-LBC (de)	LBC-Korpus (de)	absolute frequency	automatic	45,029
	K-L-RIF (de)	LBC-Korpus (de) / DeReKo	<i>chi2</i> e <i>LLR</i>	automatic	10,402
	G-LEX (de)	Dictionaries		manual	2547

Table 2: Size of the French and German LBC-Corpora

1) a keyword list based on the reference corpora integrated in the *Sketch Engine* (K-LBC). The K-LBC Lists were automatically driven by using the function *Keywords* of *Sketch Engine*. We extracted 2000 keywords and 2000 multi-words expressions (see Figure 3 and 4, which show the first ten German single and multiword keywords), representing the most typical items of both corpora.

<sup>12</sup> Alberti, Brandi, Burckhardt, Cellini, da Vinci, Gass, Goethe, Heine, Kurz, Machiavelli, Moritz, Ruskin, Stendhal, Vasari.



	Word	Frequency <sup>?</sup>		Relative freq. <sup>?</sup>		Score <sup>?</sup>
		Focus	Reference	Focus	Reference	
1	verfertigen	512	6,194	426.647	0.313	325.78
2	disegno	348	239	289.986	0.012	287.52
3	Florenz	1,475	79,843	1,229.109	4.031	244.52
4	Medici	495	14,431	412.481	0.729	239.21
5	Florentiner	373	7,634	310.819	0.385	225.08
6	Cosimo	342	6,170	284.987	0.311	218.06
7	Filippo	367	8,189	305.819	0.413	217.08
8	Lorenzo	750	37,771	624.971	1.907	215.34
9	diligenza	249	3	207.49	0	208.46
10	florentinisch	264	3,629	219.99	0.183	186.77

	Word	Frequency <sup>?</sup>		Relative freq. <sup>?</sup>		Score <sup>?</sup>
		Focus	Reference	Focus	Reference	
1	Santa Maria	261	0	217.49	0	218.49
2	Mutter Gottes	115	0	95.829	0	96.83
3	großer Teil	111	0	92.496	0	93.5
4	anderes Ding	106	0	88.329	0	89.33
5	Maria Del	91	13	75.83	0.056	72.75
6	kleine Figur	86	0	71.663	0	72.66
7	ganzes Werk	82	0	68.33	0	69.33
8	ausgezeichneter Maler	75	0	62.497	0	63.5
9	Vasari Giorgio	69	0	57.497	0	58.5
10	Leben der Ausgezeichnetsten Maler	69	0	57.497	0	58.5

Figure 3 and 4: Screenshots of first 10 keywords of the K-LBC (de) (single and multi-words expressions)

As typical items of both corpora we can see LSP-Words (*assise*/base; *verfertigen*/produce), Italian words (*disegno*/drawing; *diligenza*/diligence for German LBC), the name of the city of Florence (*Florence-Florenz*) and the connected adjective (*florentin*/Florentine; *florentinisch*/Florentine), proper names (e.g. Medici, Cosimo, Filippo, Lorenzo, Giotto). But in going down the list we also find obsolete or sophisticated words, such as *brasses*/fathom (obsolete) in K-LBC (fr) and *woselbst*/where (sophisticated) or *heutetags*/today (obsolete) in K-LBC (de). The extracted multiword expressions again show proper names of artists and monuments (Santa Croce; Vasari Giorgio) or possible candidates for collocations (*marbre blanc*/white marble; *grande chapelle*/big chapel; *Mutter Gottes*/Blessed Virgin Mary; *ausgezeichneter Maler*/excellent painter). Differences can be seen in the greater presence of Italian words considered as Keywords in the German Corpus.

2) a lemmatized frequency list (L-LBC). The L-LBC Lists were automatically extracted through the functionality *Word List*<sup>13</sup> of *Sketch Engine*. We decided to set the minimum frequency of  $x > 1$ , since not only the most frequent terms, but also the terms recurring only once (i.e. the *hapax legomena*) could be of interest for our final entry list. Both lists, L-LBC (fr) and L-LBC (de) have in common the fact that articles, conjunctions, prepositions, auxiliary verbs etc. occupy the first positions, while LSP-Terms (*art*/art; *peinture*/picture; *artiste*/artist; *Skulpturensammlung*/sculpture collection; *Marzocco-Löwen*/Marzocco-Lions; *unpoliert*/unpolished) occupy the lowest ones. The same holds true for Italian proper names (Cosimo, Medici, Bargello etc.) and Italian words (among others *palazzo*/building, *loggia*/lodge in the French one and *non-finito*/not finished, *chiesa*/church in the German one). Even if there are many similarities, we also note some differences, such as the greater incidence of French Equivalents of Italian names (Michel-Ange, Médicis, Raphaël).

3) a keyword list based on the reference corpora of the languages involved (K-L-RIF). First, for French a reference list was extracted from *L'Est Républicain* (L-RIF) (fr) by using the *AntConc* software. The list obtained of lemmatized forms arranged by frequency (145.644) was then compared with our L-LBC (fr), calculating their relative frequency. For German we used a slightly different procedure: the K-L-RIF (de) was automatically extracted<sup>14</sup> by comparing our *ad hoc* compiled LBC-Corpus (de) and *DeReKo* with the reference corpus for the German language. *DeReKo* is the world's largest linguistically motivated collection of electronic corpora for German and contains different types of corpora from the present and the recent past, corresponding to different types of texts (including articles from daily newspapers and magazines, literary texts, specialised texts)<sup>15</sup>. The result of the procedure was an excel-list<sup>16</sup> (figure 5), whose ranking can be changed on the basis of two statistical measurements ( $\chi^2$  und LLR, see Dunning 1993).

<sup>13</sup> We chose the option 'Lemma'.

<sup>14</sup> At this point we want to thank the *Leibniz-Institut für Deutsche Sprache* and in particular Rainer Perkuhn for his support. The use of IDS internal tools was fundamental for the comparison between the two German corpora.

<sup>15</sup> See <https://www.ids-mannheim.de/kl/projekte/korpora/> (02.02.2020)

<sup>16</sup> For this procedure see also Flinz/Perkuhn 2018: 962.



		List N.2		Only G-LEX		
rank	dereko-iso	LBC-Lemma-iso	winner	llr	ch2	
1.	6176	10080	LBC-Lemm	185.688,89	181.550.018,58	seine
2.	3172	3951	LBC-Lemm	71.432,76	63.655.268,36	ihre
3.	903	1801	LBC-Lemm	33.575,37	34.842.795,42	ander
4.	7906	3106	LBC-Lemm	50.747,12	25.442.971,33	meine
5.	0	497	LBC-Lemm	10.215,30	14.436.557,14	e
6.	0	348	LBC-Lemm	7.152,71	10.108.494,74	disegno
7.	0	324	LBC-Lemm	6.659,42	9.411.357,17	andern
8.	0	262	LBC-Lemm	5.385,07	7.610.418,45	della
9.	0	262	LBC-Lemm	5.385,07	7.610.418,45	con
10.	0	249	LBC-Lemm	5.117,87	7.232.802,27	diligenza

Figure 5: Screenshot of the K-L-RIF (de) List

A very high degree of association with simultaneously low frequency in *DeReKo* suggests that it can be an artefact in the procedure or even in the primary source (such as for typing errors). High associative measurements (especially LLR) show good candidates for keywords of our corpus. If we consider the first ten positions, there was a particularly striking presence of possessive pronouns (*seine*/his; *ihre*/hers; *meine*/my) and adjectives in the comparative form (*lieber*/nicer; *besser*/better); however, these could be explained by a lemmatization error of the *Sketch engine*. As keywords of our German LBC-Corpus Italian words (*e*/and; *disegno*/drawing; *della*/of; *con*/with; *diligenza*/diligence), proper names (Jacopo; Arezzo; Giovan; Vasari), obsolete spelling variants (*seyn* vs. *sein*/to be; *giebt* vs. *gibt*/gives) and LSP-items (*Bauten*/buildings; *mediceisch*/as Medici etc.) were also signalled.

4) a technical word list from a monolingual lexicographical resource (G-LEX). These lists were extrapolated from central monolingual and bilingual dictionaries (TLFi for French, Duden online and Zanichelli 2009 for German<sup>17</sup>). All entries (1,806 for French and 2,439 for German) were entered into an Excel table.

In a second step, all mentioned lists (K-LBC; L-LBC; L-RIF; G-LEX) were automatically compared and merged<sup>18</sup> with formulas and functions from Excel (including CERCA.VERT):

- From the comparison of the keywords lists K-LBC and the dictionary lists G-LEX two lists resulted (see figure 6):
  - list 1, which comprises all items present both in our corpus and in a lexicographical resource
  - 'ONLY in K-LBC' with items that only appear in LBC. From this list additional technical terms could be identified.

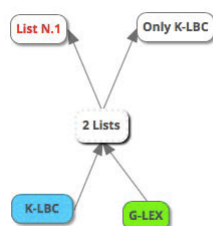


Figure 6: Lists resulting from the merging of K-LBC and G-LEX

- From the comparison of the frequency L-LBC list and the G-LEX dictionary list we obtained two lists (see figure 7):
  - list 2, with items present in both lists
  - 'Only in G-LEX': items which are uniquely present in the lexicographical resource but which are missing from our list

<sup>17</sup> The entry lists of these lexicographical resources were used also as secondary sources in our project.

<sup>18</sup> This procedure was adopted for each language.



Figure 7: Lists resulting from the merging of L-LBC and G-LEX

- From the comparison of the L-LBC frequency list and the reference corpus of each language we obtained two lists (see figure 8):
  - list 3, with items which appear in both lists
  - 'LBC not in L-RIF', with items that occur only in the corpus LBC but not in the reference corpus

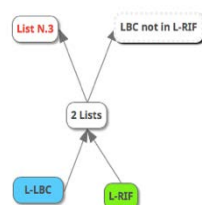


Figure 8: Resulting lists from the merging of L-LBC and L-RIF

From list 3 it was possible to filter out specialized lexemes which also occur in common language; from 'LBC not in L-RIF' all words could be extracted which are not included in the reference corpus of the German language, because they are e.g. archaisms, foreign words, proper names or errors.

All lists were then manually refined thanks to a fine-grained qualitative analysis: for example, all common language terms and other 'irrelevant terms' (such as place names, literary works, religious names etc.) were removed. To create the provisional entry list, Lists 1-3 were merged. Duplicates, and triplicates were deleted; variants of a term (among others *Perugino* – *Pérugin*, *Pisa* – *Pise*, *empattement* – *empatement*, *bastimens* – *bâtiment*; *terre-cuite* - *terracotta* for the French List; *chiaroscuro* – *chiaro-scuro*; *Lorenzo il Magnifico* - *Lorenzo de' Medici* - *Lorenzo Vecchio de' Medici*; *Mahlerei* - *Malerei*; *Piazza della Signoria* - *Piazza der Signoria* - *Piazza der Signore*; *Scultur* - *Skulptur* for the German List etc.) were noted, and the most usual variant was chosen as the lemma.

With this procedure we were able to create the French and German provisional entry lists, which were rechecked and recleaned (table 3):

Language	Provisional entry list
French	1,200 entries
German	1,355 entries

Table 3: Merging of the lists to create the final French and German provisional entry list

Even if the two lists are different for numbers of items (see table 3) we can also see some similarities. Beyond the fact that



we can find mostly LSP-terms, they also have following aspects in common<sup>19</sup>:

1. the presence of many Italian names:

- in both lists: among others Agostino, Ambrogio, Andrea, Angelo, Angelico, Annunziata, Antonio, Arnolfo
- in only one list: among others Agnolo (fr), Alessandro (fr), Alessio (de), Alesso (de), Ambrogio (fr), Annibale (fr), Antonello (fr), Ascanio (fr)

2. the presence of many Italian surnames:

- in both lists: among others Alberti
- in only one list: among others Acciaiuoli/Acciaiuoli (de), Albizzi (fr), Aldobrandini (fr), Allori (fr), Antinori (fr)

3. the presence of many Italian city names, rivers:

- in both lists: among others Arno (in German we find the compound *Arnobücke*/bridge of the Arno)
- in only one list: among others Arezzo (fr)

4. the presence of denomination of Italian institutions and monuments:

- in both lists: among others *Accademia*
- in only one list: among others *Academia* (de)

5. the presence of multiword items, which are both first names and surnames of Italian artists, such as Agostino Chigi (with the variant Agostino Chisi), Alesso Baldovinetti, Andrea di Cione, Angelo Poliziano, Arnolfo di Cambio (with the variants Arnolfo di Lapo, Arnolfo Lapi)

6. the presence of collocations such as: among others *adoration des bergers*/adoration of the Shepherds (fr), *adoration des mages*/adoration of the Magi (fr), *Auferstehung Christi*/ Christ's resurrection (de), *Ausgießung des Heiligen Geistes*/outpouring of the Holy Spirit (de).

The provisional entry lists are the result of the current situation, but will be extended and supplemented in the future. For each entry list we also extracted keywords in context (KWICs), which after a qualitative analysis will be offered to the user (see Billero/Cetro/Farina et al. in prep. for French and Flinz/Ballestracci/Bufagni et al. in prep. for German). In the LBC-Dictionary the extracted KWICs will be useful in different ways: for determining the collocations and usual word combinations of the entries, as examples and as translations of collocations in case of equivalence.

## 4 Conclusions

The lexicographical process of the planned internet LBC dictionary has gotten under way, and the aim of this paper is to reflect on the data collection and the data preparation phase, which, as is usual for internet dictionaries, are open-ended, so what we present in this paper is only a snapshot of the actual situation.

Primary sources of our LBC dictionary are *ad hoc*-created comparable LSP-corpora, which are also freely accessible for other aims, in addition to the lexicographical one presented in this paper (see LBC-Platform). For their construction we used works from major Renaissance authors (see 2.1), both text in original language and translations, because the future aim of this research group is also to set up parallel corpora.

With an alternation of corpus-driven and corpus-based procedures we were able to extract the above discussed provisional entry list for French and German (3.1); the combination and merging of different types of lists and the consequent fine-grained qualitative analysis enable us to focus not only on the most frequent lexemes of our corpora but also on the lesser ones (i.d. *hapax legomena*), not only on the typical terms and multiword expressions according to web corpora (TenTen corpora) but also according to reference corpora. In addition, we also involved existing monolingual and bilingual dictionaries as secondary sources. KWICs of the entries have already been automatically extracted, and after a meticulous qualitative work which aimed at removing all the non LSP-ones will be freely accessible (see Billero/Cetro/Farina et al. in prep. for French and Flinz/Ballestracci/Bufagni et al. in prep. for German). A selection of them will also play a central role in modelling the lexicographical data in the database structure: they will be used for filtering out typical collocations and their examples, but also to reflect on equivalent structures.

## 5 References

- Analyse et traitement informatique de la langue française - UMR 7118 (ATILF), Cognition, Langue, Langages, Ergonomie - UMR 5263 (CLLE-ERSS) (2020). *Corpus journalistique issu de l'Est Républicain* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). [www.ortolang.fr](http://www.ortolang.fr), [https://hdl.handle.net/11403/est\\_republicain/v3](https://hdl.handle.net/11403/est_republicain/v3) [04.05.2020].
- Ballestracci, S., Bufagni, C., Flinz, C. (in prep.). Das deutsche LBC-Korpus: Zusammenstellung und Anwendung. In A. Farina, C. Nicolás Martínez, R. Billero (eds.) *Corpora LBC*. Firenze: Firenze University Press.
- Billero, R., Cetro R., Farina, A. et al. (in prep.). *Lexique français de l'art basé sur le corpus LBC (Lessico dei Beni*

<sup>19</sup> For this article we decided to focus, as examples, only on the lemmas beginning with the letter A.



- Culturali*). Firenze: Firenze University Press.
- Billero, R., Farina, A., Nicolás Martínez, C. (in prep.). *Conclusioni: Dati numerici attuali e bilanciamento dei corpora*. In A. Farina, C. Nicolás Martínez, R. Billero *Corpora LBC*. Firenze: Firenze University Press.
- Billero R., Nicolás Martínez, M.C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. In *CHIMERA Romance Corpora and Linguistic Studies*, Vol.4, No. 2, pp. 203-216.
- Duden online. Accessed at: <https://www.duden.de/woerterbuch> [04.05.2020]
- Dunning, T. (1993). Accurate methods for statistics of surprise and coincidence. In *Computational Linguistics*, 19(1), pp. 61-74.
- Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. 4. überarb. u. erw. Aufl. Tübingen: Stauffenburg.
- Farina, A. (2015). Guideline Proposal for the Description and Translation of Proper Nouns in a Multilingual Cultural Heritage Dictionary of Florence. In O. M. Karpova, Faina I. Kartashkova (eds.), *Life Beyond Dictionaries*, Newcastle: Cambridge Scholars Publishing, pp. 122-132.
- Farina, A. (2016). Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique”. In *PUBLIF@RUM*, vol. <24 [http://publifarum.farum.it/ezine\\_articles.php?id=335](http://publifarum.farum.it/ezine_articles.php?id=335)> [04.05.2020]
- Farina, A. (in prep.). *Le corpus LBC français*. In A. Farina, C. Nicolás Martínez, R. Billero *Corpora LBC*. Firenze: Firenze University Press.
- Farina, A., Billero, R. (2018). Comparaison de corpus de langue « naturelle » et de langue « de traduction » : les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues. In *JADT 2018 - International Conference on Statistical Analysis of Textual Data*, Roma, 12-15 giugno 2018. Rome: UniversItalia, pp. 108-116.
- Flinz, C. (2019). Der lexikographische Prozess bei Tourlex (ein deutsch-italienisches Fachwörterbuch zur Tourismussprache) für italienische DaF-Lerner. In A. Klosa, A. Storrer, J. Taborek (Hrsg.) *Internetlexikographie und Sprachvermittlung. Jahrbuch Lexicographica*. Berlin: de Gruyter, pp. 9-35.
- Flinz, C., Ballestracci, S., Buffagni, C. et al. (in prep.). *Deutsche Lexik der Kunst auf der Basis des Korpus LBC (Lessico dei Beni Culturali)*. Firenze: Firenze University Press.
- Flinz, C., Perkuhn, R. (2018). Wortschatz und Kollokationen in ‚Allgemeine Reisebedingungen‘. Eine intralinguale und interlinguale Studie. In S. Krek, et al. (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Context*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 959-967.
- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: C. Fellbaum (Hg.) *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum, pp. 23-41.
- Geyken, A., Lemnitzer, L. (2016). Automatische Gewinnung von lexikografischen Angaben. In A. Klosa, C. Müller-Spitzer (Hrsg.), *Internetlexikografie: Ein Kompendium*. Berlin/Boston: de Gruyter, pp. 195-241.
- Hunston, S. (2008). Collection strategies and design decisions. In A. Lüdeling, M. Kytö, (Hrsg.) *Corpus Linguistics. An International Handbook*. Volume 1. Berlin/New York: de Gruyter, pp. 154-168.
- Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release vom 08.03.2017). Mannheim: Institut für Deutsche Sprache. PID: 10932/00-0373-23CD-C58F-FF01-3 [04.05.2020].
- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In R. Gouws et al. (Hrsg.) *Dictionaries. An International Encyclopaedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin u.a.: de Gruyter Mouton, pp. 517-524.
- Klosa, A. (2020). The lexicography of German. In P. Hanks, G.M. de Schryver (Hrsg.) *International handbook of modern lexis and lexicography*. Berlin: Springer, pp. 1-21.
- Kupietz, M., Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L.M. Eichinger (Hrsg.) *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven* (Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/Boston: de Gruyter, pp. 297-322.
- Il nuovo dizionario di tedesco. Dizionario tedesco-italiano, italiano-tedesco (2009). Ediz. bilingue. Con CD-ROM (Italiano). Torino: Zanichelli.
- Lemnitzer, L., Zinsmeister, H. (2015). *Korpuslinguistik. Eine Einführung*. 3. Aufl. Tübingen: Narr.
- Lessico dei Beni Culturali*. Accessed at: <http://www.lessicobeniculturali.net> [04.05.2020].
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing.
- Trésor de la Langue française informatise (TLFi)*. Accessed at: <http://atilf.atilf.fr> [04.05.2020].
- Wiegand, H.E. (1998). Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband. Berlin/New York: de Gruyter.
- Zotti, V. (2017). L'integrazione di corpora paralleli di traduzione alla descrizione lessicografica della lingua dell'arte: l'esempio delle traduzioni francesi delle Vite di Vasari. In V. Zotti, A. P. Alamán (cur.) *Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali*. Firenze: Firenze University Press, pp. 105-134.







# To discriminate between discrimination and inclusion: a lexicographer's dilemma

Petersson S., Sköldberg E.

University of Gothenburg, Sweden

## Abstract

The overall theme of this paper is the balance between descriptive adequacy and discrimination in dictionaries. More specifically, the purpose is to describe the process of revising dictionary articles related to the grounds of discrimination in the forthcoming edition of the Swedish monolingual dictionary *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy], which is expected to be published in 2020. The focus of the article is on the semantic fields related to sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation and age. Updates of the lemma list, based on a more diverse data set, are presented. Furthermore, revisions of definitions and linguistic examples, motivated by the new data and principles of inclusion, are shown. We also discuss usage labels of negatively charged words and explore cross-references and their role in facilitating non-discriminatory word choices. Moreover, methodological questions are raised, and the role of corpora and other data gathering methods are considered.

**Keywords:** critical lexicography; Swedish; the Swedish law of discrimination

## 1 Introduction

In the last few decades, there has been a distinct increase of public awareness of the ways language is intertwined with systems of power, and the ways power relations are embedded in everyday language use. In Sweden, the public concern about the relationship between language and power has increased after the implementation of a new law on discrimination in 2009. The law covers seven grounds of discrimination: sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation and age (The Equality Ombudsman 2020).

In this paper, we wish to address lexicographical challenges with respect to words like *hora* ('whore'), *blatte* ('wog'), *rödsinn* ('redskin'), *funktionsvariation* ('functional variation'), *miffo* ('freak'), *bög* ('gay', 'faggot') and *hedersrelaterad* ('related to honour'). Our examples will be drawn from the ongoing update of *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy] (henceforth SO) – a two-volume, mainly corpus-based, definition dictionary consisting of roughly 65,000 words (also available online and as app). The first edition of the dictionary was published in 2009; the updated edition (SO2) is scheduled for publication in 2020.

The dictionaries of the Swedish Academy have a special status among lexicographic resources in Sweden. Since 2017, SO, as well as two other Academy dictionaries, are freely available at a web interface (<https://svenska.se/>), with an increasing number of views. Furthermore, in contrast to most other lexicographic resources of contemporary Swedish, SO is based on a database unique in its scope, developed at a university department (the Department of Swedish, University of Gothenburg). The vocabulary covered by SO, and how this vocabulary is presented, have, undoubtedly, an impact on language policy and planning, which is noted by e.g. Josephson (2018:249-251). Relatedly, the edition from 2009 has several merits, but it is also clear that much can be improved and modernized. Large parts of its contents were developed in the 1980s. The editorial team's judgment is that some of the material ought to be revised, if the aim is to provide a correct description of contemporary Swedish, which will be useful for the intended users (both first and second language speakers) in supporting reception and production (cf. Malmgren 2009). The editors wish to provide a resource that is both descriptively adequate and free from discrimination, and should be understood with these facts in mind.

## 2 Data and Method

In 2009, the editorial team in office at the time worked on lexicographical issues related to gender. The current team, including the authors of this paper, has a somewhat broader perspective. It focuses not only on gender, but also on other areas. The grounds of discrimination mentioned above have the role of semantic fields demarcating the focus of the practical work. There are other fields of controversial words, for example the fields of body shape, political views, social class, etc. (see e.g. Veisbergs 2002 for a discussion of the definitions of political terms; cf. also Hartevelt & van Niekerk 1996), but the grounds of discrimination offer a non-arbitrary method for the identification of problems.

Idealizing to some extent, the following steps of the editorial process can be distinguished. First, a semantic field is chosen, e.g. sexual orientation, ethnicity or some of the other grounds of discrimination. In some cases, the work is made from scratch. Then, the lexicographers pick words from lists of novel word candidates. In other cases, a relevant word is already present in the database. In such situations, the lexicographers may have to reconsider hypotheses and decisions made by earlier members of the editorial team. This sometimes results in revisions of dictionary articles. Differences in judgements can depend on examined data, changes in language use, or choices of theories, studies and guiding principles. The editorial team uses several methods for data gathering, for example the following:



- Corpus searches in Swedish texts, including newspaper articles from various papers and magazines, Wikipedia entries, social media text, etc.<sup>1</sup>
- Google searches
- Discussions within the editorial team
- Discussions with colleagues at seminars, conferences, etc.
- Examinations of informal word lists and dictionaries from relevant organizations, such as the Swedish Federation for Lesbian, Gay, Bisexual, Transgender, Queer, and Intersex rights (RFSL)
- Readings and discussions of relevant studies

SO is to a large extent based on corpus data. An overwhelming part of the corpora used for the 2009 edition (and its predecessors) contained newspaper texts and novels (see Introduction, SO 2009: XII). The data investigated by the current team is more diverse, and contains, among other genres, social media texts. The choice of data has consequences for the lists of novel word candidates, and the uses of words in the data bear on the dictionary descriptions.

### 3 Previous and Related Research

Nunberg (2017) argues that derogatory aspects of words, in particular of so-called “slurs”, depend on the social and historical background of the expressions. Consider *redskin*, a slur used of native Americans. According to Nunberg, the derogatory aspect is not a component of the meaning of the word; but the word has been used, and is still used, within social groups with negative or stereotypical attitudes towards native Americans; furthermore, the history of the word matters for its derogatory aspect. Our treatment of usage labels is in line with this reasoning, as will be shown in what follows.

Moon (2014) focuses on English learners’ dictionaries and issues relevant for non-native speakers of English. Her area of interest is ideologically loaded words. The article investigates themes such as age, sexuality and ethnocentrism. Chen (2019) suggests that the study of dictionaries can be subsumed under the broader research program of critical discourse analysis (see e.g. Fairclough 2010, van Dijk 2015). Dictionaries are sometimes ideologically loaded and reproduce illegitimate power structures in society, Chen argues. Wojahn (2015) discusses language planning and linguistic activism, i.e. attempts to change language with the purpose of changing, for instance, inequality between women and men, between persons with different sexual orientations or between gender identities. In the editorial team’s work on controversial words, a different, perhaps more modest, approach is taken, compared to the outlook of Moon, Chen and Wojahn. Our aim is to provide a discrimination free product, but the editorial team has no (further) ambition of addressing social problems. Our theoretical viewpoint is primarily descriptive. However, some questions concerning the design of the database are not purely descriptive, as will be discussed below.

### 4 Review and Revision

During the editorial work on dictionary articles thematically related to the grounds of discrimination, the lexicographers in the SO project focus on the following information categories and issues:

- Lemma list: What words are included in the lemma list? And what words are not included but should be?
- Meaning descriptions: How are lemmas related to the grounds of discrimination defined? What words are used in the definitions of these words?
- Usage labels: How are labels like “derogatory” and “can be perceived as derogatory” used? How can distinctions between different usage labels be motivated?
- Examples: What linguistic examples (of compounds, phrases and sentences) are presented in the dictionary?
- Cross-references: Which lemmas and meanings are cross-referenced? What lemmas could but should not be linked?

Examples from each of these information categories are presented below.

#### 4.1 Lemma list

A revised lemma list is of course an integral part of the forthcoming revised edition of SO. In addition to the challenges mentioned above, several words pertaining to semantic fields like disabilities, sexual orientation and ethnic background, often possess such an emotive charge that their inclusion, preservation or exclusion equally is the cause of media attention and user indignation.

When it comes to e.g. the semantic field disabilities (both physical or mental), SO (2009) includes several lemmas related to this theme: *adhd* (‘ADHD’), *assistansersättning* (‘assistance compensation’), *funktionshinder* (‘functional disability’), *gruppboende* (‘group home’), *hyperaktivitetssyndrom* (‘hyperactivity syndrome’) etc.

In the dictionary, you also find the lemma *invalid* (‘invalid’, noun) and compound examples including the noun (e.g. *invalidbil* ‘car for invalids’, and *invalidbostad* ‘house for invalids’). These words now have a more negative emotive charge than they had when the first edition of the dictionary was published, and the frequency of them in the corpora is significantly lower today. In the SO2-database, a usage label marking the archaic tone of *invalid* is provided, and the compounds are excluded.

<sup>1</sup> Primarily provided by Språkbanken Text (2020), Mediearkivet (2020) and Kungliga biblioteket (2020).



A new lemma in the next edition will be *funkofobi* ('funcophobia') with the meaning 'prejudice or discrimination against people with functional disabilities'. In 2014 the members of Förbundet Unga Rörelsehindrade ('The association for young persons with disabilities') campaigned for greater recognition and dissemination of this word, which had been created in analogy with words like *homofobi* ('homophobia') and *xenofobi* ('xenophobia'). The campaign was successful: the usage of the word has increased and is now included in several dictionaries.

Furthermore, numerous words related to the fields of sexual orientation and gender identity have been added to the SO2-database, e.g. *cisperson* ('cisgender' or 'cis person'), *hbtq* ('LGBTQ'), *hen* ('they', singular), *heterosexualitet* ('heterosexuality'), *icke-binär* ('non-binary'), *könsbekräftande* ('gender confirming'), *könsdysfori* ('gender dysphoria'), and *polyamöros* ('polyamorous'). By including these lemmas, it could be said that the existence of the sexual orientations etc. that these words denote is acknowledged. At the same time, it should be noted that not only expressions referring to norm-breaking conditions, etc. are added, which *heterosexualitet* ('heterosexuality') exemplifies. It may seem strange that the noun *heterosexualitet* has not been recorded before, but it probably has to do with frequencies in corpora. The word *homosexualitet* ('homosexuality'), which is included in SO (2009), occurs 83,353 times in all Swedish texts in Mediearkivet, the largest digital news archive in the Nordic region. However, the word *heterosexualitet*, which is now listed in the SO2-database, occurs only 2,903 times in corresponding material.

Moreover, it can be noted that the newly established Swedish pronoun *hen*, which has attracted international attention (see e.g. *The Guardian* 2015 and *La Vanguardia* 2015), has developed radically in Swedish since the compilation of the first edition of SO. The pronoun, which is now included in the SO-database, has two meanings, namely 'gender-neutral expression denoting someone referred to, explicitly or implicitly, in the discourse context' and 'expression denoting persons who do not want to, or cannot, unambiguously categorize themselves as men or women'.

## 4.2 Meaning Descriptions

Some definitions from SO (2009) are also revised during the review process. For example, the meaning descriptions of the noun *kön* ('sex' or 'gender') in SO (2009) have been revised. In the forthcoming edition, a new meaning referring to self-perceived identity is recorded, in addition to the already included meaning of biological gender. This identity-related meaning of gender is relevant in lexicalized compounds like *könsidentitet* ('gender identity'), and in phrases such as *psykologiskt kön* ('psychological gender').

Furthermore, some meaning descriptions in the forthcoming edition of SO will be more inclusive than the corresponding ones in the first edition. This is the case with several words describing family relations, e.g. *svärdotter* ('daughter-in-law'). In SO (2009), it is defined as 'the wife of one's son', but in SO2 the definition will be as follows: 'the female partner of one's adult child'. The new definition thereby encompasses both heterosexual and homosexual relationships. Other lemmas redefined according to the same principle are, for example, *svärson* ('son-in-law'), *svärfar* ('father-in-law'), *svärmor* ('mother-in-law'), *man* ('husband'), *make* ('husband'), and *hustru* ('wife').

In connection with this work, the lexicographers have regarded the discussion in Moon (2014), which reports on a study of how different sexual orientations are represented in five major English learning dictionaries. Moon discusses more inclusive meaning descriptions and non-heteronormative examples in the dictionaries (see section 4.4 below). Furthermore, the Swedish version of the user-generated dictionary Wiktionary, which is relatively progressive in this respect, has served as a source of inspiration.

The semantic aspect of the editorial work also concerns words used in definitions of lemmas, and the connotations of these definition words. For example, in the forthcoming edition of SO, the noun *ras* ('race') is replaced with other words or phrases such as *etnisk bakgrund* ('ethnic background') and *hudfärg* ('skin colour') in the definitions of lemmas like *apartheid* ('apartheid'), *eskimå* ('Eskimo'), *halvblod* ('half-breed'), and *mörkhyad* ('dark-skinned') (cf. the discussion of racist terms in dictionaries in Cloete 2013).

## 4.3 Usage Labels

The precise set of usage labels is not an obvious fact. They also tend to be fairly "square", which makes them difficult to apply to words with unstable or context-dependent meanings (see e.g. Norri 2000; Schutz 2002). In the editorial team's recent work, the labels have been carefully examined and the set of usage labels modernized. Hopefully, they are thereby clearer to the dictionary users.

The example *bög* ('faggot' or 'gay')<sup>2</sup> can illustrate our process of developing suitable usage labels. SO (2009) and the current SO2-database provide the following information about the word:

*bög* [...] homosexuell man <ngt vard.> *bögskräck*; *rätten för bögar att gifta sig och adoptera barn* [...]

'gay/faggot [...] homosexual man <somewhat informal> *fear of gays/faggots; the right of gays to get married and adopt children* [...]' (SO 2009)

*bög* [...] homosexuell man <något vardagligt; kan uppfattas som nedsättande> *bögpar*; *rätten för bögar att gifta sig och adoptera barn* [...]

'gay/faggot [...] homosexual man <somewhat informal, can be perceived as derogatory> *gay couple; the right of gays to get married and adopt children* [...]' (SO2-database).

As is seen above, the usage label 'somewhat informal' is attached to the noun *bög* in SO (2009). This is a comment about

<sup>2</sup> The emotive charge of the Swedish word is context dependent; it is, therefore, difficult to translate.



style: the word is marked for register and is not appropriate in more formal genres. In SO (2009) we find, furthermore, a more elaborated style comment, according to which *bög* is a reclaimed word, which previously was derogatory. It is now a perfectly neutral expression, according to the comment.

However, we disagree with the style comment in SO (2009). If a homosexual man uses the word about himself, it is probably not derogatory but neutral. But the word is still used in contexts, especially among boys and young men, where it is possible, even likely, that the word is associated with homophobic attitudes. Following the reasoning in Nunberg (2017), the word can be said to be used in social groups with negative attitudes towards homosexuality. Native and non-native dictionary users should be informed of this fact. Therefore, the style comment has been removed, and the usage label ‘can be perceived as derogatory’ has been added.

Interestingly, the current design of the dictionary article differs from the recommendations of the Swedish Federation for Lesbian, Gay, Bisexual, Transgender, Queer, and Intersex rights (RFSL). In accordance with SO (2009), RFSL holds, in their informal online dictionary, that *bög* is a reclaimed, neutral word.<sup>3</sup> The difference between our judgment and RFSL’s probably depends on viewpoint and aim. RFSL’s approach is arguably linguistic activism, in contrast to our more descriptive outlook.

Furthermore, we have changed one of the language examples in the lexical entry *bög*. Instead of the mildly homophobic *bögskräck* (‘fear of gays/faggots’), we have chosen the expression *bögpar* (‘gay couple’), which, in addition, is more frequent in our corpora.

It is debatable whether the usage label ‘can be perceived as derogatory’ is the best option. The label is written from the perspective of the hearer. One could argue that the relevant perspective is the speaker’s: words can be said with different intentions. However, we have chosen to keep the listener’s perspective, for taxonomic reasons. A revision of the taxonomy would be too time-consuming. Moreover, there are theoretical reasons for focusing on other aspects than speaker intentions. Nunberg (2017) argues, as we saw above, that the social or historical background is what matters most for a word’s negative or positive charge; given that account, speaker intentions do not seem to be the only central factor. It can, finally, be discussed whether *bög* has one main meaning, and an emotive charge depending on context, or, alternatively, if the neutral and the negative use constitute two different meanings. Given our taxonomy and standard design, it is more parsimonious to start from the descriptive meaning, rather than the emotive charge. But admittedly, it is difficult to see any strong theoretical reason for or against the alternatives. One can, therefore, put less weight on the descriptive definition, and let the emotive charge motivate two meanings, if one prefers that kind of design of dictionary articles.

#### 4.4 Examples

In some dictionaries, examples illustrating the use of derogatory words have been suppressed (see e.g. Harteveld & van Niekerk 1996). Within the SO project, no such general decision has been made. However, compounds, phrases and sentences included in SO (2009) have been the subject of a detailed examination and many examples have been replaced. Since SO is rather a corpus-influenced dictionary than a strictly corpus-based resource, the editorial team puts a lot of effort into adapting the examples to the dictionary.

Nikula (2008) has examined how age is represented in Swedish dictionaries. She observes the following:

Today elderly people in many ways radically differ from those of former generations. One consequence of this is that terms like *pensioner* (Sw. *pensionär*) with their connotations are often not felt to be adequate any more. The same concerns the stereotype of old people as generally poor, ill and disabled. The lexicographic examples in the entries of Swedish monolingual dictionaries to an astonishingly great extent repeat this stereotype. (Nikula 2008: 337).

Nikula questions whether the one-sided description of the elderly can be defended, from an ethical point of view (see also Moon 2014 for a discussion on ageism in English dictionaries).

In connection with this semantic field, the adjective *gammal* (‘old’) can be mentioned. There are several well-established Swedish word pairs including this particular adjective, e.g. *gammal och grå* (‘old and grey’), *gammal och trött* (‘old and tired’) and *gammal och ful* (‘old and ugly’), i.e. phrases with negative connotations, painting a grey and depressing picture of older people and of ageing. However, the word combinations are lexicalized in Swedish; therefore, they are included in the dictionary. This fact does not prevent the SO2-lexicographers from including some new examples under headwords like *pigg* (‘healthy, alert, lively’), e.g. the collocation *pigg pensionär* (‘active pensioner’) and the lexicalized expression *vara pigg för sin ålder* (‘be nimble for one’s age’).

In accordance with the previous reasoning, the lexicographers of the project also aim to provide a more multifaceted and varied picture of sexual orientations and related fields. For example, consider the lemma *sexualitet* (‘sexuality’). In SO (2009), the article contains two compounds illustrating the usage of the word: *heterosexualitet* (‘heterosexuality’) and *homosexualitet* (‘homosexuality’). In the SO2-database, these two compounds have been supplemented with the following expressions: *asexualitet* (‘asexuality’), *bisexualitet* (‘bisexuality’) and *hypersexualitet* (‘hypersexuality’) (these words are also lemmas in the current database).

In the next edition of SO, several phrases and sentences describing homosexual relationships will also be given (cf. Moon 2014). For example, in the dictionary article *blivande* (‘future’) the users will find the example *hon träffande sin blivande hustru i USA* (‘she met her future wife in the United States’) and under the idiom *lära känna någon/något* (‘get to know somebody/something’) (in the article *lära*), they will find *han lärde känna sin blivande make i Paris* (‘he got to know his

<sup>3</sup> See <https://www.rfsi.se/hbtqi-fakta/begreppsordlista/> [accessed 24/04/2020].



future husband in Paris’).

However, it should be noticed that examples of this kind are few in number, relatively speaking. They mainly appear in dictionary articles where the definitions have been modified (cf. the discussion of the new definitions of e.g. the lemmas *man* ‘husband’, *make* ‘husband’ and *hustru* ‘wife’ in section 4.2 above). It is much more common to incorporate gender-neutral phrases and sentences like *han har förlovat sig med sin nya kärlek* (‘he is engaged with his “new love”’). In this example, the gender of the “new love” is unknown. Another relevant example is *flytta ihop med sin flickvän* (‘move in with one’s girlfriend’) under the lemma *flytta* (‘move in’). This latter type of shorter example is common in the SO2-database and is in line with the examples already included in SO (2009).

#### 4.5 Cross-References

Finally, the SO editors have updated the guidelines with regards to cross-references between different dictionary articles. In the project, the lexicographers aim to give cross-references from derogatory words to more neutral synonyms etc., but not the other way around (cf. e.g. Coffey 2010). However, it is debatable, if this strategy is compatible with the dictionary’s primarily descriptive approach and its aim to give an exhaustive description of the lexical relations between Swedish words.

The strategy has been applied to several examples in the semantic field of ethnic background. In SO (2009) there is a cross-reference from the neutral *same* (‘Sami’) to the derogatory *lapp* (‘Lapp’). In the SO2-database, this reference is deleted. However, there is a link from *lapp* to *same*. In the case of the neutral noun *rom* (‘Rom, Romani’), the more controversial expression *zigenare* (‘gypsy’), is referred to, in SO (2009); in SO2, there is no link in that direction, but the cross-reference from *zigenare* to *rom* is still there.

The dictionary user is thereby offered an alternative form of expression, if she searches for *lapp* and *zigenare* (see Malmgren 2009 for a discussion about information in SO that supports production). By offering a cross-reference in one direction, from negatively charged alternatives to neutral ones, but not in the other direction, the user is guided from derogatory formulations to the expressions that the ethnic groups referred to prefer. This guidance is motivated by our aim to provide a discrimination free product, rather than the ambition to develop a resource that describes Swedish objectively.<sup>4</sup>

#### 5 Final Remarks

This paper reports on ongoing work on words related to different grounds of discrimination in the development of a new edition of the Swedish monolingual dictionary *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy] (SO), which is expected to be published in 2020. Words related to semantic fields associated with the grounds of discrimination are, in relation to the dictionary’s list of 65,000 lemmas, few in number. However, the words are important in the social debate, and many of them also have a symbolic value to dictionary users. As was mentioned at the beginning of the paper, the overall goal of the work is to make the content of the dictionary more inclusive and more descriptively adequate. The lexicographers in the project aim to develop a product that is as discrimination free as possible, and to construct a resource that correctly describes the Swedish language.

In the paper, we have presented various aspects of the work. The lemma list is revised. And the lexicographers are making efforts to use as neutral words as possible in the meaning descriptions and to make the definitions more inclusive. Furthermore, the set of usage labels is reviewed. The examples from the first edition are examined and new ones, including ones based on norm-breaking relationships, have been added. Finally, the guidelines for cross-references have been updated: the editors add references from negatively charged words to neutral ones, but not in the opposite direction. The decisions made in the process are sometimes difficult, and balanced judgments that, on the one hand, respect the Swedish law of discrimination and, on the other hand, live up to demands of descriptive adequacy, are not always easy to attain.

One important issue, which is only partially addressed in the article, is how the lexicographers find words that are – or can be perceived as – discriminatory in the database they are revising. Certain lemmas are, through usage labels etc. noted beforehand, but language is developing rapidly and a word that has previously been relatively neutral can quickly become controversial (for example through influence of other languages).

Another central question is how negatively charged (or biased) words are identified. The lexicographers may use sources such as e.g. RFSL’s glossary, but informal dictionaries do not always fit the purpose of the professional lexicographer. Here, appeals to linguistic intuitions may emerge as central, but such methods are not free from risks. Another question is how lexicographers become aware of the stereotypes that can be found in other seemingly harmless articles (cf. Moon 2014). These reflections, or developments of them, can perhaps serve as research questions for future lexicographic work, by our team or others interested in balancing between descriptive adequacy and discrimination.

#### 6 References

- Chen, W. (2019). Towards a discourse approach to critical lexicography. In *International Journal of Lexicography* 32(3), pp. 362-388.
- Cloete, A. E. (2013). The treatment of sensitive items in dictionaries. In Rufus Gouws & Franz J. Hausmann (eds.) *Dictionaries. An International Encyclopedia of Lexicography Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: de Gruyter Mouton, pp. 482-486.

<sup>4</sup> It is instructive to compare with the Swedish version of Wiktionary, where derogatory synonyms are shown in the articles of *same* and *rom*.



- Coffey, S. (2010). 'Offensive' items, and less offensive alternatives, in English monolingual learners' dictionaries. In Anne Dykstra & Tanneke Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress*. Leeuwarden, pp. 1270-1281.
- van Dijk, T. A. (2015). Critical discourse analysis. In Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds.) *Handbook of Critical Discourse Analysis*. Second edition. Vol. I. Malden, Massachusetts: Wiley Blackwell, pp. 466-485.
- The Equality Ombudsman (2020). Protected grounds of discrimination. Accessed at: <https://www.do.se/other-languages/english/protected-grounds-of-discrimination/> [24/04/2020].
- Fairclough, Norman (2010). *Critical Discourse Analysis. The Critical Study of Language*. New York: Routledge.
- The Guardian* 2015. Sweden adds gender-neutral pronoun to dictionary. Published: 23/3/2015. Accessed at: <https://www.theguardian.com/world/2015/mar/24/sweden-adds-gender-neutral-pronoun-to-dictionary> [24/04/2020].
- Harteveld, P. & A. E. van Niekerk (1996). Policy for the Treatment of Insulting and Sensitive Lexical Items in the Woordboek van die Afrikaanse Taal. In Martin Gellerstam et al. (eds.) *Proceedings of the VII EURALEX International Congress*. Göteborg, pp. 390-402.
- Josephson, O. (2018). *Språkpolitik*. Stockholm: Morfem.
- Kungliga biblioteket (2020) = Kungliga biblioteket. Sök bland svenska dagstidningar. Accessed at: <https://tidningar.kb.se/> [24/04/2020].
- La Vanguardia* (2015). Suecia oficializa el pronombre neutro. Published: 28/3/2015. Accessed at: <https://www.lavanguardia.com/vida/20150328/54429302071/suecia-oficializa-pronombre-neutro.html> [24/04/2020].
- Malmgren, S.-G. (2009). On production-oriented information in Swedish monolingual defining dictionaries. In Sandro Nielsen & Sven Tarp (eds.) *Lexicography in the 21st Century: In honour of Henning Bergenholtz*. (Terminology and Lexicography Research and Practice 12.) Amsterdam/Philadelphia: John Benjamins, pp. 93-102.
- Mediearkivet 2020 = Mediearkivet. Accessed at: <http://web.retriever-info.com.ezproxy.ub.gu.se/services/archive> [24/04/2020].
- Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. In Andrea Abel, Chiara Vettori & Natascia Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano, pp. 85-105.
- Nikula, K. (2008). Pensionär – aldrig i livet. De äldre i ordböckerna. In Ásta Svavarsdóttir, et al. (eds.) *Nordiska Studier i Lexikografi 9*. Akureyri, pp. 337-352.
- Norri, J. (2000). Labelling of derogatory words in some British and American dictionaries. In *International Journal of Lexicography* 13(2), pp. 71-106.
- Nunberg, G. (2017). The Social Life of Slurs. In Daniel Fogal, Daniel W. Harris & Matt Moss (eds.) *New Work on Speech Acts*. Oxford: Oxford University Press, pp. 238-291.
- Schutz, R. (2002). Indirect Offensive Language in Dictionaries. In Anna Braasch & Claus Povlsen (eds.) *Proceedings of the X EURALEX International Congress*. Copenhagen, pp. 637-641.
- SO = *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy], 2009. Accessed at: <https://svenska.se/> [24/04/2020].
- Språkbanken Text (2020) = Språkbanken Text. Accessed at: <https://spraakbanken.gu.se/> [24/04/2020].
- Svenska.se = Svenska Akademiens ordboksportal. Accessed at: <https://svenska.se/> [24/04/2020].
- Veisbergs, A. (2002). Defining Political Terms in Lexicography: Recent Past and Present. In Anna Braasch & Claus Povlsen (eds.) *Proceedings of the X EURALEX International Congress*. Copenhagen, pp. 657-667.
- Wiktionary = *Wiktionary, den fria ordboken*. Accessed at: <https://sv.wiktionary.org/> [24/04/2020].
- Wojahn, D. (2015). *Språkaktivism. Diskussioner om feministiska språkförändringar i Sverige från 1960-talet till 2015*. (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 92.) Uppsala.



# The MorfFlex Dictionary of Czech as a Source of Linguistic Data

Štěpánková B., Mikulová M., Hajič J.

Charles University, Prague, Czech Republic

## Abstract

In this paper we describe MorfFlex, the Morphological Dictionary of Czech, as an invaluable resource for exploring the formal behavior of words. We demonstrate that MorfFlex provides valuable and rich data allowing to elaborate on various morphological issues in depth, which is also connected with the fact that the MorfFlex dictionary includes words throughout the whole vocabulary range, including non-standard units, proper nouns, abbreviations, etc. Moreover, in comparison with typical monolingual dictionaries of Czech, MorfFlex also captures non-standard wordforms, which is very important for Czech as a language with a rich inflection. In the paper we also demonstrate how particular information on lemmas and wordforms (e.g. variants, homonymy, style information) is marked and structured. The dictionary is provided as a digital open access source available to all scholars via the LINDAT/CLARIAH-CZ language resource repository. It is available in an electronic format, and also in a more human-readable, browsable and partly searchable form.

**Keywords:** Morphology; Dictionary; Czech; Lemma; Wordform; Tag

## 1 The Morphological Dictionary of Czech: MorfFlex

MorfFlex (Hajič et al. 2020a) is a dictionary of Czech wordforms with detailed morphological information (see a more detailed description in Hajič 2004, Hlaváčová et al. 2019, Mikulová et al. 2020). The MorfFlex dictionary represents more than 100 million wordforms and more than 1 million lemmas. It has been developed gradually since 1988 and the latest electronic version will be published in 2020 (Hajič et al. 2020a). It is also available in a more human-readable, browsable and partly searchable form via online MorphoDita tool.<sup>1</sup>

MorfFlex has the following goals, namely providing:

- a basis for consistent morphological annotation of the Prague Dependency Treebanks (see more Hajič et al. 2017, Hajič et al. 2020b) which serve as a training data for various NLP tasks (tagging and lemmatization, cf. the tool Morphodita (Straková et al. 2014));
- a basis for tagging and lemmatization of other synchronic corpora of Czech, e.g. the Czech National Corpus (Hnátková et al. 2011) and the web corpora Araneum (Benko 2014);
- a resource for linguistically-oriented research, particularly for describing morphological characteristics of Czech.

In this paper, we concentrate on the last aspect.

## 2 MorfFlex as a Resource for Linguistic Research

In typical monolingual dictionaries, which focus on the meaning of lexical units, there is usually a brief morphological description followed by the definition of the meaning and examples for each word. In current Czech dictionaries (e.g. *Slovník spisovného jazyka českého*, *Slovník spisovné češtiny*) the morphological description contains the part of speech, gender (for nouns), grammatical aspect (for verbs), and typically one or more supporting inflectional suffixes (endings), mostly genitive singular ending for nouns and 1st singular present and 2nd singular imperative endings for verbs, which help to identify, for a speaker of Czech, the complete inflectional paradigm. The macrostructure as well as the microstructure of MorfFlex is completely different, since it specializes in morphology. MorfFlex is not a dictionary of words, but of wordforms. Each entry is represented by a triple composed of a wordform, lemma, and tag. Wordforms are organized into paradigms according to their formal morphological behavior. The paradigm is identified by a unique lemma. For each wordform, full inflectional information is encoded in a tag. We can search the dictionary by lemma or by wordform.

The morphological system plays an important role in Czech, which is, like other Slavic languages, highly inflected. As the results from the European Survey of Dictionary Use and Culture (cf. Kosem et al. 2019, the Czech Republic local dataset) confirm, grammatical information is one of the pieces of information most searched for by the users of Czech monolingual dictionaries.

This fact probably reflects the Czech language situation. Bermel (2000) describes it as quasi-diglossic, i.e. a situation which is characterized by the existence of two varieties used by a single language community.<sup>2</sup> Besides the main variety representing Standard Czech, the other variety is also significant for Czech – it is a non-standard variety, which covers

<sup>1</sup> <http://lindat.mff.cuni.cz/services/morphodita/run.php>

<sup>2</sup> Bermel follows Ferguson's description of diglossia (1959), specifying the term quasi-diglossia: "In contrast to classic diglossic system, the high and low codes are mutually comprehensible." (Bermel 2007: 51).



most of the Czech language area, mainly of Bohemia.<sup>3</sup> This variant is used mainly in spoken informal communication, the so called *obecná čeština* (Common Czech, sometimes also *Colloquial Czech*; cf. Hoffmannová 2013). This variety is present in both the lexicon (lemmas) and the morphology (wordforms), e.g. most adjectives have a complete paradigm of Common Czech endings (cf. also examples 1-4 below). While Czech monolingual dictionaries are traditionally focused on description of standard Czech, MorfFlex captures the morphology (and to a lesser extent the lexicon) of Standard Czech, Common Czech and to some extent some dialects.

- (1) lemma: *kývat* (standard) vs. *kejvat* (Common Czech) [infinitive: 'to sway']
- (2) lemma: *okno* (standard) vs. *vokno* (Common Czech) [nom. sg. neuter: 'window']
- (3) wordform: *kývají* (standard) vs. *kejvají* (Common Czech) [3rd pl. present: 'sway']<sup>4</sup>
- (4) wordform: *mladých* (standard) vs. *mladejch* (Common Czech) [loc. pl. adj.: 'young']

### 3 Generation: From Lemma to its Wordforms

MorfFlex covers all possible types of words (tokens) that occur in real Czech texts, i.e. Czech words, loan words, foreign words, proper nouns, abbreviations, parts of words, and numerals. For each lemma, the following information is captured:

- paradigm
- stylistic characteristics
- homonymy
- semantic labels
- derivative relation

Unlike paradigms which are represented by a set of wordforms and tags, the labels and indexes are not a part of the tag but refer to the whole lemma. Note that in MorfFlex, neither the semantic label, nor the stylistic characteristic, nor the homonymy indexing are transposed from any Czech monolingual dictionary, but they are provided by manual annotation (cf. Hajič et al. 2020b).

#### 3.1 Paradigms Comprising all Wordforms, including Non-standard Ones

MorfFlex captures both the singular and the plural set of wordforms of all inflected words, even of proper nouns. As mentioned above, MorfFlex is not only focused on Standard Czech, therefore the paradigms also provide non-standard variants and capture the stylistic characteristics of wordforms. The distinction between standard and non-standard wordforms is captured by a number on the last, 15th position in the tag (see Sect. 4.1.3 for another use of the 15th position in the tag). No value (in the tag) indicates the primary wordform, numbers 1-5 mark standard variants, and numbers 6-9 non-standard variants. See Table 1, where variant wordforms of the instrumental plural of the name *Thales* are shown; the first two variants belong to the standard variety, the last wordform represents the non-standard one.

Thalesi	Thales_Y	NNMP7-----A----
Thalety	Thales_Y	NNMP7-----A---1
Thalesema	Thales_Y	NNMP7-----A---6

Table 1: Standard and non-standard variants.

#### 3.2 Homonymy of Lemmas

Unlike typical monolingual dictionaries, MorfFlex does not capture any differences in meanings of homonymous words;<sup>5</sup> it however distinguishes lemmas with the same spelling but different formal morphological behavior. Each homonymous lemma is marked by an index, e.g. *drát-1* (the noun 'wire'), *drát-2* (the verb 'to pluck'). In some cases, however, essential syntactic characteristics are taken into account: for example, homonymous forms of uninflected words are considered to be different, and therefore represented by two lemmas, e.g. *přece* which is in accordance to its behavior/function in a sentence interpreted as a conjunction 'despite' (lemma *přece-1*) or as a particle 'after all' (lemma: *přece-2*), and it has two lemmas with different indexes and different tags in the dictionary.

#### 3.3 Stylistic Characteristics

Although MorfFlex is primarily focused on morphology, in certain cases the stylistic characterization of a word (lemma) is also provided. This information is consistently marked for variants which have the same declension or conjugation but different stylistic characteristics. In such cases, one lemma is selected as the basic one and the others are

<sup>3</sup> Sometimes this variety is classified as an interdialect (Šipková 2017).

<sup>4</sup> Often various combinations of standard and Common Czech are possible, e.g. *kývají*, *kývají*, *kejvají*, *kejvají*.

<sup>5</sup> Thus, in MorfFlex we do not distinguish e.g. the feminine noun *matka* ('mother') from *matka* ('nut'), even though the former is animate and specific possessive forms can be derived.



marked and linked to it. Still, the rule applies that meanings of words are not taken into account. We use a set of labels for distinguishing standard and non-standard variants. (See Table 2.) In examples 5-7 several variant lemmas are mentioned and the way they are linked is shown: two standard variants of the name *Thalés* and *Thales*, the noun *býk* 'bull' and its non-standard, colloquial variant *bejk*, and the noun *večer* 'evening' and its dialect variant *večír*.

(5) *Thalés*\_:Y\_,s\_^(^DD\*\**Thales*) → *Thales*\_:Y

(6) *bejk*\_,h\_^(^GC\*\**býk*) → *býk*

(7) *večír*-l\_,n\_^(^GC\*\**večer*-l) → *večer*-l

Stylistic characteristics		Label
standard	literary	s
	archaic	a
non-standard	dialectal	n
	non-standard, Common Czech	h
	expressive	e
	slang, argot, cant	l
	offensive, vulgar	v

Table 2: List of stylistic labels.

### 3.4 Semantic Labels

Some nouns are also marked by the so-called *semantic label* (see Table 3), i.e. a label (or more labels) classifying them into a particular semantic group. (See Table 3 for the list of the semantic labels.) It is primarily used for nouns starting with a capital letter, both Czech ones and those integrated into the Czech morphological system. For example, in Table 1, the label Y following the lemma *Thales* indicates a person name. Semantic labels help to tell homonymous words apart – e.g., they distinguish the animateness of nouns, as e.g. in *McIntosh*-l\_:Y the Y refers to the animate behavior of the word, and G and m in *McIntosh*-2\_:G\_:m to its inanimate behavior. Semantic labels also serve to verify that the first capital letter is used properly, i.e. all noun lemmas starting with a capital letter have to be marked by a semantic label.

Code	Definition
Y	person names (given, family, etc.)
E	nationalities, citizen, ethnic, and other named groups
G	geographical names of any kind
m	product, organization, company and other proper names
U	medical, chemistry and natural science terms

Table 3: List of semantic labels.

### 3.5 Derivative Relations

The word-formation relations in Czech has been delegated to derivational data sources, such as Derinet (Vidra et al. 2019).<sup>6</sup> In MorfFlex, the lemma contains information about the base lemma it is derived from only in case of regular derivations. For example, lemmas of possessive adjectives (e.g. lemma: *otcův*\_^(\*3ec)) contain information about the noun they are derived from: *otcův* 'father's' → *otec* 'father'. The originating lemma is (for space saving reasons only) written in the form of a rule. For example, derivation information \*3ec in lemma *otcův*\_^(\*3ec) means remove 3 characters, add *ec* to get *otec*.

<sup>6</sup> <https://ufal.mff.cuni.cz/derinet>



## 4 Analysis: From Wordform to its Detailed Morphological Description

For each wordform, the structured morphological information is captured in its tag. Each position of the tag captures a different aspect of the wordform, e.g. its case or number, the style of the inflectional variant, or a characteristic of the lemma equivalent to information given in typical monolingual dictionaries, e.g. part of speech, detailed features of the particular part of speech (such as possessivity of pronouns, verbal aspect, animateness of nouns etc.).

### 4.1 Special Parts of Speech

The values on the individual tag positions reflect the morphological focus of the dictionary. Therefore, in addition to the standard POS set, several special categories are used (cf. Hlaváčová et al. 2019; Mikulová et al. 2020): foreign word, segment, abbreviation, and isolated letter. The new categories of POS allow to describe the diversity of the language much more precisely.

#### 4.1.1 Foreign word

Foreign word (F at the POS tag position) identifies a word that is not subject to the Czech inflectional system, often creates a part of a longer foreign phrase in a Czech text and has no meaning of its own in Czech.

#### 4.1.2 Segment

Segment (S at the POS tag position) describes a part of a word that creates a complete meaningful unit only when joined with another component. In MorFlex, we distinguish two types of segments. First, the so called *prefixal segments*, which stand at the beginning of a word and which express no morphological categories. Secondly, the so called *postfixal segments*, which may express all morphological categories of a particular part of speech. Table 4 shows the analysis of the segments of the tokenized compound adjective *tchaj-pejský* ('*Taipei[s]*'). The segment *pejský* behaves as an adjective and expresses the morphological features of the compound adjective.

Wordform	Lemma	Tag
tchaj	tchaj	S2-----A----
pejský	pejský_^(tchaj-pejský)	SAMS1----1A----

Table 4: Segments.

#### 4.1.3 Abbreviation

Abbreviations (B at the POS tag position) composed of capital letters and representing a multi-word unit (e.g. *ČR* for *Česká republika* 'the Czech Republic') are considered to be separate parts of speech. In contrast, the abbreviations abbreviating a one-word term are captured as a special wordform in the paradigm of the term, i.e., by the letter *b* at the 15th position of the tag (e.g. *čt.* for *čtvrtek* 'Thursday'). See examples in Table 5.

Wordform	Lemma	Tag
USA	USA_;G_^(United_States_of_America)	BNXXX-----A----
ČT	ČT_;m_^(Česká_televize)	BNXXX-----A----
čt	čtvrtek	NNIXX-----A---b
m	minuta	NNFXX-----A---b

Table 5: Abbreviations.

#### 4.1.4 Isolated letter

Isolated letters (Q at the POS tag position) stand for many meanings but it is not clear for which of the many alternatives. We do not distinguish between an abbreviation (e.g. *A. Franklin*) and a label (e.g. *skupina A* 'group A', *A-konto* 'A-account') and between the other meanings such as for sorting a list (e.g. *a), b)*) or as a graphical separator in a text (e.g. *o o o o o o o*). See examples in Table 6. The introduction of new POS for isolated letters does not mean that standard POS such as conjunctions or prepositions are not distinguished for one-letter word (e.g. letter *a* in *otec a matka* 'father and mother' is considered a conjunction POS).



Wordform	Lemma	Tag
(skupina) A	A-33	Q3-----
A. (Franklin)	A-33	Q3-----

Table 6: Isolated letters.

## 4.2 Homonymy of Wordforms

Searching by wordform also helps in the recognition and interpretation of homonymous forms, both within the lemma and across the whole dictionary. As it is evident from Table 7, the wordform *sil* is analyzed as the genitive plural form of two different nouns, *silo* 'silo' and *síla* 'power', as the masculine singular past participle of the verb *sít* 'to plant', and as two forms of the imperative of the verb *sílit* 'to strengthen'.

Wordform	Lemma	Tag
sil	sít_^(zasévat [semena,...])	VpYS----R-AAI--
sil	silo_^(pro úschovu např. krmiva; raket)	NNNP2-----A----
sil	síla_^(fyzická, vojenská; moc)	NNFP2-----A----
sil	sílit_^(získávat sílu)	Vi-S---2--A-I--
sil	sílit_^(získávat sílu)	Vi-S---3--A-I-4

Table 7: Homonymy of wordforms.

## 5 Conclusion

We have demonstrated the possibilities of the exploitation of the MorfFlex dictionary for linguistic research purposes. In contrast to other Czech monolingual dictionaries or grammar handbooks, MorfFlex contains not only much more morphological data, described in detail (as expected), but it also covers a wider range of words, including non-standard wordforms and lemmas. Furthermore, the dictionary is extended by adding semantic labels and stylistic labels and other complementary tools, which firstly serve to specify and clarify morphological data, and secondly to simplify the orientation in the dictionary for users. Although the dictionary also provides some wordforms that are only potential (i.e. unattested), due to a significant proportion of manual annotation and consequent unification the results are relatively reliable, therefore the dictionary could serve as a resource for diverse linguistic research, and as a morphological support for the creation of other dictionaries.

## 6 References

- Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds): *TSD 2014*, LNAI 8655. Springer International Publishing, pp. 257–264.
- Bermel, N. (2000). *Register Variation and Language Standards in Czech*. Studies in Slavic Linguistics, 13. Muenchen: LINCOM EUROPA.
- Bermel, N. (2007). *Linguistic authority, language ideology, and metaphor: the Czech orthography wars*. Berlin-New York: Mouton de Gruyter.
- Ferguson, Ch. A. (1959). Diglossia. *Word*, 15, pp. 325–340.
- Hajič, J. (2004). *Disambiguation of Rich Inflection. (Computational Morphology of Czech)*. Prague: Karolinum.
- Hajič, J., Hajičová, E., Mikulová, M., Mirovský, J. (2017). Prague Dependency Treebank. In *Handbook on Linguistic Annotation*. Dordrecht: SpringerVerlag, pp. 555–594.
- Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánková, B. (2020a, in press). *MorfFlex CZ*. Institute of Formal and Applied Linguistics, LINDAT/CLARIAH-CZ, Charles University, Prague, Czech Republic, LINDAT/CLARIAH-CZ PID: <http://hdl.handle.net/11234/1-3186>.
- Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., Štěpánková, B. (2020b). Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 5208–5218.
- Hlaváčová, J., Mikulová, M., Štěpánková, B., Hajič, J. (2019). Modifications of the Czech morphological dictionary for consistent corpus annotation. *Jazykovedný časopis/Journal of Linguistics*, 70 (2), pp. 380–389.
- Hoffmannová, J. (2013). Česká hovorovost a hovorová čeština (v kontextu dalších slovanských jazyků). *Slavia* 82, pp. 125–136.
- Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S., Dorn, A. et al. (2019). The image of the



- monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32 (1), pp. 92-114.
- Mikulová, M., Hlaváčová, J., Hajič, J., Hana, J., Hanová, H., Hladká, B., Štěpánková, B., Zeman, D. (2020). *Manual for morphological annotation, Revision for the Prague Dependency Treebank - Consolidated 1.0*. Technical Report TR-2020-64, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic. In press.
- Skoumalová, H., Hnátková, M., Petkevič, V. (2011). Linguistic Annotation of Corpora in the Czech National Corpus. In Zacharov, V.: *Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2011" (Proceedings of the International Conference "Corpus Linguistics – 2011")*. St.-Petersburg State University, Institute of Linguistic Studies, Sankt-Petěrburg, Russian State Herzen Pedagogica, pp. 15-20.
- Slovník spisovné češtiny*. (1978) (Second, revised edition 1994; third, revised edition 2003). Prague: Academia.
- Slovník spisovného jazyka českého*. (1960–1971). Prague: Academia.
- Straková, J., Straka, M., Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*. Association for Computational Linguistics, Baltimore, pp. 13-18.
- Šipková, M. (2017). Interdialekt. In P. Karlík, M. Nekula, J. Pleskalová (eds.), *CzechEncy - Nový encyklopedický slovník češtiny*. <https://www.czechency.org/slovník/INTERDIALEKT> [20/05/2020].
- Vidra, J., Žabokrtský, Z., Ševčíková, M., Kyjánek, L. (2019). DeriNet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic, pp. 81–89.

### Acknowledgements

The research and language resource work reported in the paper has been supported by the LINDAT/CLARIAH-CZ projects funded by Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).



# Announcing the Dictionary: Front Matter in the Three Editions of Furetière's *Dictionnaire Universel*

Williams G.<sup>1</sup>, Galleron I.<sup>2</sup>, Stincone C.<sup>2</sup>

<sup>1</sup> Grenoble-Alpes University, France

<sup>2</sup> Paris 3 – Sorbonne Nouvelle University, France

## Abstract

The front matter of a dictionary provides important information as to the background to a work and what is to be expected inside. Although they can be read as standalone texts, it is only when linked to the actual dictionary content that their full potential is realised. This is very much the case for the prefaces to Furetière's *Dictionnaire Universel*, first published posthumously in 1690 and then to go through two major revisions in 1701 and 1725/27. As Furetière left no preface, we start with his *factums*, texts that details his fight with the *Académie Française* who wanted to impede publication. We then have the preface by Bayle of 1690 and then the front matter produced by the two revisers, Basnage de Beauval and Brutel de la Rivière. This was a highly innovative dictionary as both an encyclopaedic work and one with a pedagogical intention. We explore the declarations in the prefaces and the encyclopaedic and linguistic content concentrating on the 1701 edition that is currently being fully digitised in XML-TEI. Citations from largely contemporary texts were used to illustrate entries leading to a very wide knowledge network of late seventeenth century science. Basnage also experimented in illustrating usage through examples, grammatical and pronunciation information.

**Keywords:** Front matter; historical dictionaries; digitisation; history of ideas

## 1 Introduction

“Il n’y a point de Livres qui apportent plus d’utilité au Public, et moins de gloire à l’auteur”. Antoine Furetière (1619-1688), *Factum*. (1859a: 3.)

(There is no book that brings more value to the public and less glory to the author).

Although it is possible to treat prefaces as standalone texts, the important information they contain only acquires its full sense when seen in the light of a systematic analysis of the dictionary contents. This is very much the case for the three principal editions of Antoine Furetière's *Dictionnaire Universel*, henceforth DU, published at the end of the 17<sup>th</sup> century. This paper will look in detail at the claims made in the front matter of these works, whilst cross linking these to the fully digital versions actually in preparation.

The history of the DU from its initial publication in 1690 to its final edition in 1727 displays an interesting series of actions and reactions. The first action is the publication itself which provoked a hurried first edition of the *Dictionnaire de l'Académie Française* (1694), accompanied by the *Dictionnaire des Arts et des Sciences* (1694). Unwilling to lose a profitable product, the publisher, Leers, commissioned a new edition under the editorship of Henri Basnage de Beauval, this being published in 1701. In order to undermine the Protestant publishing industry, the Jesuits of Trévoux immediately started on a plagiarised version (Trévoux 1704), cleansed of so-called protestant heresy. This was brought out in 1704 and was the beginning of a long series of dictionaries that gradually distanced themselves from the Basnage's version of the DU. On the Protestant side, *Trévoux* was met with a partially revised version in 1708. This was more a reprint than a revised edition, as changes were minor. The front matter does not change as a truly revised version was planned but this was brought to a stop by the death of Basnage in 1710. A revised Trévoux dictionary came out in 1721, followed by another edition of the DU in 1725/1727, under the editorship of Brutel de la Rivière, an edition that will prove to be the last one, because of the publishers' death. All of this was documented in fierce exchanges in the scholarly journals of the time, and reflected in the front matter of the three editions.

Front matter is abundant in the three editions, as this includes the *privilege*, the right to publish, and, notably in the case of the Brutel edition, comments from the publishers. Working in the context of the French nationally funded BasNum project,<sup>1</sup> our aim here is to use this front matter to study the development of a work that was highly influential in the rise of encyclopaedic dictionaries and of encyclopaedia in the eighteenth century. After a presentation of the background and of the general principles of the three major contributors to the three main versions of the DU, we will concentrate on two aspects: the encyclopaedic nature of the dictionary and the pedagogical dimension of this work. Working from the fully digitised texts, we shall show just how powerful a lexical and encyclopaedic model was thus gradually being built.

## 2 The Three Editions of the *Dictionnaire Universel*

By the time the *Dictionnaire Universel* was published in 1690, its author, Antoine Furetière, had already been dead for two years. It was thus the Protestant émigré philosopher Pierre Bayle who wrote the preface. However, Furetière had already

<sup>1</sup> <https://anr.fr/Projet-ANR-18-CE38-0003>



published his *Essai* (1684), a proof of concept destined to show that he had not plagiarised the, as yet far from published, dictionary of the Academy, and a series of *factums* (Furetière 1859), pamphlets against members of the French Academy who were attempting to prevent publication of his dictionary.

Dedicated to the King of France, Louis XIV, the *Essai* (1684) is quite clear in what Furetière had set out to achieve. The aim is to build an encyclopaedia that will be usable by foreigners and will ensure a place for the French language in posterity. This was not to be a rival dictionary to that of the Academy, but an overview of the French language of the late seventeenth century with all the necessary terms from the arts, crafts and sciences. Such a presentation was vital in defending Furetière's right to publish (the *privilège*), as the *Académie* had a monopoly on the creation of a dictionary of general language usage. Nevertheless, the *Essai* brought forth fierce opposition from the Academy, which had hitherto failed to produce its own more limited work. Furetière's response was the above-mentioned *factums*. The main thrust of the first one (Furetière 1859a), printed in Amsterdam in 1685, is a defence of his right to a privilege to create and publish a dictionary in his own name. In the second, he uses sarcasm to the full and attacks individual members of the *Académie*. In the first *factum*, he denies the demand of the *Académie* for a monopoly on all monolingual dictionaries for a period of twenty years on the basis that they have not opposed their *privilège* in other cases. He also refutes the accusations of plagiarism from the *Académie* challenging to prove that that had not consulted other works as those of Vaugelas and Ménage. Both are essentially legalistic in nature and not our main concern here, as this has been amply done by Rey in his biography of Furetière (Rey 2006).

Furetière lost the following legal battle, which led him to contact the great Dutch publishing house Leers, who published the work posthumously in 1690. This was not so unusual procedure as many French intellectuals, Catholic and Protestant, used the Dutch publishers as a means of getting around the censor. The dictionary was well-received, including by Louis XIV himself, despite the embargo from France, and was read widely. Indeed, it is this edition that was used by Bluteau when writing his dictionary of Portuguese (Verdelho & Silvestre, 2007) and the work was also present in the library of Matthias Moth, who was engaged in his dictionary of Danish (Eegholm-Pedersen, personal communication).

Whilst the writings of Furetière were essentially defending his work, the preface by Bayle in 1690 introduces the finished dictionary. What Bayle insists upon is the advantage the French language is taking upon ancient and modern languages, that do not benefit from such an extensive description as that of the DU. Comparing Furetière's work to the admirable lexicons produced by the Estienne and other Latin and Greek scholars, he underlines that the DU offers "*la langue de tous les jours*" (everyday language), which is irremediably lost for the ancient languages, since it did not pass in the books and other writings we have inherited from these old times. The DU becomes therefore a model for all the nations, and Bayle invites scholars from all over Europe to build similar works about other idioms. While it is not clear if, in Bayle's opinion, all languages are worthy of such a complete description, nor what his or Furetière's view were on the evolution of the language, the approach is extremely distinct to the one taken by the *Académie*. The DU appears thus not as an instrument for freezing the language in a point of perfection, which is what the academy was vainly hoping to achieve, but as a snapshot of the ways of speaking in France at a certain moment in time.

To this idea of capturing the general use rather than the "*bel usage*" Furetière adds another, promised to a great future, that of describing the objects, notions and entities the lexical units are designating, as much as glossing upon the words. The dictionary offers thereby a simple means to acquire accurate knowledge about a wealth of domains, so that one can engage into an informed conversation with the specialists of the various disciplines, trades, arts and crafts. This is all the more interesting as Furetière, Bayle states, has managed to convey a great deal of information without being pedantic or 'dry'. There is something 'curious' to be read in each entry, instruction and agreement going hand in hand in this impressive work. The dictionary being a great success, a totally revised and corrected second edition, with a new preface in addition to that of Bayle, was produced by the Protestant refugee Henri Basnage de Beauval (1657-1710) in 1701. The 1701 preface demonstrates why there was a need for revision and how this was carried out. He underlines the extent of his revisions, but never claims the work as his own.

Basnage's first goal is, of course, to revise and correct the numerous errors of Furetière's work, answering thus the call for improvement launched by Bayle in his *Préface* of the 1690 edition:

Pour conclusion on avertit le public, qu'on est bien éloigné de croire qu'il manque rien à cet Ouvrage. Un Dictionnaire est un de ces livres qui peuvent être améliorés à l'infini ; & quoy qu'on ne les gâte que trop souvent dans les dernières Editions, il faut pourtant convenir, qu'en general la première n'est qu'une ébauche en comparaison de celles qui la suivent [...] (Furetière 1690)

(To sum up, we warn the public that we are fully aware about everything this work is still missing. A dictionary is one of those books that can be continuously improved; and even if, in fact, the last editions do more harm than good, one must agree that the first one is more of a sketch of the following ones [...])

In fact, Basnage goes far beyond correcting and completing the entries, and generally tends to perform these activities with a new vision about what the dictionary should be. The most noticeable initiative is probably the decision to reinstate the "bel usage" that Furetière has left aside as being the task of the *Académie*:

On a cru que pour bien remplir le titre de *Dictionnaire universel*, il fallait qu'on y pût y apprendre à parler poliment, aussi bien qu'à parler juste, & dans les termes propres à chaque Art. (Basnage 1701)

(We believe that in a fully universal dictionary one should find learning material for speaking politely, as well as for speaking in an accurate manner, and with the appropriate terms from each domain.)

Writing largely after the quarrel between Furetière and the *Académie*, and in a foreign country, Basnage has not to worry about the *privilège* protecting the prestigious French institution. He can afford to diverge from his predecessor in his strategy to underline the specificity of the DU, that he places elsewhere than in the somewhat artificial distinction between 'specialised' and the 'common' language put forward both by Furetière's *factums* and by Bayle's preface. So great is his



certitude about the quality of the product he delivers, that he can even afford to quote the *Dictionnaire de l'Académie* as a resource for this very 'polite language' whose re-inclusion is announced in his preface. However, he reserves also the right to mention points of view about the correctness of a word, genre, spelling or pronunciation, etc., that are not in line with those of the *Académie*:

peut-être aussi que l'on ne sera point fâché de revoir les raisons de douter; ces sortes de contestations forment, & raffinent le bon goût: ce n'est pas peu de chose que de sçavoir douter par raison. (Basnage 1701)

(maybe one would not be unhappy to find reasons to doubt; this kind of discussions form and strengthen good taste: it is important to base ones' doubts on reasoning.)

Added to the fact that all over the dictionary Basnage pays attention to other sociolinguistic specificities (regionalisms and expressions of a specific trade being underlined as often as the fact that such or such word is *bas* (low) or *familier* (colloquial), these disagreements show that, rather than supporting the ideal of a 'perfect' French language, the reintroduction of the 'polite speech' is the result of a larger interest for the idiomatic variety, of which *l'usage de la Cour* (use of the Court) is just one example.

A second innovation mentioned in the *Préface* are the quotations from "the most excellent authors". Actually, their works as building materials of the DU are mentioned on the title page of the first edition, in 1690, but while Furetière hints towards a writer from time to time, Basnage decides to introduce short abstracts from their works in a large number of entries. On the one hand, his goal is to counteract in a personal way the 'dryness' of the alphabetic work, an aspect for which we have seen above that Furetière has already received much praise from Bayle. On the other hand, this addition aims less to set models for the right or the elegant use of a word, even if quotations could have such an effect, as an aftermath, but to help differentiating between various senses of the same word, a particularly difficult point in Basnage's view. Finally, Basnage is clearly less interested in the 'excellency' of the quoted authors, than in their efficiency in illustrating a semantic nuance, and while keeping the same claim on the title page, out of reverence for Furetière, he confesses in the *Préface* that he quoted everybody without praising or blaming, "*je les cite tous également*" (Basnage 1701).

As we will see in the next part of this paper, these innovations are far from covering the entirety of the changes performed by Basnage; nor do they occupy the entire *Préface*, in which many other aspects are discussed, such as the use of a 'historical orthography', in opposition to those appealing to adopt a phonetic orthography for French, or the elimination of the 'proper names', better suited in the *Dictionnaire historique* of Moreri. But they are clearly salient elements allowing to perceive and to understand Basnage's reinterpretation of the idea of an "universal" dictionary. While following Furetière's intuition about the necessity to offer a book in which the words and the things they designate are described together, Basnage is clearly re-equilibrating the balance in favour of a more lexicographic approach, attentive to the many layers and aspects of the language, as well as to the contexts, practices, ideas and mentalities that leave their imprint on it. Perfectly conscious about the impossibility to produce an entirely 'universal' dictionary, Basnage has a more holistic vision of it, aiming to create a work that accurately reflects the complexity and the intricacy of the language. This is quite different to Furetière who appears, maybe perforce, more interested in the 'margins' and the non-conventional, and who approaches the vocabularies of sciences, arts and crafts with quite the same picturesque vein that he uses to describe the manners of the bourgeois living rue Mouffetard.<sup>2</sup>

The Basnage edition was reprinted in 1702 and 1708, and certainly several times in between, but there was no revised edition until 1725. Basnage was fully aware of the shortcomings of his edition of the DU and was busy revising it at the time of his death in 1710, having only reached the beginning of the letter E. As the dictionary was not just considered by its publishers as a work of intellectual importance, but also of great commercial value, finding a new editor was vital. The choice fell on another refugee, Jean-Baptiste Brutel de la Rivière. This period from 1710 to publication in 1725 is detailed by Brutel in his preface where he explains how work stopped at the death of Reinier Leers in 1714 and was only picked back up in 1721 at the instigation of consortium of six Dutch publishers who had bought the rights from Leers' inheritor. The first volume appeared in The Hague in 1725 with the other two volumes in 1727.

The new edition benefits from a Dutch *privilege* and is dedicated to the Prince William VIII of Hesse-Kassel. The new *Préface*, that follows the two previously commented ones, shows that Brutel remains globally faithful to the idea of a universal dictionary, but rather in Furetière's sense. He reminds and even insists on the French language as the courtly *lingua franca* of the period, a key argument in Bayle's preface too, but somewhat less important for Basnage whose approach is more 'philosophical' in the terms of that time. He announces the introduction of new terms from a variety of arts and sciences, and the title page displays, indeed, new domains, such as mythology, dance, fencing and economy.<sup>3</sup> However, in addition he considers fit to add proper names and information about the various 'sects',<sup>4</sup> that Basnage has excluded. He is also insisting on the increases he made in the number of the *vieux mots* (old words) covered by the dictionary, as well as in the number of *termes des relations* (foreign terms, introduced in France by ambassadors and tradesmen). This disparate list of changes the preface covers suggests that additions have been made less because they were completing the picture about the complexity of the language, and more because they appear 'curious' and offering the proper kind of instruction and recreation to the *honnête homme*. This figure is clearly the target Brutel has in mind when revising the dictionary, with a very strict definition about what is 'honest' and what is not. Indeed, whilst Basnage was a lawyer with very wide interests, Brutel was a pastor and man of letters who was keen to impose his view, notably on moral issues.

<sup>2</sup> This is the setting of his 1666 novel, *Le Roman bourgeois*.

<sup>3</sup> He eliminates, in exchange, weaponry, maybe because he considers it not enough differentiated from other *arts mécaniques* (mechanical arts).

<sup>4</sup> Following the tolerant spirit of Basnage, he emphasises though the neutrality of his approach when increasing matter pertaining to all religions.



Therefore, while he continues Basnage's habit of quoting, he pays increased attention to the 'excellency' of the authors:

J'ai puisé les exemples que je cite dans nos meilleurs Auteurs; j'ai choisi ceux qui en fixant l'usage de la Langue, contiennent quelque pensée fine, quelque trait ingénieux, ou quelque maxime importante, propre à éclairer l'esprit, ou à purifier le cœur. (Brutel 1725)  
(I have picked up the examples I quote from our finest authors; I have chosen those that settle the right use, while conveying some delicate thought, some clever thinking, or some moral recommendation, so as to enlighten the spirit or to purify the heart.)

He becomes therefore quite angry with the publishers that put back in the texts quotations from authors he disapproves of, such as Rabelais, and headwords that he excluded, and he insists that the reader does not consider these as being his doing. In addition, Brutel is concerned with a level of correct usage that goes towards the pedantic, and has been criticised by Basnage in his own foreword. In order to update spelling to the latest conventions, he made use of the most recent, 1718, edition of the *Dictionnaire de Académie Française*, which he praises as being "à un point qui n'est pas éloigné de la perfection" (at a point that is close to perfection). He draws upon Desmarais' and Buffier's grammars, and claims even to have made wide use of the *Trévoux* dictionary - a rival work, but corresponding to his views about correctness in language use and moral mission of a book.

As announced by its front matter, this third version appears therefore far more prescriptive. Its audience is also more restricted, as Brutel dedicated his work to "le monde poli et savant" (the polite and educated world). To a certain extent, both the dedicatory epistle and the *Avertissement des Libraires* try to counterbalance this, by insisting on the superiority of the DU as opposed to the *Dictionnaire de l'Académie* and to the plagiarising *Trévoux*. However, this is clearly the end of the line in a very interesting lexical and encyclopaedic endeavour, and the following dictionaries in the 18<sup>th</sup> century will not pick up the thread, while often making use of the materials Furetière and Basnage gathered.

### 3 The Dictionary as an Encyclopaedia

As underlined above, the encyclopaedic dimension is important to the three main authors of the DU (Furetière, Basnage and Brutel). In his second *factum* (Furetière 1859a), Furetière points out that whilst the members of the *Académie* might be able to judge polite literary language, they were utterly incompetent when it comes to the usage of terms from trades and professions, which were also part of everyday usage when conversing, to use his examples, with an architect, a military person, a courtier or a lawyer. This argument is stated again by Bayle in the *Préface* of the first edition:

On ne sera plus réduit, comme le sont tant de gens, dans les matières même les plus communes, à recourir au mot vague de chose, de pièce, & à faire des postures de mains & de pieds, (manières qui passent avec raison pour rustiques) afin d'exprimer la figure, la situation, & l'étendue de ce dont on parle. Cet Auteur apprend à tout le monde, non seulement la nature des choses par leur matière, leurs usages, leurs espèces, leurs figures, & leurs autres propriétés, mais aussi les termes propres dont il se faut servir pour les décrire.  
(One will not be obliged any more to use the vague words of 'thing', 'piece', as most of people do, even when speaking about very common things, or to gesticulate with hands and feet (all manners that are rightly considered as rustic) so as to indicate the figure, the position or the size of what one is talking about. This author teaches everybody the nature of each thing, its materials, its usages, the different sorts it comes in, but also the appropriate terms for describing these.)

The same idea is obviously integrated by Basnage and Brutel to their own approach, even if they do not insist on it in their preliminary material, referring the reader to what has been previously written on the subject.

However, the methods for building the encyclopaedic contents are quite different, as we will see in what follows. Also, this is a dictionary and not a terminology, and consequently, it is not systematic in its coverage. To take but an example, for the verb *abatre* (Furetière 1701) gives seven senses, only two of which it designates clearly as being terms by supplying a domain name. In other cases, as with the verb *battre*, there are no terminological uses at all, although examples could be deemed as giving specialised usage. Still, the evolution of title pages and the term coverage indicate that Basnage and Brutel tried to improve the situation as compared to Furetière, and while one cannot talk, in the case of the DU, about a full "knowledge tree", one can see this being gradually sketched as the dictionary grows.

#### 3.1 Building the Encyclopaedic Contents

Furetière, as Basnage and later Brutel, builds on a great deal of bibliographical material, that is gradually being identified within the BASNUM project. He insists on the increase and the update of the information he provided from the most recent specialised works in the various arts and sciences he covers. However, Basnage also took a different road, since, in addition to quotes and new ideas from books, he calls upon a medical doctor with knowledge of medical and natural sciences to rewrite specialised entries, as well as upon a mathematician:

Je ne mets pourtant pas sur mon compte les articles d'Algèbre. Cette science m'est inconnue. Je ne m'approprie point non plus ce qui regarde la Médecine, l'Anatomie, la Pharmacie, la Chirurgie, & la Botanique. Je n'ai point voulu me fier à moi-même là-dessus. Un habile Mr. Régis, Médecin à Amsterdam homme s'en est chargé.  
(I do not declare as mine the entries about algebra. I do not know this science. I do not claim either anything pertaining to medicine, anatomy, pharmacy, surgery and botanical sciences. I did not want to trust myself on these matters. A knowledgeable man, Mr. Régis, doctor in Amsterdam, took charge of these.)

Basnage anticipates thus one of the most powerful features of the *Encyclopédie* by Diderot and d'Alembert (1751), whose interest comes to a great extent from the fact that a team of more than 200 contributors provided specialised knowledge for the various entries (Proust 1962). This work has been explored by Leca Tsiomis (1999) who details the debt to the *Dictionnaire Universel* of Furetière.

The play off between Furetière, Basnage and Régis is interesting to observe and will be done through authorial studies as



the full text becomes available. The question is already receiving a partial analysis in looking at trees and gardens. Insofar as he had to work quickly, Basnage often just revised sections of Furetière's work, or simply copied verbatim. This can be illustrated with the word *JARDINAGE* (gardening), where the italics show the additions by Basnage:

*JARDINAGE*. subst. masc. L'art de cultiver les jardins. [...] Le jardinage a été mis depuis peu de temps en un haut point de perfection par le Sr. Le Nôtre. *La Quintinie est encore allé plus loin, & nous a donné une ample instruction sur le jardinage. Mr. Fatio a donné, depuis quelques mois (1699) au public un livre sur le jardinage où il enseigne les moyens d'employer utilement les reflexions du soleil.*

Whilst Furetière only mentions André Le Nôtre, Basnage adds two more recent publications, that of the agronomist Jean Baptiste de La Quintinie (1626-1688) (*La Quintinie* 1690) and a highly influential paper published by the Royal Society from the astronomer Nicolas Fatio de Duillier, (1664-1753) dealing with the orientation of walls and greenhouses for fruit growing (Fatio 1699).

In cases of botanical interest, such as *COIGNASSIER* (Quince), we find a total rewrite of the entry, which clearly points to an intervention from Dr Régis.

*COIGNASSIER*, ou *COIGNIER*. s. m. Arbre qui porte les coins, & qui ne devient jamais fort grand à cause de la pesanteur de son fruit qui fait pancher ses branches vers la terre. Son bois est tortu, pâle & blanc par dedans, assez ferme & égal. Ses feuilles sont semblables à celles du pommier, fort cotonnées sur le dos, lisses & vertes de l'autre côté : elles ne sont point decoupées sur les bords. Ses fleurs ressemblent à celle des roses sauvages : elles sont composées de cinq feuilles presque rondes & de couleur de chair. Sa semence est renfermée dans son fruit : elle rend l'eau dans laquelle on la fait tremper, épaisse & mucilagineuse. Son fruit est appelé *coin*, il en sera parlé en son lieu. Quelques Jardiniers disent que le *coignier* est le mâle, & le *coignassier* la femelle. La Quintinie pretend qu'il n'y nulle difference. On a donné au *cognassier* le nom de *cydonia*. Ce mot vient de Cydon ville de Candie, d'où ce fruit fut porté en Grece. On l'appelle aussi *malus cotonea*. Les meilleures especes viennent de Nevers & d'Orleans.

In the above example, the underlined text is all that remains from Furetière.

In the cases cited above, Basnage gives the full name of the author, in other cases an abbreviation is used. In the front matter to the 1701 edition, Basnage provides a list of abbreviations for works and authors cited in the text. We have as yet no idea as to why or how this list was made as it is far from complete consisting as it does of some 117 authors, twelve works and one institution – the *Académie Française*. Given that Basnage was working sequentially through the letters of the dictionary, taking the revised pages immediately to the printer, it is probable that this list represents the works and authors seen on a first reading of Furetière's edition. One of the tasks being carried out during mark-up is to list all the cited sources and to complete a prosopographical profile for each author, linking when possible to their ISNI code so as to reduce ambiguity. Current work has led to 240 cited authors and 32 printed sources. This will increase when the named entity analysis is carried out with a post-sorting between authors cited and persons simply mentioned.

Works cited show the extent of interaction amongst members of the Republic of Letters. It also shows a widening of the net from Furetière's essentially Paris-centred knowledge base, to a Europe-wide one known by the Protestant diaspora. It also brings to the fore the influence of the learned academies such as the Royal Society and the Berlin Academy, both of which Basnage was a member.

### 3.2 Term Coverage

Bayle was quite clear as to what differentiated the DU from that of the *Académie*:

Mais pour Monsieur Furetiere, il ne s'est pas proposé les termes du beau langage, ou du stile à la mode, plus que les autres. Il ne les a fait entrer dans sa compilation que comme des parties du tout qu'il avoit enfermé dans son dessein. De sorte que le langage commun n'est icy qu'en qualité d'accessoire. C'est dans les termes affectez aux Arts, aux Sciences, & aux professions, que consiste le principal. (M. Furetiere did not intend to concentrate on words from the polite language, or the fashionable style. He put them in his dictionary only insofar as they are parts of a whole that he intended to cover. The common language can be found in this work at a very accessory place. The largest part of it concerns the terms from arts, sciences and crafts.)

The title page of the 1690 edition offers an impressive list of arts, sciences and professions to be covered by the DU, as it can be seen in the following figure:



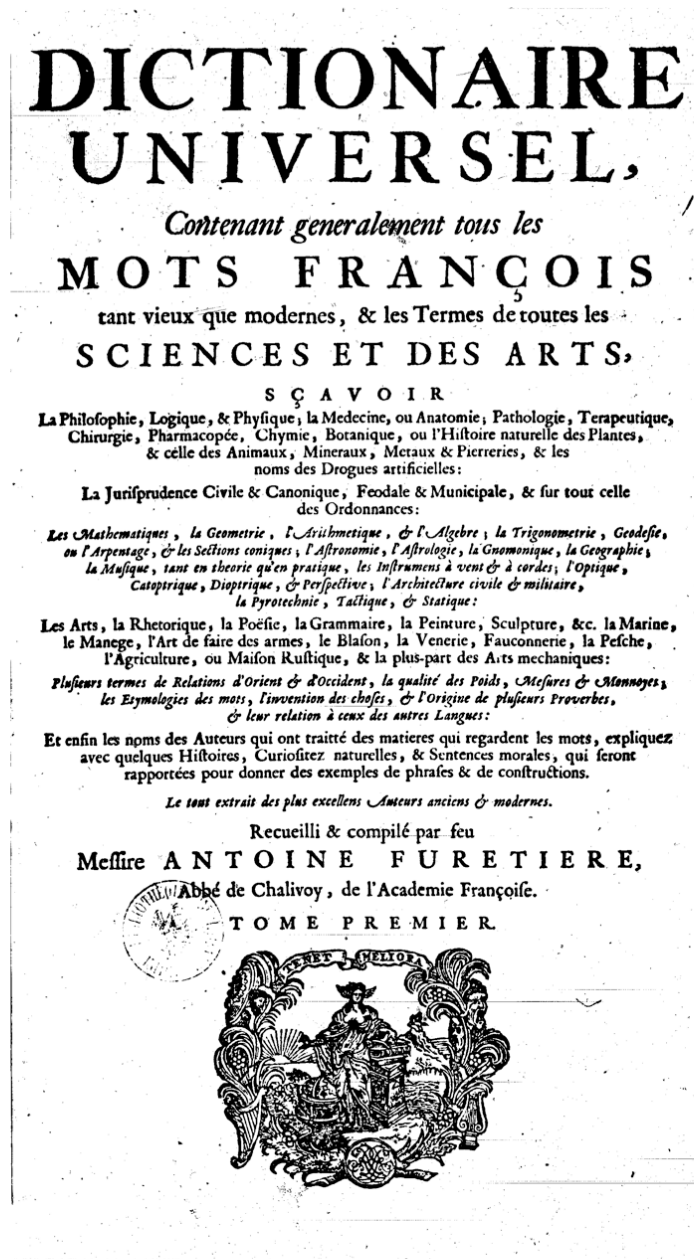


Figure 1. Frontispiece of Dictionnaire Universel 1690.

From a modern point of view, this long enumeration appears as somewhat disorganised for several reasons: disciplines going together from our point of view are quoted in different places (see, for instance, ‘physics’ put near ‘logic’, while ‘optic’ appears after ‘music’ and ‘wind and string instruments’), some distinctions are puzzling (‘mathematics’ is quoted apart from ‘geometry’, ‘arithmetic’, ‘algebra’ and ‘trigonometry’, as well as ‘conical sections’), the global order does not correspond to our modern separation between ‘hard’ and ‘soft’ sciences, and we do not immediately perceive what several labels, such as ‘artificial drugs’, or ‘mechanical arts’, refer to. However, by looking closer one starts to perceive some principles of organisation. Italics and roman fonts alternate as the list is split in several paragraphs, so as to ease the reading and to create disciplinary blocks. To these, we can add the attention paid to the punctuation, with commas, semicolons and colons carrying semantic baggage. While the first indicate that the arts and sciences they separate are intellectually close, the second are triggers, a list within a list, as for the various applications of mathematics, with trigonometry and astronomy commanding each a different set of subdomains. Finally, columns underline an epistemic gap, as the one between the paragraph stating with *philosophie*, and the second one gathering the juridical sciences. Also, inside each ‘paragraph’ Furetière goes from the more speculative sciences to the applied ones: by the ideas of the time, physics is, indeed, related to philosophy, and the medical sciences are based on it, while civil or military architecture, as well as tactics and



pyrotechnics, derive from mathematics.<sup>5</sup> As for the particular place legal disciplines receive in this title page, this is to be related both to Furetière's specialisation, and to its historical place and development in the higher education institutions of the time.

Apart the addition of his name and role, Basnage de Beauval does not change this title page, even if in fact he is not strictly following the domain repartition thus listed. The digitisation of his DU shows that he deals with more than 370 declared term fields, from *accoucheur* (midwife, also *sage femme*) to *voltigeur* (acrobat), a figure that indicates the extension he gives to Furetière's view about the variety of arts, sciences and crafts. However, since this remodelling of the knowledge map is not declared as such in the front matter, we will not elaborate further on it in this paper.

In 1725, the divergence from Furetière is such that Brutel feels the need to remodel the title page, while staying as close as possible to the original work. Certain disciplines are now printed in capitals, while others remain in lower case, while the line breaks add further meaning about the place of each domain in the organization of knowledge. The first 'paragraph' of the list therefore reads:

LA PHILOSOPHIE, LOGIQUE ET PHYSIQUE, LA MEDECINE;  
Anatomie, Pathologie, Therapeutique, Chirurgie, Pharmacie, Chymie, Botanique ;  
l'Histoire naturelle des Plantes, des Animaux, Minéraux, Métaux & Pierrieres,  
& celle des Drogues naturelles & artificielles:

Together with other changes that one can observe between this list and that of Furetière (such as the modification of *pharmacopeia* in *pharmacie*, or the elimination of the coordination between *botanique* and *histoire naturelle des plantes*), this reorganisation is telling about the evolution of the encyclopaedic contents of the DU, both respectful of the intentions of the founding father and brought up to date thanks to extensive reading.

This does not mean that the imbalance one can find in the 1690 edition is entirely corrected. Some domains continue to get better coverage than others, with legal language being very well represented (since both Furetière and Basnage were lawyers by training), even if Brutel obliterates the importance of jurisdiction by spelling it in lower case in his title page. As we have already seen, agronomy, and related fields, get detailed treatment as it was very much a fashionable subject at the time. Architecture (see Williams 2020) was another area that received great attention, as well as maritime matters at a time when sea travel had such great importance.

Mapping terms to domains is a complex task, especially in areas as law where the interrelation of different fields is far from easy to anyone not acquainted with seventeenth century French legal practice. The same goes for numerous domains so one essential task is to relate the domain names to current spellings, to the domain names used in *Encyclopédie* (Diderot and d'Alembert 1751), and also to a modern natural ontology that will facilitate access to terms in the new digital edition. The initial stage in mapping the terminological usage is by identifying the formulae that link senses to a domain. The most frequent are "*terme de* [domain]" (term of [domain]), "*en termes de* [domain]" (in terms of [domain]). There are 3997 of the former and 2828 of the latter, which gives an idea of the scope of dictionary. In addition to these, specialised usage may also be given using the formula "*en* [domaine]". This is too vague for automatic extraction until the definitive list of domains has been established, but, for example, there are 396 terms for *medicine* (medical practice, note that the accent on *médecine* was not used at the time) and a further 180 for *en médecine*. This leads to another problem as *medecine* refers to general practice, but we also find related fields as *chirurgie* (surgery), *anatomie* (anatomy) as well as *arracheur de dents* (puller of teeth) and other practitioners as *apothicaire* and *herboriste* or the *accoucheur*, mentioned above. Sometimes we have both a domain and a practitioner as in *jardinage* (gardening) and *jardinier* (gardener) or *charpenterie* (carpentry) and *charpentier* (carpenter). Why the authors use one or another is not necessarily clear as yet, but it is to be hoped that a fully digitized version of the three editions will help shedding some light over the reasons behind the lexicographic practices.

#### 4 The Dictionary as a Learning Tool

Judging by the *Factums* and what Bayle says in the *Préface*, Furetière intended the DU as a learning tool, addressed on the one hand to foreigners that need to speak the most widely used language in Europe at the time, and to the other hand to the "honnêtes hommes" interacting with specialists from the various disciplines and trades. Basnage is going even further, making the continuous study of the language a characteristic of any fully educated man, beyond any utilitarian aim:

On sçait bien qu'il ne faut pas être trop pointilleux, & qu'on énerve, ou qu'on desseche le discours à force de le limer, & de le polir. Une régularité trop grammaticale a quelque chose de pedantesque : mais aussi le mauvais choix, ou même la trop grande négligence des expressions, est un défaut beaucoup moins supportable. Il n'y a pas grand honneur à bien sçavoir sa langue, & il y a de la honte à ne la sçavoir pas. Ensorte que si les observations, ou si l'on veut, les minuties de Grammaire dont ce Dictionnaire est rempli, ne sont pas fort essentielles pour parler, quand on ne parle que pour se faire entendre, elles ne sont pas tout-a-fait meprisables pour ceux qui se piquent de parler exactement, poliment, & noblement.

(One knows that one should not be too strict, and that speeches become bland and diluted when they are too polite. Being too grammatically rigorous is pedantry: but picking up the wrong expressions, or in a too negligent way, it is an even greater error. There is no honour in knowing one own's language, but there is shame not to know it. As a consequence, the observations, or even the minute grammatical details this dictionary is full of, are not essential for speaking, when one speaks only to communicate, but they are not

<sup>5</sup> To give another example from the contents of the DU, in ship building and maintenance, carpentry is often to be found related to *marine* (maritime, 1042 terms) and *mer* (sea, 145 terms), as well as in fortification, as this is the period of the great military architect and engineer, Sébastien Le Prestre, marquis de Vauban (1633-1707), generally known simply as Vauban.



altogether useless to those who try to communicate in an exact, polite and noble way.)

In both cases, this means that the contents of the dictionary have to be modelled so that linguistic regularities and particularities of French are clearly displayed.

Although Basnage eschews grammatical pedantry and recognises that modern languages are more easily learned through usage than through grammar, his attention to the grammatical and pedagogical aspect of his dictionary is undeniable.

For Basnage grammar is, together with use, the tool that allows one to master the language. This point is made all the clearer when one reads the statement (which is already in Furetière):

REGIME, en termes de Grammaire, est la syntaxe ou concordance que des mots doivent avoir les uns avec les autres suivant les règles de la Grammaire, ou l'usage de la Langue. (*REGIME*)

(REGIME, in terms of Grammar, is the syntax or concordance that words must have with each other according to the rules of Grammar, or the use of Language.)

In fact, his work appears to be a form of learner's dictionary through various facets that emerge clearly from the comparison with the Furetière's edition. First of all, on several occasions Basnage simplifies the more tortuous thought of his colleague: for example, what Furetière had expressed through a metaphor is replaced by the corresponding concept (*son fruit* in Furetière becomes *un enfant* in Basnage under the entry *MÈRE*). In other cases, the simplification concerns Latin sentences. In fact, Basnage translates into French particularly meaningful passages of Scripture that Furetière had reported in Latin. For instance, in *DIEU* entry, Basnage's *Je suis qui je suis* translates Furetière's *Ego sum qui sum*.<sup>6</sup> Moreover, the so-called grammatical words such as personal pronouns and possessive adjectives, mostly ignored by Furetière, find ample space in Basnage as well as the treatment of prepositions, although they are present in Furetière, is much better articulated in Basnage.

As stated above, the Basnage edition is undoubtedly the most inclusive of the three because it aims to describe the French language in its entirety. This description always relates to the reader of the dictionary to whom Basnage provides orientation tools in such a vast universe, both in the process of reception and in that of production of the language. Indeed, the lexicographer gives, where they exist, the orthographical variants of the entries, the connotations of the senses (*il est neutre* – it is neutral), the register indications (*ce mot est populaire* – this is a people's word), the diachronic specifications (*ce mot est vieux* – old word) as well as the diatopic ones (*mot Picard* – word from Picardy). In short, the reader of Basnage knows exactly in which context to use a certain term, in the book of which author can meet such or such archaism, in which region of France to expect a certain word, etc.

Understanding the structure of the language being studied is certainly a first step towards its acquisition. It is not surprising, therefore, the presence in the dictionary of the suffixes used in the formation of the French language's lexicon (e.g. *IEL. IEN. IER. IEZ*). This clarifies Basnage's desire not to overlook any aspect of the subject he is dealing with and consequently, to be useful to as many readers as possible. Some paragraphs are pointedly dedicated to a specific section of his audience, for instance *ces remarques ne regardent que le Poëtes* (the following observations concern only poets).

Quite often, in the various senses of an entry, the reader encounters suggestions for use (e.g. *il faudroit dire* – one should say) and constructions to avoid (e.g. *J'aimerois mieux l'éviter en disant* – I'd rather avoid this formula by saying), very similar to those a teacher would give to his or her students.

Two other elements contribute in a decisive way to characterise Basnage's work as a learner's dictionary: the information he gives about pronunciation, and verb inflections he provides. Being Basnage a pioneer in the adoption of both innovations, it is inevitable that an accurate and systematic work cannot be expected. Still, he provides a wealth of material to reflect upon.

The 1701 edition records the pronunciation of about 780 lemmas. Although Basnage indicates the pronunciation in various ways through the work, three patterns can be recognized which sometimes combine with each other:

- The invitation to read or not to read one of the letters contained in the headword: in this case, one often encounters formulae of the type *x se prononce* and *x ne se prononce pas* (such and such letter is to be pronounced or not). In the case of *h*, he invites to aspire it (*aspirez l'h* – aspire the *h*).
- The indication of the pronunciation of the whole word in capital letters, introduced by *Prononcez*.
- The indication of the headword's syllabic quantity by means of sentences of the type *la première syllabe est brève* (the first syllable is short).

Sometimes, in order to save time and space, Basnage adopts the expedient of indicating the pronunciation within a single entry and making it count for all the headwords of the page. For example, under *HAILLON* one reads: "L'h de ce mot & de tous ceux qui sont dans la page suivante, s'aspire & se prononce." (The 'h' of this word and all the following ones on the next page are to be pronounced).

As for the inflections, the Basnage edition contains the conjugation of about 240 verbs. The lexicographer's idea was probably to provide conjugation only for irregular verbs, unfortunately there is no uniformity in the treatment of inflections. Perhaps Basnage intended to indicate in the dictionary the only forms that in his view were more difficult to infer from a general knowledge about verb inflection in French. The cost of printed paper has probably also played a role in the selection of entries to be completed with such grammatical information. All in all, the result is quite unsystematic, with some verbs inflected exclusively at the first singular person of indicative present (*On conj. J'accourcis* – One inflects *J'accourcis*, entry *ACCOURCIR*), others illustrated with the first singular person of imperfect and present perfect (*Je m'accouplai, je me suis accouplé*, entry *ACCOUPLER*), or the first singular person of indicative present, imperfect and simple future (*Je m'accoude; je m'accoudai, je m'accouderai*, entry *ACCOUDER*) and so on. In addition, in the cases of derived verbs, the

<sup>6</sup> This may also be important from another point of view since it highlights the divergent approach to the Bible texts of Catholics and Protestants. However, this aspect will not be discussed here.



inflection sometimes is given (see *REDIRE*), and sometimes not (see *ADJOINDRE*).

Regardless of the result, the effort of inserting headwords' pronunciation and verbs' inflections within the dictionary could denote a certain closeness to the linguistic and cultural needs of the French community in exile in the Netherlands, of which Basnage himself was a part.

## 5 Conclusion

In conclusion, it is clear that the *Dictionnaire universel* was a highly innovative work, and that the changes that Basnage tried to bring in were way ahead of his time. Comparing the three series of front matter allows us to seize the originality of the work, but only by linking the points raised in the prefaces to what is actually happening in the dictionary itself can the degree of innovation be grasped. The very title page and the *factums* of Furetière underline the encyclopaedic nature of the work, considering the terms of arts, sciences and trades that are covered. However, only through access to the full dictionary in a digital format can the breadth of coverage be seen and a qualitative analysis of the degree to which individual domains are handled. What is not immediately obvious is the breadth of the knowledge base developed by Basnage. This started a tradition that led to the work of Chambers, and then onto the *Encyclopaedia* of Diderot and d'Alembert. The influence was real, but is not always recognised. The failure of the consortium, who bought up the rights to the dictionary from Leers, to develop, promote and distribute the work, inevitably led to the rival Trévoux dictionaries taking the limelight.

The other oft overlooked innovation concerns the pedagogical aspect of the dictionary. The prefaces give general information as to a type of audience, the learned French for Furetière who was working still in the context of a developing academy tradition. What is clear though is that Basnage's work as editor of a learned journal, coupled with the fact that he was working in a non-French speaking country, opened vistas in the way the work of Hornby in Japan would lead to modern learner's dictionaries. Basnage was developing a clear linguistic model that analysis brings to the fore. He was interested in teaching the language and therefore developing means for production of language rather than just a description of language. Basnage was working to tight deadlines and so never had the opportunity to go further with the innovative changes he introduced and Brutel was clearly far more normative in his outlook. As work goes on, it will become clearer just how ahead of his time Basnage was.

## 6 References

- Académie Française (1694). *Le dictionnaire de l'Académie françoise, dédié au Roy. T. 2. L-Z. 1ère.* Paris: Vve J. B. Coignard. <http://catalogue.bnf.fr/ark:/12148/cb35153876f>.
- Accademia della Crusca (1612). *Vocabolario degli Accademici della Crusca.* Venezia: Giovanni Alberti. <http://www.accademiadellacrusca.it/en/digital-shelves/crusca-online>.
- Corneille, T. (1625-1709). (1694). *Le dictionnaire des arts et des sciences. Tome 1 / par M. D. C...* Paris: Vve J. B. Coignard. <http://gallica.bnf.fr/ark:/12148/bpt6k50507s>.
- Desroches, N. (1687). *Dictionnaire des termes propres de la marine.* Paris: Amable Auroy. <https://books.google.fr/books?id=HcoeZRLKsC&printsec=frontcover#v=onepage&q&f=false>.
- Diderot, D., d'Alembert (Le Rond), J. (1751). *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers.* 17 vol. Paris. <http://encyclopedia.uchicago.edu>.
- Fatio de Duillier, N. (1699). *Fruit-walls improved, by inclining them to the horizon, or, A way to build walls for fruit-trees whereby they may receive more sun shine, and heat, than ordinary / by a member of the Royal Society.* Ann Arbor: Text Creation Partnership, 2011. Edited by R. Everingham, J. Taylor, at the Sign of the Ship, in St. Paul's Church Yard. <http://name.umdl.umich.edu/A40990.0001.001>.
- Furetière, A. (1684). *Essais d'un dictionnaire universel.* Amsterdam: chez Henri Desbordes. <http://gallica.bnf.fr/ark:/12148/bpt6k575529>.
- . (1685). *Factum pour Messire Antoine Furetière, abbé de Chalivoy, contre quelques-uns de l'Académie française ([Reprod.]).* Amsterdam: chez Henri Desbordes. <http://gallica.bnf.fr/ark:/12148/bpt6k575420>.
- . (1690). *Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes et les termes des sciences et des arts.* La Haye, Rotterdam: Arnoud et Reinier Leers. <http://gallica.bnf.fr/ark:/12148/bpt6k50614b>.
- . (1701). *Dictionnaire universel, contenant généralement tous les mots françois tant vieux que modernes, & les termes des sciences et des arts. Tome 1 / ,... par feu messire Antoine Furetière,... 2e édition revue, corrigée et augmentée par M. Basnage de Bauval.* La Haye, Rotterdam: Arnoud et Reinier Leers. <http://gallica.bnf.fr/ark:/12148/bpt6k5841680f>.
- . (1859a). *Recueil des factums d'Antoine Furetière, de l'Académie françoise, contre quelques-uns de cette Académie ; suivi des preuves et pièces historiques données dans l'édition de 1694.* Edited by Charles Asselineau. Vol. Tome 1. Alençon: Poulet-Malassis et de Broise.
- . (1859b). *Recueil des factums d'Antoine Furetière, de l'Académie françoise, contre quelques-uns de cette Académie ; suivi des preuves et pièces historiques données dans l'édition de 1694.* Edited by Charles Asselineau. Vol. Tome 2. Alençon: Poulet-Malassis et de Broise.
- La Quintinie, de, J. (1690). *Instruction pour les jardins fruitiers et potagers. [Volume 1] / , avec un Traité des orangers, suivy de Quelques réflexions sur l'agriculture, par feu M. de La Quintinye,... Tome 1. [-II.].* Paris: Claude Barbin.
- Leca-Tsiomis, M. 1999. *Écrire l' "Encyclopédie". Diderot : de l'usage des dictionnaires à la grammaire philosophique.*



- Oxford: Voltaire Foundation.
- Proust, J. (1962). *Diderot et l'Encyclopédie*. Paris: Armand Colin.
- Rey, A. (2006). *Antoine Furetière : Un précurseur des Lumières sous Louis XIV*. Paris: Fayard.
- Richelet, P. (1680). *Dictionnaire françois : contenant les mots et les choses, plusieurs nouvelles remarques sur la langue françoise, ses expressions propres, figurées et burlesques, la prononciation des mots les plus difficiles, le genre des noms, le régime des verbes...* Geneva: J.-H. Widerhold.
- Trévoux. (1704). *Dictionnaire universel françois et latin, contenant la signification et la définition tant des mots de l'une et l'autre langue, avec leurs différents usages, que des termes propres de chaque estat et de chaque profession ; la description de toutes les choses naturelles et artificielles... ; l'explication de tout ce que renferment les sciences et les arts... Avec des remarques d'érudition et de critique...* E. Ganeau (Trévoux).
- Verdelho, T.d.S., Silvestre P.J. (2007). *Dicionarística portuguesa: inventariação e estudo do património lexicográfico*. Aveiro: Universidade de Aveiro.
- Williams, G. (2020). Architecture in the 1701 *Dictionnaire Universel*: Encoding and analysing architectural terminology with digital humanities methodologies. In A. Pano Alamán, V. Zotti (eds.) *The Language of Art and Culture Heritage: a Plurilingual and Digital Perspective*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 190-207.

### Acknowledgements

The authors are grateful to the support received from the Consortium Cahier, part of the national HumaNum infrastructure that allocated seed money enabling work to begin on the digitalisation of the 1701 *Dictionnaire Universel* and to the French National research council for the BasNum project. The BasNum project is a four year French nationally funded BasNum programme (ANR-18-CE38-0003-01). We are also grateful to Andrés Echevarría, a research student in the Master's in Digital Publishing at the Université de Bretagne Sud Lorient for his work in encoding the front matter which is now freely available on the website of the LiCoRN Research Group - <http://www.licorn-research.fr/Basnage.html>





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Terminology and Terminography**







# Term variation in terminographic resources: a review and a proposal

Cabezas-García M., León-Araúz P.

University of Granada, Spain

## Abstract

Term variation or the coexistence of different terms to name the same concept (e.g. *contamination* and *pollution*) is frequent in specialized language (Fernández-Silva 2018). Since variants are not always interchangeable, language users such as translators or terminologists need to know when and why a variant should be used in preference to another. Terminographic resources should facilitate this task by including different variants as well as the criteria guiding their selection. However, variants are not usually fully covered and, when they are included, indicators regarding semantics, pragmatics, or usage are not often provided. This paper investigates the representation of term variation in terminographic resources. Our goals were (i) to confirm whether term variants are underrepresented and usage indications are not usually provided, (ii) to collect the data categories and fields employed in the description of term variants, and (iii) to propose a model of representation of term variation in the terminological knowledge base EcoLexicon. Our results showed that, despite the prevalence of term variation, terminographic resources do not usually describe the different possibilities and/or the criteria guiding their selection. In contrast, those which attempt to add pragmatic information do not show this kind of data in a parameterized way.

**Keywords:** term variation; terminology; terminographic resources

## 1 Introduction

Term variation occurs when different designations are used to name the same concept (e.g. *environmental contamination* and *environmental pollution*). Although to a lesser extent than general language (Freixa 2006; Sanz-Vicente 2011), specialized discourse exhibits a considerable degree of term variation, which has been explored by means of corpora (Fernández-Silva 2018).

Discovering the causes or types of variation may reflect the mental processes involved in the selection of one term instead of another. Furthermore, this information is helpful to terminologists or translators since they need to know when and why one variant should be used in preference to another (Candel-Mora & Carrió-Pastor 2012). Accordingly, terminographic resources should reflect the different variants that may designate each concept as well as their conceptual and communicative implications, since this will affect linguistic productions.

Unfortunately, the different possibilities and the criteria guiding the selection of variants are not usually described (Kerremans 2017). Frequency alone cannot be the sole criterion of classification, since other motivations can be involved in term selection (León-Araúz & Cabezas-García *in press*). Therefore, those data should be enhanced with structural, semantic, pragmatic, and usage information in terminographic resources (Faber & León-Araúz 2016; Giacomini 2018). This would improve a sound use of variation in texts.

This paper explores how term variation is currently represented in terminological resources with a view to (i) confirming whether term variants are underrepresented and usage indications are not usually provided, (ii) collecting the data categories and fields employed in the description of term variants when represented, and (iii) proposing a model of representation that acknowledges the tendencies found and meets the needs in EcoLexicon (a terminological knowledge base on the environment). For this purpose, several multilingual terminological resources, generally aimed at translators (who need to handle variants properly), were examined (e.g. IATE, TERMIUM Plus, MuLex, VariMed), focusing on the presence of variants, their location, and the information provided.

The rest of this paper is organized as follows. Section 2 explains term variation and relevant aspects from the domains of Translation and Terminography. Section 3 analyzes the representation of term variants in terminographic resources. Section 4 presents a proposal for the description of term variation in EcoLexicon, and Section 5 summarizes the conclusions of this study and future research lines.

## 2 Term Variation in Translation and Terminography

Term variation, or the coexistence of more than one term to name the same concept, can be frequently found in general and specialized language. In both types of discourse, but particularly in the latter, it must be properly understood and treated for the sake of an accurate communication. It thus affects both comprehension and production stages in the translation workflow. On the one hand, different variants in the same text must be understood as pointers to the same concept. On the other hand, the deliberate or random use of variants must be identified and distinguished. Term variation is directly related to the notion of “equivalence”, which can be intra- or interlinguistic (i.e. term variants in the same language, and translations, respectively). While for terminologists, equivalence is present when two or more terms refer to the same concept, other linguists, such as translators, adopt a broader view of equivalence. They consider equivalence at the sentence or text level, rather than at the term level, and thus accept broader mechanisms as equivalent facilitators,



such as hypernymy or variants reflecting different conceptualizations. These views often underlie the notion of term variation in terminography, and consequently, its description in terminographic resources.

However, even when variants represent exactly the same concept and thus seem interchangeable, the use of one of them may be preferred depending on different contextual factors. Accordingly, ascertaining the reasons of variation can help us make the correct choice. As Freixa (2006) states, causes for term variation can be (1) dialectal, caused by different origins of the authors (*lift*, *elevator*); (2) functional, resulting from different communicative registers (*hepatoma*, *cancer of liver*); (3) discursive, due to different stylistic and expressive needs of the authors (*motor vehicle pollution*, *car pollution*); (4) interlinguistic, caused by contact between languages (*nursery school*, *kindergarten*); and (5) cognitive, resulting from different conceptualizations and motivations (*climate change*, *climate emergency*). Several of these reasons can also co-occur.

Term variation can thus have consequences in communication. When there is only a change in the form but not in the meaning, term variants do not have cognitive effects, as in *marine product* and *sea product*. On the contrary, these consequences occur when term variation implies a shift in perception, as in *sea product* and *fishing product* (Fernández-Silva et al. 2009). Its potential results should be considered when selecting a term variant.

Along these lines, Faber & León-Araúz (2016: 12-13) proposed a classification of term variants that encompasses different proposals and integrates the types, causes and consequences of term variation with a view to describing them in a terminological knowledge base (TKB). This inventory delves into the particularities of the different terms and specifies whether semantics or communicative situation are affected, thus facilitating term selection. The classification includes: (1) orthographic variants (*aesthetics*, *esthetics*); (2) diatopic variants (*groyne*, *groin*); (3) short form variants (*laser*, *Light Amplification by Stimulated Emission of Radiation*); (4) diaphasic variants (*H<sub>2</sub>O*, *water*); (5) dimensional variants (*Gutenberg's discontinuity*, *core-mantle boundary*); (6) metonymic variants (*water*, *sea*); (7) diachronic variants (*carbonic anhydride*, *carbon dioxide*); (8) non-recommended variants (*mental retardation*, *intellectual disability*); and (9) morphosyntactic variants (*wave action*, *the action of the waves*). Every category is further specified with additional levels (e.g. diaphasic variants can be scientific, informal, or domain-specific).

Since translators mostly use terminographic resources as their primary tool for finding equivalence, it would be useful for them to find variation-related information, which would accelerate and enhance their translations. However, the different variants that may designate a single concept are not always included in terminographic resources and, when they are, these are not usually enriched with usage information or data are not provided in a systematic fashion (e.g. variants may be lemmatized or not, they may appear in different microstructural positions, etc.) (Giacomini 2018). Different studies have highlighted the need for enhancing the representation of variants by means of semantic, pragmatic, and usage information. One of these studies is Louw's (1998), who argues that, in bilingual dictionaries, the different possible target terms should be accompanied by usage indicators of these variants since contextual constraints may apply. He also emphasizes the role of a coherent marking system of term variants. Precisely this marking system is analyzed in Mari (2017), who states that the tags used in lexicographic resources to account for variation are not often supported by an explanation of how these tags are used. He also highlights the need for systematizing these labels, while acknowledging the complexity of this task.

The importance of enriching the representation of variants for practical users, such as translators or specialized language learners, has also been addressed in Abekawa & Kageura (2008), who complain about the lack of accurate descriptions and exemplifications of variation in electronic terminological English-Japanese dictionaries, despite the large amount of these resources. Since this hinders the translators' tasks, they ask for more usage notes of term variants and develop a tool, which can be accessed from a translation aid system and explores term variants in context using the web. Although these variants are often not very frequent (there is no frequency threshold) and their occurrences are not verified (the system is automatic), as acknowledged by the authors, it is a promising system that may be helpful thanks to the high number of variants provided (mostly based on additions and reductions to the main term).

Along this same line, Alves Costa & Fernández-Silva (2018) explore explicit term variation in Brazilian dictionaries and also complain about its underrepresentation in terminological resources, due to the prescriptive tradition in terminology. They design a proposal for representing term variation in a specialized dictionary for learners, in which they claim that users should be able to: i) understand term variants; ii) distinguish the usage differences of every variant; iii) recognize the possible causes of variation; and iv) verify its cognitive consequences. To this end, the following information is provided: i) term; ii) part of speech; iii) definition; iv) knowledge-rich contexts; v) term variants (followed by an explanation of how variation occurs: relative synonym by inclusion, relative synonym by intersection, no synonym, abbreviation); and vi) explanatory notes. This valuable proposal can guide users when there is no total equivalence.

Giacomini (2018) devises a frame-based model for the representation of multiword terms and their variants in a technical e-dictionary. Aimed at facilitating text production in the native language, her proposal includes multiword terms and their variants, as well as an indication of the type of variation (e.g. partial morphological variation + syntactic variation), the variation pattern followed (e.g. paraphrase + explication + transposition), and usage contexts of every variant. This is also a useful representation that does not limit to the mere inclusion of variants.

Accordingly, Kruse & Giacomini (2019) make a proposal for an electronic specialized dictionary, focusing on synonymy relations from an orthographical, morphological and syntactic perspective. Entries will include definitions, abbreviations, equivalents, collocations, and usage examples, as well as information on semantically related terms (e.g. synonyms, antonyms or hyponyms). They claim that, to facilitate text production, terminological resources should include term variants accompanied by information such as the type of text where they are found and its author.

Other studies that have focused on term variation are Janssen's (2006), who explores the representation of orthographic variation, usually presented in the form of cross-references; as well as different works by Freixa (2006), León-Araúz (2015, 2017), Fernández-Silva (2016, 2018), Daille (Daille 2017; Hazem & Daille 2018), and L'Homme (2020), among



other authors, on different aspects of term variation.

In conclusion, the representation of the different term variants in terminographic resources becomes central in descriptive settings, contrary to what has traditionally been done. Evidently, users, such as translators, need to know when to use each variant as well as its conceptual and communicative implications, since this will affect the receiver's interpretation of the message. Otherwise, they can actually over-standardize, creating consistency in places where the use of variants was deliberate and well-reasoned (Bowker & Hawkins 2006: 80). Consequently, besides describing different types of variants, which is undoubtedly important, the added value of a linguistic resource lies in the enhancement of those data with additional information, such as semantic, pragmatic, and usage aspects. This thorough representation demands an in-depth analysis as well as a homogenization effort with a view to providing enriched, consistent data.

### 3 Exploring the Representation of Term Variants in Terminographic Resources

In order to explore how term variation is usually represented in terminographic resources, several publicly available multilingual resources were examined (i.e. IATE, TERMIUM Plus, MuLex, VariMed). Different aspects were analyzed: (i) the presence of term variants and their scope (e.g. what is considered a term variant and how many of them are included in each entry); (ii) their location and prominence in the resource (e.g. whether term variants are described with the same detail as main entry terms); (iii) the information provided for each variant (e.g. type of usage-related constraints included in their description). The analysis was carried out by browsing through term entries where an indexed list was available (i.e. TermiumPlus, MuLex and VariMed) or by searching for a list of terms that were likely to show variants in the domain(s), mostly based on previous knowledge of the concepts (i.e. IATE). The resources analyzed are all aimed at translators, they represent different domains (Law in MuLex and Medicine in VariMed) or multiple domains (i.e. IATE and Termium Plus) are developed by researchers or institutions and are bilingual or multilingual, where English is the only common language. They thus cover different settings in term management and provide an overall view on different ways of dealing with term variation.

#### 3.1 IATE

IATE (<https://iate.europa.eu/>; Fontenelle & Rummel 2014) is the TKB of the European Union, which includes terms of multiple specialized domains in its official languages. IATE is a concept-oriented resource, that is, entries represent concepts (term variants are thus presumably to be found in the same entry). The entry level in IATE (i.e. the top level, which includes information that applies to the concept and, thus, the whole entry) shows the following data categories: i) concept ID; ii) domain(s); iii) domain note; iv) owner (e.g. Council, European Parliament); v) primary entry (in the case of duplicates); vi) origin (for country-specific concepts); vii) origin note (only in English); viii) life-cycle (historical, proposed or abandoned); ix) cross-references; and x) attachments (documents or graphics). Then, the language level, which holds information relevant to a language, includes the following fields: i) language code; ii) anchor language (according to which all other languages will be attached, usually the source language of the text where the term occurred); iii) definition; iv) definition reference; v) language note; vi) language note references; and vii) attachments. Finally, the term level in IATE (i.e. the basic level, which holds data that is specific to a term) shows the following categories, which are mostly optional and should be used to describe term variants, as will be analyzed below: i) terms; (ii) term type (term, abbreviation, phrase, formula, short form, lookup); iii) evaluation (preferred, admitted, deprecated, obsolete); iv) term reference; v) reliability (not verified, minimum reliability, reliable, very reliable); vi) note; vii) context (i.e. an excerpt); viii) context reference; ix) language usage; x) regional usage; xi) customer (e.g. Translation Centre, European Environment Agency); and ix) grammatical information (part of speech, gender, and number).

After reviewing the structure of this database, we focused on its representation of term variants. To this end, the presence of term variants and their scope was first analyzed. Overall, a "strict" notion of term variation is observed in IATE, that is to say, variants convey the same concept whereas terms conveying slightly different concepts (e.g. hypernyms) are not usually considered term variants, with a few exceptions. Some examples of this vision are *air pollution* and *atmospheric pollution*, which are represented as different denominations of the same concept. However, in some cases a broader vision is adopted. For instance, the term *air pollution episode* is accompanied by its variants *pollution episode* and *episode*, which, strictly speaking, do not convey exactly the same concept although translators could resort to them for stylistic reasons. Nevertheless, it can be concluded that term variants in IATE are most often conceptually equivalent, as confirmed by Kerremans (2015, 2017). As for the number of variants included in IATE entries, we can consider it to be intermediate. In other words, although variants are not extensively covered (as can be the case in TERMIUM Plus, see below), there is a reasonable inclusion of them, with most entries including at least two different denominations in every language.

Regarding the location and prominence of term variants, these appear at the term level (i.e. inside every concept entry, in the language in question) and are described with the same detail as main entry terms. This includes the term-level fields presented above, such as reliability, evaluation, term level note and regional usage, among others. Usage-related constraints can be found in these fields, which must be highlighted since, as mentioned above, this is not the norm. Different types of information are provided: from the evaluation of the term (e.g. preferred, admitted, deprecated) to its geographic details (e.g. UK, internationally), as well as other data, which are very useful in comprehension and production tasks.

However, this information is not provided whenever there are term variants, it is distributed among different fields (despite the fact that a predetermined and thorough set of data categories is available) and it is not always systematically organized. Although this does not prevent users from finding the details, a consistent organization is considered to enhance user experience. For example, some variants (e.g. symbols) are not presented in the *term* field, but in other data



categories, such as the *term level note*. This occurs, for instance, with the *O2* symbol. Additionally, certain types of usage-related information do not appear in the corresponding field. This is the case, for example, in the French term *acide carbonique*, which is said to be an inappropriate denomination, as well as in the French term *mongolisme*, which is considered obsolete and thus not recommended. These evaluation-related details, instead of appearing in the *evaluation* field, are presented in the *term level note* category. Accordingly, the Spanish term *anhídrido carbónico* is also considered obsolete and not recommended, which again is not indicated in the *evaluation* field; on the contrary, this time the *regional usage* field is used.

Sometimes, different types of indications are provided in the same data category. For instance, the *language usage* field includes regional information such as the following on *Down's Syndrome*: *Although "Down's Syndrome" is still commonly used in the UK, "Down syndrome" is becoming prevalent internationally and is also found in the UK*. In other cases, this field includes conceptual information that would certainly be more appropriate in this field, such as the following on *ocular ulcer*: *The term ocular ulcer is seldom used when referring to animals, although it occurs in CELEX:32006R1950/EN where it is used for horses. When referring to humans, this term is used to denote a broader concept, encompassing both corneal and eyelid ulcers*. In conclusion, the accurate description of term variation in IATE must be acknowledged. However, even if IATE offers many options to record different types of variants and specify their use, many of the fields are left empty (Kerremans 2015, 2017). It would be ideal if these rich details were provided in as many variants as possible since this would help make better linguistic decisions. Besides, an ordered and homogenous distribution would also be beneficial in such a useful resource.

### 3.2 TERMIUM Plus

TERMIUM Plus (<https://www.btb.termiumplus.gc.ca/>) is the TKB of the Government of Canada, which represents millions of concepts from different specialized domains in English, French, and to a lesser extent Spanish and Portuguese. It is one of the largest TKBs in the world. It is also concept-oriented, although the three-level structure is not as clear as in IATE, since some conceptual (i.e. subject field) and term-related information (i.e. usage observations) are stored at the language level.

There is no information at the entry level. At the language level, available categories are what they call textual support: i) definition; (ii) context (where the terms are employed in official sources, especially the main entry term); (iii) observation (where more information is provided regarding the concept or terms); and (iv) phraseologism (for collocations); and v) key terms, which cover spelling or semantic variants and masculine, feminine, singular or plural forms. It is striking that these fields are shown at the language level since, except for the definition, all of them describe the particular use of terms. It is also surprising that key terms are used to include spelling or semantic variants, instead of including them as full-fledged term entries. This is the case of the pollution equipment French entry, which includes four terms as variants (*matériel antipollution*, *matériel de dépollution*, *matériel de lutte contre la pollution* and *équipement antipollution*) and five more as key terms (*matériel anti-pollution*, *équipement de lutte anti-pollution*, *équipement de lutte contre la pollution*, *équipement anti-pollution* and *équipement de lutte antipollution*).

At the term level, available categories are the following: (i) acceptability rating (correct, avoid, unofficial, no rating); (ii) temporal labels (former name, archaic, obsolete); (iii) origin (chemical abstracts service number, classification system code, form code, ISO item number, occupational code, publication code, formula, latin, legal origin, trademark, and proposals, which are equivalents proposed by terminologists, translators or specialists but cannot be found in a written source); iv) linguistic parameters (anglicism, barbarism, calque, deceptive cognate, pleonasm); v) reference (pointing to the field of observation); vi) parts of speech; vii) gender; viii) number; ix) geographic parameters (Africa, Antarctica, Canada, France, Great Britain, NATO, intergovernmental, international, regional); x) frequency (less frequent, rare); xi) sociolinguistic parameters (familiar, jargon); xii) semantic parameters (generic, pejorative, specific); and xiii) official status parameters (standardized, officially approved).

Regarding the presence of variants, Termium Plus covers a wide range of variants per entry. For instance, the entry of *photochemical smog* contains, as opposed to IATE, which does not include any other English variant, three different variants in English (*oxidant smog*, *Los Angeles smog* and *photochemical oxidant smog*) and four in French (*smog photochimique oxydant*, *smog oxydant*, *brouillard photochimique oxydant* and *brouillard photo-oxydant*).

As for their scope, TermiumPlus has a broader vision on term variation, as the field *semantic parameters* (generic, pejorative, specific) implies. For example, many entries include as term variants a hypernym, as those of *environmental pollution* or *water pollution*, which also include *pollution* as a variant (although it is not accompanied by the label *generic*).

In terms of location and prominence, variants appear at the term level and are described in the same detail as main entry terms, with the exception of those included as key terms, as previously mentioned. The preferred term is displayed first followed by variants, including abbreviations.

The description of variants is scattered through different data fields. This is only natural if we take into account the fact that there are different variation parameters and that the description of term variation sometimes happens at the term level and some others by comparison to others. For instance, information related to the geographic origin of a term only needs a geographic label assigned to the term. In contrast, semantic parameters or certain usage constraints depend on the comparison of several variants in different contexts.

Variation parameters at the term level include status (acceptability and official parameters), source (origin), geography (geographic parameters), time (temporal labels), term formation devices (linguistic parameters), frequency, diaphasic parameters (sociolinguistic parameters), and semantic distance (semantic parameters).

However, as was the case in IATE, not all of these fields are filled in each entry. Moreover, there seems to be a certain



overlap among these fields. Acceptability is related to the status of terms (correct, avoid, unofficial), but official status parameters seem to point in the same direction (standardized, officially approved). The same happens with geographic parameters and origin. For instance, within geographic parameters Termium Plus includes *intergovernmental* or *international*, which could be considered origin parameters. Likewise, within origin parameters we can find *proposal*, which would actually be related to acceptability or source.

The most interesting field disambiguating the use of term variants is that of observation, at the language level. For instance, in the *pollution* entry, two observations are found: (1) *Pollution is very often used in the more general sense of "environmental pollution".* (2) *In water pollution, the term contamination is often used as a synonym (...) However, a distinction should be made between "pollution" and "contamination": the latter implies a health danger (particularly to humans).* However, this information is not included in the *water pollution* entry.

Nevertheless, this field is not always included nor it always deals with variation. The problem lies in the fact that the information contained in this field varies in nature and is not systematized. Sometimes it includes conceptual information and some others information related to the terms. For example, in the French entry of *pollution abatement*, two of these types of observations are made: (1) *Le terme français "abatement", dans le domaine de la pollution, est extrêmement douteux: il n'est attesté dans aucun ouvrage spécialisé que nous possédons* (term usage); (2) *Les nombreux moyens de lutte contre la pollution visent comme objectifs quantitatifs soit la suppression totale des polluants (dépollution), soit leur réduction. D'autres mesures antipollution visent plutôt des objectifs qualitatifs comme l'élimination sélective des polluants, par exemple, les plus dangereux ou les plus cancérigènes* (conceptual expansion).

When this information concerns the use of one particular term variant, the label *see observation* is included at the term level, but several observations may be included in the entry. The user must then guess which observation was referred to. We must acknowledge the considerable effort made by this resource to represent term variation in terms of coverage and thoroughness. It is an extremely valuable TKB, but the main drawbacks regarding term variation are related to the lack of information (not all fields are filled) and systematicity (some fields overlap and different types of information are included within a single field, such as that of observation).

### 3.3 MuLex

MuLex (<http://mulex.altervista.org/>) is a TKB that describes English and Italian terms related to the legal area of victims of crime. The TKB contains a total of 149 English terms and 197 Italian terms. It was developed in Peruzzo (2012), a study that pays special attention to term variation. MuLex is also concept-oriented, although when non main terms are selected, language-level information is not shown, so it can be viewed as a hybrid approach. Moreover, the searches are based on terms rather than concepts. At the entry level, data categories are: i) subject; ii) subfield; iii) concept field; iv) concept map. At the language level, data categories include i) definition; ii) term variants; iii) notes on term variation; iv) equivalent term; v) note. Finally, the following fields are included at the term level: i) part of speech; ii) usage label (main term, uncommon); iii) category (graphic variant, short form, full form, graphic variant, initials); iv) regional label (UK, EU, CoE); v) style label (official, official EU); vi) lexica (found in IATE); vii) legal system (UK, US); viii) synonymy degree (~, <), only for entries which are not main terms; ix) phraseology (e.g. to claim, to obtain compensation from the offender); x) grammar (e.g. the term was only found in its plural form); xi) context (i.e. an excerpt).

The fact that notes on term variation are at the language level, and only shown when the main term is searched for, implies that these notes are included when different variants are compared. Information characterizing variants individually are included in the fields of usage label, category, regional label, style label, legal system and context. The synonymy degree, when included, is always used with regards to the main term, but not to the rest of the variants. Therefore, the main term has a privileged role in the TKB over the term variants. In contrast, the number of variants included in MuLex is high and their scope is broad. For instance, the entry of *victim* includes variants such as *victim of criminal conduct*, *victim of a crime*, *victim of the offence* and *crime victim*. They are accompanied by the following note: *When used in national texts, the terms "victim", "victim of a crime", "victim of the offence" and "crime victim" refer to national legislation and can therefore be considered to have a narrower meaning compared to the meaning they have in EU texts. The same holds for "victim of criminal conduct", which is only used in national texts and has a narrower meaning compared to the main term "victim", since the mentioned criminal conduct constitutes an offence specifically regulated by British law.*

These notes are extremely useful, especially considering the diverse nature of legal entities depending on the legal system and/or the institutions producing the texts. They always deal with issues related to conceptual asymmetries, vagueness, and regional differences, which, in this domain, are of paramount importance when disambiguating the use of variants, both in comprehension or production tasks. The problem is that these notes are not found in all entries where several variants coexist. For instance, the entry of *mediation in criminal cases* has nine different variants but no notes on variation are found. In those cases, only usage, regional, style and lexica labels may help to discriminate the use of each variant.

### 3.4 VariMed

VariMed (<http://varimed.ugr.es/>) is a TKB developed in the University of Granada and other institutions (Tercedor-Sánchez et al. 2014), which includes terms on the medical domain in English and Spanish, focusing on term variation as a cognitive and communicative phenomenon. VariMed is also concept-oriented but with a particular focus on the description of terms. At the entry level data categories are: i) definition; ii) concept type; iii) related concepts; iv) organs affected; v) images. There is no language-specific information and at the term level data categories are: i) part-of-speech; ii) language; iii) register (formal, informal, jargon, neutral, children); iv) conceptual dimension highlighted (affects, agent/result, visual attribute, non-visual attribute, composed of, time, intensity, discoverer, body part, geography, metaphor/metonymy), v) geographical use (UK, US), vi) familiarity degree (a Likert scale); related variants



(abbreviated form of, full form of, often confused with) and vii) other labels (abbreviation, short form, French calque, English calque, eponym, false friend, misspelling, neologism, ICD-10 nomenclature, MeSH nomenclature, greek/latin origin, English borrowing, borrowing from other languages, acronym, culture-specific term, frequent term, non-recommended term, obsolete term, orthographic variant).

Most fields at the term level are aimed at the description of term variants (i.e. register, conceptual dimension, geographical use, familiarity, and other labels). As opposed to Termium Plus, where overlapping fields were found, in VariMed different types and causes of variation converge in an all-purpose field, that of other labels. In this field term formation devices coexist with other parameters such as frequency, time, origin or reliability.

VariMed includes a myriad of variants in both English and Spanish. For instance, type 2 diabetes has up to 7 terms in English (*adult-onset diabetes*, *diabetes mellitus type 2*, *NIDDM*, *non-insulin dependent diabetes mellitus*, *non-insulin-dependent diabetes mellitus*, *T2DM*, *type 2 diabetes*) and 9 terms in Spanish (*diabetes insulinoresistente*, *diabetes mellitus de inicio adulto*, *diabetes mellitus de inicio lento*, *diabetes mellitus estable*, *diabetes mellitus no insulino dependiente*, *diabetes mellitus resistente a la cetosis*, *diabetes mellitus tipo 2*, *DM-2*, *DMNID*).

It follows a "strict" vision of variation (i.e. no hypernyms are included), but it shows a wide range of variants in terms of register (i.e. many informal variants and jargon), and morphosyntactic variants are included as full-fledged term entries. They are thus all described with the same level of detail although, as what seems to be a trend, not all fields are always filled. For example, among all variants of type 2 diabetes, *adult-onset diabetes* is categorized as neutral from the register viewpoint, *diabetes mellitus type 2* is categorized as formal and pertaining to the MeSH nomenclature, *NIDDM* is categorized as a formal acronym, *non-insulin dependent diabetes mellitus* as a formal orthographic variant, *non-insulin-dependent diabetes mellitus* as a formal variant pertaining to the MeSH and ICD-10 nomenclatures, *T2DM* as a formal acronym and *type 2 diabetes* as formal. Therefore, more information would still be needed for the disambiguation of all seven variants for text production purposes.

### 3.5 Summary

Despite the prevalence of term variation, terminographic resources do not usually describe the different possibilities and/or the criteria guiding their selection. In contrast, those which attempt to add pragmatic information do not show this kind of data in a parameterized way. For instance, IATE does not always include this information consistently (i.e. the same kind of data are shown in different fields). TERMIUM Plus shows a rich selection of term variants in many entries. Furthermore, in the field Observations it includes useful information regarding the use of term variants. However, it would benefit from a more systematic approach, since the data provided are not structured around a predetermined set of data categories, they are included at the language level and different types of information (i.e. conceptual and usage-related) are included in the same field. Something similar happens with MuLex, where the most interesting feature disambiguating term variants is placed at the language level but cannot always be found. VariMed features multiple labels to characterize term variants and includes a wide range of them, from specialized nomenclatures to colloquial forms, but they are not always systematically described.

Commonly found data categories are those related to register, reliability, time, origin, geographical use, formation device or status, but the most useful information in terms of usage disambiguation is often found in the form of notes.

## 4 Representing Term Variation in EcoLexicon: a Proposal

To improve the current description of term variation in EcoLexicon (a comprehensive list of variants but only occasionally described in the form of notes), we designed a preliminary, enhanced model of representation of term variants in this TKB (León-Araúz et al. 2020). This proposal incorporates and improves the tendencies found in other terminographic resources, as well as new approaches, and presents them in a consistent, systematic way. It was tested on multiword terms (e.g. *doubly-fed induction generator*), which are frequent combinations in specialized discourse that are especially prone to variation (Cabezas-García *in press*).

The internal organization of EcoLexicon is essential to understand this model of representation. Based on the TBX standard, EcoLexicon is also structured in three levels: (1) entry level, (2) language level, and (3) term level. When representing term variation in a TKB, terminographers need to decide at which level they will record each type of information. Previous classifications of term variation are not specifically conceived for the design of a TKB, because the patterns observed refer to both the description of a single term (e.g. borrowings, scientific name, etc.) or the description by comparison to a particular form (e.g. reductions, lexical changes, etc.). Therefore, from the types, causes and consequences of variation analyzed in León-Araúz & Cabezas-García (*in press*) and the data categories found in the resources analyzed in this paper, a set of descriptive fields was devised. Some of them are included in the description of individual terms (i.e. term level) and some others are a set of criteria according to which all term variants of a concept could be grouped and compared (thus, at the language level).

### 4.1 Variation Fields at the Term Level: Term Entries

Term entries in EcoLexicon contain the following fields so far: language, term type (main entry term, variant, acronym), part-of-speech, gender, and note. However, for an accurate representation of term variation, other values and fields needed to be added.

Table 1 shows the structure of a new term entry proposal for the TKB, including data categories and their values (their type and possible options, whether they are mandatory or optional and whether they admit single or multiple values).



Data category	Values
Term type	Picklist ( <i>main term, variant</i> ); single value, mandatory
Formation device	Picklist ( <i>borrowing, adapted borrowing, calque, blending, acronym, abbreviation, formula, symbol, eponym, culture-specific</i> ); multiple values, optional
Source	Free text (e.g. UN, corpus EurLex); multiple values; optional
Use_geographical	Free text (e.g. Spain, Mexico, Australia, etc.); multiple; optional
Use_status	Picklist ( <i>admitted, deprecated, standardized, non-recommended</i> ); single value, optional
Use_register	Picklist ( <i>scientific name, jargon, formal specialized, formal semi-specialized, informal</i> ); single value; optional
Use_context	Free text; multiple values; optional
Use_translation context	Free text; multiple values; optional
Notes	Free text; multiple values; optional

Table 1: Data fields at the term level.

Figures 1 and 2 illustrate how the new fields would describe at the term level the Spanish terms *ozono a nivel del suelo* and *ozono troposférico*, two variants of *tropospheric ozone*, also known as *ground-level ozone*, *surface ozone*, and *low level ozone*.

Figures 1 and 2: Term entries for *ozono a nivel del suelo* and *ozono troposférico*.

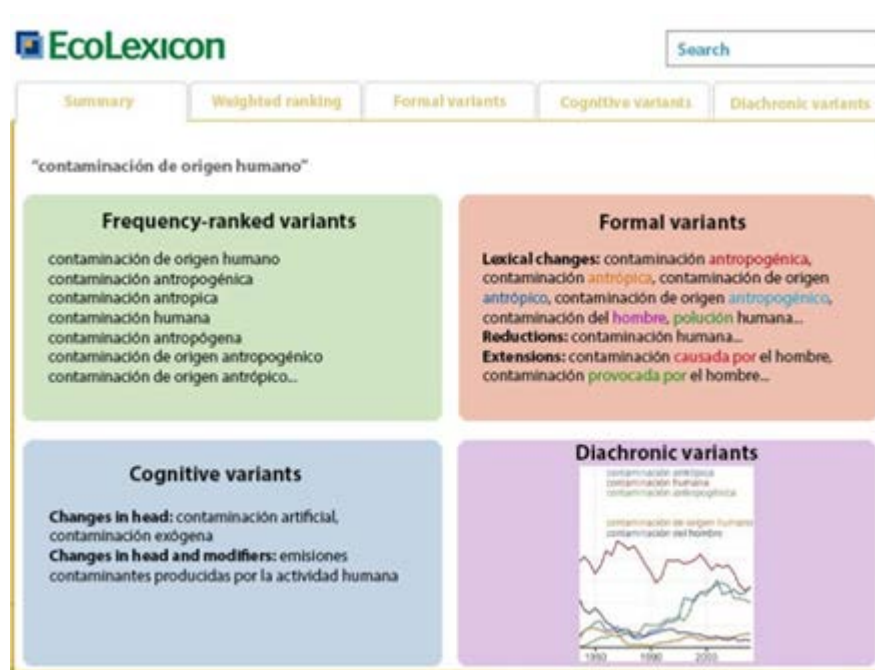
The Use\_context includes any information about the nuances that a particular variant may have in comparable corpora (i.e. a context of original production rather than translation), whereas the Use\_translation context is filled when clear patterns are found regarding the asymmetries of equivalence in parallel corpora (i.e. translations).

For example, although *ozono troposférico* is clearly the most frequent Spanish variant designating TROPOSPHERIC OZONE, in the comparable corpora, *ozono a nivel del suelo* and *ozono superficial* seem to be preferred when the term is related to human health issues. In turn, in the parallel corpora, we found that while *ozono troposférico* was usually translated by *tropospheric ozone* or *ground-level ozone*, *ozono a nivel del suelo* and *ozono superficial* clearly preferred *ground-level ozone*, even though it was exactly the same concept. Consequently, the field Use\_translation context allows us to establish interlinguistic variation preferences whereas the field Use\_context serves the same purpose for intralinguistic variation.

## 4.2 Variation Fields at the Language Level: Contrastive Views

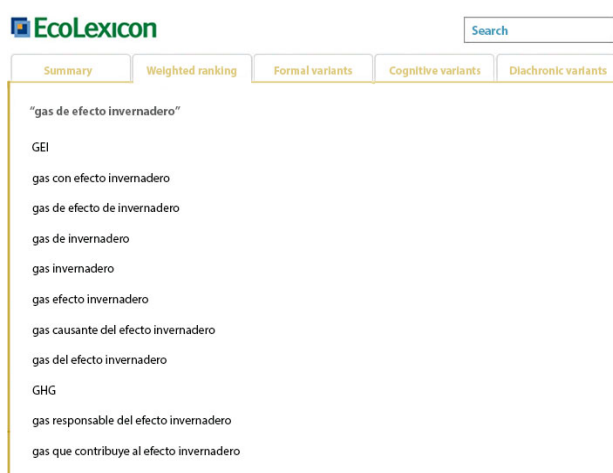
Regarding the criteria for the grouping and comparison of all term variants of a concept, which is essential for proper language production, a new module in EcoLexicon was devised (Figures 3-7). This module enhances the representation of term variants by grouping them together and highlighting their differences based on frequency, meaning, form, and usage trends over time. For this reason, it belongs to the language level, since information is not term-centered. It is divided into five tabs.



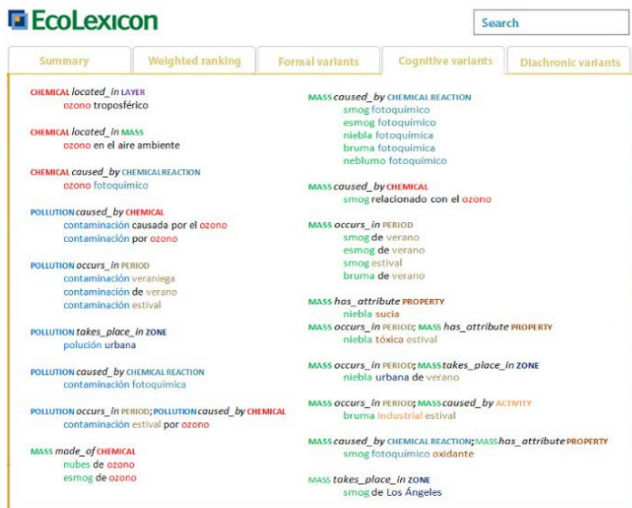
Figure 3: Summary view for *anthropogenic pollution* Spanish variants.

The first tab contains a summary of the comparison with regards to a previously established main term. Figure 3 summarizes the variants of *contaminación de origen humano* (*anthropogenic pollution*). Figures 4-7 show different types of variant classification: frequency-ranked variants, formal variants, cognitive variants, and diachronic variants. All examples are illustrated with different groups of term variants that best exemplify the approach.

Figure 4 ranks the most representative term variants of *gas de efecto invernadero* (*greenhouse gas*), according to a procedure developed in León-Araúz et al. (2020), which identifies the most established variants of a term. Figure 5 classifies term variants of *emisión de gases de efecto invernadero* (*greenhouse gas emission*), based on the type of formal changes as compared to the main term, highlighting their differences. In this case, only morphological and morphosyntactic changes, reductions and extensions apply, but other classifying parameters could also be used, such as lexical or graphical changes. Figure 6 shows the term variants of *smog fotoquímico* (*photochemical smog*), in regard to the conceptual categories and semantic relations codified in every term. Finally, Figure 7 depicts the term variants of *agotamiento del ozono* (*ozone depletion*) in a diachronic graph drawn from the Google N-gram viewer.

Figure 4: Frequency-ranked *greenhouse gas* Spanish variants.Figure 5: Formal *greenhouse gas emission* Spanish variants.



Figure 6: Cognitive *photochemical smog* Spanish variants.Figure 7: Diachronic view for *ozone depletion* Spanish variants.

## 5 Conclusions

Properly handling term variation is essential in order to understand and produce quality texts. Terminographic resources play a major role in this respect, as the linguistic resources most used for these tasks. Nevertheless, when resorting to them, users often find, at best, various synonyms with no indication (or unstructured data) on which term should be used in a particular context; or at worst, a lack of coverage of the different forms of naming the same concept. The resources analyzed (i.e. IATE, TERMIUM Plus, MuLex, and VariMed) were not among the worst-case scenarios, since they satisfactorily cover term variants and often provide indications on their usage. However, an in-depth analysis revealed some inconsistencies, which shows that there is still room for improvement regarding the representation of term variation.

Apart from the different views of term variation (some are broad, considering variants terms that do not convey exactly the same concept; other are stricter, acknowledging variation only when the same concept was evoked), a different coverage of term variants was observed (which is in line with the variation scope adopted). Since these resources were mostly concept-oriented, term variants were included in concept entries, and were usually described with the same detail as main entry terms. Nevertheless, it is the information provided for each variant, as well as the data categories chosen, that differs most among these resources. Their inclusion of pragmatic information about variants is helpful, although this does not appear for every single variant contained in the resource. Different fields and data categories are obviously used in every resource. In addition, data are not always systematically presented (sometimes the same information is presented in distinct fields), which could hinder access to information. Finally, a new model of representing term variation was devised in the TKB EcoLexicon (León-Araúz et al. 2020), which provides users with more usage-related, consistent information (the weakness of most terminographic resources). What remains to be done is to fill the new fields for all variants in the TKB, which will undoubtedly be a time-consuming task.

## 6 References

- Abekawa, T., Kageura, K. (2008). QRcep: a term variation and context explorer incorporated in a translation aid system on the web. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress, Euralex 2008, 15-19 July 2008*. Barcelona: Universitat Pompeu Fabra, pp. 915-922.
- Alves Costa, L., Fernández-Silva, S. (2018). A variação denominativa explícita na Lexicografia no Brasil: pressupostos para a organização microestrutural do Dicionário de Lexicografia Brasileira. In *Ibérica*, 36, pp. 93-118.
- Bowker, L., Hawkins, S. (2006). Variation in the organization of medical terms. Exploring some motivations for term choice. In *Terminology*, 12(1), pp. 79-110.
- Cabezas-García, M. (In press). *Los términos compuestos desde la Terminología y la Traducción*. Berlin: Peter Lang.
- Candel-Mora, M.A., Carrió-Pastor, M.L. (2012). Corpus analysis: a pragmatic perspective on term variation. In *RESLA (Revista Española de Lingüística Aplicada)*, 2012, pp. 33-50.
- Daille, B. (2017). *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. Amsterdam: John Benjamins.
- Faber, P., León-Araúz, P. (2016). Specialized knowledge representation and the parameterization of context. In *Frontiers in Psychology*, 7(196), pp. 1-20.
- Fernández-Silva, S. (2016). The Cognitive and Rhetorical Role of Term Variation and its Contribution to Knowledge Construction in Research Articles. In *Terminology*, 22(1), pp. 52-79.
- Fernández-Silva, S. (2018). The cognitive and communicative functions of term variation in research articles: a comparative study in Psychology and Geology. In *Applied Linguistics*, 2018, pp. 1-23.
- Fernández-Silva, S., Freixa, J., Cabré, M.T. (2009). The multiple motivation in the denomination of concepts. In *Journal*



- of *Terminology Science and Research*, 20, pp. 1-24.
- Fontenelle, T., Rummel, D. (2014). Term Banks. In P. Hanks, G.-M. De Schriver (eds.) *International Handbook of Lexis and Lexicography*. Berlin-Heidelberg: Springer, pp. 1-12.
- Freixa, J. (2006). Causes of Denominative Variation in Terminology: A typology proposal. In *Terminology*, 12(1), pp. 51-77.
- Giacomini, L. (2018) Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary. In N. Calzolari et al. (eds.) *Proceedings of the XVIII EURALEX International Congress, Euralex 2018, 17-21 July 2018*. Ljubljana: EURALEX, pp. 309-318.
- Hazem, A., Daille, B. (2018). Word Embedding Approach for Synonym Extraction of Multi-Word Terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, 7-12 May 2018*. Miyazaki, Japan: ELRA, pp. 297-303.
- Janssen, M. (2006). Orthographic variation in lexical databases. In E. Corino, C. Mareello, C. Onesti (eds.) *Proceedings of the XII EURALEX International Congress, Euralex 2006, 6-9 September 2006*. Turin: Alessandria, Edizioni dell'Orso, pp. 167-172.
- Kerremans, K. (2015). Managing terminological and translational diversity in parallel corpora: a case study in institutional translation. In *InTRAlinea: Online Translation Journal*, 2015.
- Kerremans, K. (2017). Towards a resource of semantically and contextually structured term variants and their translations. In P. Drouin, A. Francœur, J. Humbley, A. Picton (eds.) *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins, pp. 83-108.
- Kruse, T., Giacomini, L. (2019). Planning a Domain-specific Electronic Dictionary for the Mathematical Field of Graph Theory: Definitional Patterns and Term Variation. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference, 1-3 October 2019, Sintra*. Brno: Lexical Computing CZ, s.r.o., 676-693.
- L'Homme, M.C. (2020). *Lexical Semantics for Terminology: An Introduction*. John Benjamins.
- León-Araúz, P. (2015). Term variation in the psychiatric domain: transparency and multidimensionality. In P. ten Hacken, R. Panocová (eds.) *Word Formation and Transparency in Medical English*. Newcastle-upon-Tyne: Cambridge Scholars Publishing, pp. 33-54.
- León-Araúz, P. (2017). Term and Concept Variation in Specialized Knowledge Dynamics. In P. Drouin, A. Francœur, J. Humbley, A. Picton (eds.) *Multiple Perspectives on Terminological Variation*. Amsterdam: John Benjamins, pp. 213-258.
- León-Araúz, P., Cabezas-García, M. (In press). Term and translation variation of multi-word terms. In *MonTI: Monografías de Traducción e Interpretación*.
- León-Araúz, P., Cabezas-García, M., Reimerink, A. 2020. Representing Multiword Term Variation in a Terminological Knowledge Base: a Corpus-Based Study. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation, LREC 2020, 11-16 May 2020*. Marseille, France: ELRA, pp. 2351-2360.
- Louw, P. A. (1998). Synonymy in the Translation Equivalent Paradigms of a Standard Translation Dictionary. In *Lexikos*, 8, pp. 173-182.
- Mari, I. (2017). Notes on the treatment of variation in the Diccionari català-valencià-balear (DCVB). In *Dialectologia*, special issue VII, pp. 113-131.
- Peruzzo, K. (2012). Terminological Equivalence and Variation in the EU Multi-level Jurisdiction: A Case Study on Victims of Crime. PhD thesis. Università degli Studi di Trieste, Trieste, Italy.
- Sanz-Vicente, L. (2011). Análisis contrastivo de la terminología de la teledetección. La traducción de compuestos sintagmáticos nominales del inglés al español. PhD thesis. Universidad de Salamanca, Salamanca, Spain.
- Tercedor-Sánchez, M., López-Rodríguez, C.I., Prieto Velasco, J.A. (2014). También los pacientes hacen terminología: retos del proyecto VariMed. In *Panace@: revista de Medicina, Lenguaje y Traducción*, 25(39), pp. 95-103.

## Acknowledgements

This research was carried out as part of project FFI2017-89127-P, Translation-Oriented Terminology Tools for Environmental Texts (TOTEM), funded by the Spanish Ministry of Economy and Competitiveness.



# Revisiting Polysemy in Terminology

L'Homme M.-C.

Université de Montréal, Canada

## Abstract

For many, the success of specialized communication is achieved when it is devoid of ambiguity. However, polysemy is quite common in specialized corpora and needs to be managed when compiling domain-specific resources. In this paper, we show that polysemy affects many lexical items in specialized texts and review specific cases of polysemy, some of which are seldom discussed in terminology literature. We also show how different types of polysemy can be handled in terminological resources. Methods include: 1. accounting for meaning distinctions using well known tests in lexical semantics; 2. representing links and differences between meanings with lexical relations and labelled argument structures. We also explain how Frame Semantics (Fillmore (1982) and the methodology used in the FrameNet project (Ruppenhofer et al. 2016) can provide a broader perspective on meaning distinctions in specialized fields. Methods are applied to examples found in the English versions of two terminological dictionaries in the fields of computing and the environment.

**Keywords:** terminology; polysemy; predicative units; alternation; terminological resource; semantic frame

## 1 Introduction

For many, the success of specialized communication is achieved when it is devoid of ambiguity. This is why some approaches to terminology seek to reduce polysemy as much as possible using standardization methods or by creating new names to distinguish two concepts. These methods are often necessary in knowledge representation systems in which concept classes are mutually exclusive.

Paradoxically, looking up a single-word term in a term bank can produce a surprisingly high number of hits. For instance, the noun *file* appears in 20 different records in *Le grand dictionnaire terminologique* (2020) and 15 different terms records in *LATE* (2020) and *Termium* (2020) (other term records deal with verb senses). Some records describe the same meaning, i.e. “a collection of related information in electronic form,” but then *file* is associated with different subject fields: information technology, documentation, management, etc. To account for this phenomenon, terminologists have conveniently redefined the concept of “homonymy” to refer to polysemy across domains.<sup>1</sup> Traditionally, terminologists have distinguished “homonymy” (multiple meanings in different domains) from “polysemy” per se (multiple meanings in the same domain) (Felber 1987).

Polysemy, even when it is considered within specialized domains, is quite common and the topic has been repeatedly debated in terminology literature. Part of this literature advocates ways to prevent polysemy. Another smaller portion presents examples of lexical items that can be defined in a surprisingly high number of different ways depending of the perspective taken on them. Seldom, however, are compilers of domain-specific resources offered solutions to manage polysemy.

In this paper, we focus mainly on polysemy observed within specialized fields of knowledge, as we are concerned with the management of polysemy when compiling domain-specific resources. This being said, we do not adhere to the traditional homonymy/polysemy distinction that can still be made in terminology. Looking at senses within a single subject field only provides a partial picture of the different senses lexical items can carry. Theoretically, connections between senses across domains or between specialized and general usage should also be taken into consideration since they explain how specialized meanings are situated within the lexicon of a language. In practice, however, terminologists focus on senses that are relevant in given domains and need to find ways to account for them.

In this paper, we first explain that polysemy affects many lexical items in specialized domains (Section 2). We also review specific cases of polysemy that can be observed in these domains (Section 3). Some of these cases are seldom discussed in terminology literature. We also suggest methods to handle and describe polysemous lexical items in terminological resources (Section 4). Examples are based on entries that can be found in the English versions of two terminological dictionaries in the fields of computing and the environment.

## 2 Reduced Polysemy in Terminology?

Managing polysemy is certainly a less intricate matter for terminologists than it is for lexicographers. For instance, in an environment dictionary (DiCoEnviro 2020), 1,045 meanings were identified for 882 English lexical items (a 1.18 ratio). Similarly, in a computing dictionary (DiCoInfo 2020), the ratio observed is 1.25 (1,896 meanings for 1,511 English

<sup>1</sup> This indeed differs from the way “homonymy” is defined in lexical semantics and lexicography where homonymy is distinguished from polysemy when no intersection between senses can be identified.



lexical items).<sup>2</sup> These figures can give an indication of the presence of polysemy in domain-specific resources, but it must be kept in mind that dictionary word lists reflect choices made by their compilers. In text, specialized meanings can interact with other senses that can be associated with general language of other fields of knowledge. Nevertheless, the meaning/lexical item ratio is still lower than in general language dictionaries which is around 2.0 with some variation from one dictionary to another (Cooper 2005).

A number of factors explain why polysemous items are less common in terminological resources. Firstly, the focus is placed on specialized meanings and other ones can be ignored to a certain extent. Terminologists compile domain-specific corpora in which many lexical items carry a single or a reduced number of meanings. Even when lexical items convey multiple domain-specific meanings, their number is reduced when compared to the senses recorded in general language dictionaries. Table 1 gives a summary of the senses recorded for the verb *recover* in a general language dictionary (Merriam-Webster 2020), a computing dictionary (DiCoInfo 2020) and an environment dictionary (DiCoEnviro 2020). The Merriam-Webster makes up to 12 meaning distinctions as opposed to the computing dictionary which records a single meaning and the environment dictionary that describes three different senses.

Merriam-Webster (2020)		
recover (Entry 1 of 2)	transitive verb	
	1	to get back : <a href="#">regain</a>
	2a	to bring back to normal position or condition stumbled, then recovered himself
	2b	archaic : <a href="#">rescue</a>
	3a	to make up for recover increased costs through higher prices
	3b	to gain by legal process
	4	archaic : <a href="#">reach</a>
	5	to find or identify again recover a comet
	6a	to obtain from an ore, a waste product, or a by-product
	6b	to save from loss and restore to usefulness : <a href="#">reclaim</a>
	intransitive verb	
	1	to regain a normal position or condition (as of health) recovering from a cold
	2	to obtain a final legal judgment in one's favor
recover (Entry 2 of 2)	transitive verb	to cover again or anew
Computing dictionary		
recover	transitive verb	
	1	user recovers data: <i>In a "globalizing" economy, today's work force is necessarily becoming more mobile with the need to reliably store, access, and <b>recover</b> data from any location.</i>
Environment dictionary		
recover	transitive verb	
	1	official organization recovers materials: <i>the USA landfilled 54% of MSW, incinerated 14%, and recovered, recycled or composted the remaining 32 %</i>
	intransitive verb	
	2a	species recover: <i>species and their habitats are able to survive and recover in a warmer world.</i>
	transitive verb	
	2b	human recovers species: <i>to the contrary, the intent was to conserve and recover species.</i>

Table 1: Meanings recorded for recover in a general language dictionary and two domain-specific dictionaries

The second factor which explains why polysemy is reduced in domain-specific resources is that it is customary for terminologists to collect multiword expressions. In fact, in most specialized resources, the majority of entries describe multiword nouns, such as *climate change*, *expert system*, *configuration file*, etc. A potentially polysemous item is often disambiguated when considered within a longer sequence.

Finally, the conceptual approach with which most terminologists comply often compel them to focus on nouns or noun phrases. This impacts the perspective taken on polysemy that is chiefly concerned with diverging denotations (Béjoint & Thoiron 2000). Other types of polysemy that affect other parts of speech, alternations for example, are often ignored.

<sup>2</sup> It should be mentioned that these two domain-specific resources include several single-word terms, which is not common practice in terminology as we will see further on. It is to be expected that the meaning/lexical item ratio is even lower in traditional resources.



### 3 Cases of Polysemy

Even if polysemy is reduced when considering the meaning of lexical items from the perspective of a single subject field, it can still be found in specialized corpora and needs to be managed by terminologists. Furthermore, it takes many different forms that are described in the following subsections.

#### 3.1 Domain-specific versus other meanings

Terminology textbooks often mention the fact that many terms are created on the basis of more common meanings. Adding new meanings to existing lexical items is a commonly used method for creating terms (Sager 1990; Kocourek 1991; Aldestein & Cabré 2002; L'Homme & Polguère 2008). The addition can be the result of a metaphorical extension. In computing, there are multiple cases of the sort: *client* (defined as “hardware that uses a service given by a server”); *declare* (defined as “to state the content of a variable”) (see also Meyer et al. 1997).

The original lexical item can be part of general usage or taken from another special subject field. The environmental meanings of the adjectives *green* and *clean* (“that has a low impact on the environment”) illustrate the first situation. The meaning extensions of *virus* and of its collocates *infect* and *contaminate* in computing borrowed from medical terminology illustrate the second one.

In practice, however, this first case of textbook polysemy is not the most difficult that terminologists must tackle since, as was said above, they usually focus on domain-specific meanings.

#### 3.2 Multiple Meanings in the Same Field

Polysemy also occurs within domains and these are the cases that will need to be managed in practice. For instance, *environment* can designate “a global set of meteorological, biological conditions ...” or “a place where species carry out activities”. Both meanings are linked to the more general field of the environment. Similarly, *extinct* can mean “that is no longer active” or “that no longer exist”. (Examples from corpora are given in Table 2 for each of these meanings.)

Term	Example
<i>environment<sub>1</sub></i>	<i>the government's broader environmental vision aimed at supporting a healthy environment and a competitive economy</i>
<i>environment<sub>2</sub></i>	<i>many endangered freshwater fish and mussels need clean, clear, cold water to survive, and are sensitive to changes in their aquatic environment</i>
<i>extinct<sub>1</sub></i>	<i>species must be considered extinct if they are listed as endangered for 15 or more years.</i>
<i>extinct<sub>2</sub></i>	<i>This extinct volcano has woken up</i>

Table 2: Polysemous lexical items in the environment

##### 3.2.1 Regular Polysemy

Within special subject fields, different meanings can be more closely connected than those mentioned in Table 2 and lead to regular polysemy (Apresjan 1974;<sup>3</sup> Barque 2008). Different cases of regular polysemy in computing and the environment are illustrated below:

- Activity – result: *pollution<sub>1</sub>* (these include forest fires, floods, oil spills and pollution of waterways); *pollution<sub>2</sub>* (extensive inshore and coastal pollution).
- Concrete – abstract: *server<sub>1</sub>* (Computers are linked together, or “networked”, many of the programs and files can be stored centrally on a more powerful computer called a “server”); *server<sub>2</sub>* (In the client-server model, the term “server” describes the application that offers a service that can be utilized by any other application over the network).
- Whole – part: *sea<sub>1</sub>* (containers can be transported by sea); *sea<sub>2</sub>* (the coastal seas); *email<sub>1</sub>* (do not send us email asking for information); *email<sub>2</sub>* (a programming student sent this email to some friends).
- Entity – instrument: *email<sub>2</sub>* (a programming student sent this email to some friends); *email<sub>3</sub>* (email is a means of sending messages from one person to another using the Internet as the transmission mechanism).

It is likely that some cases of regular polysemy are more productive or occur more specifically in given domains. For instance, the concrete – abstract polysemy is quite prevalent in computing. In the environment, lexical items can first designate a natural entity and a resource exploited by men:

- (1) a fish<sub>1</sub>: Do not release snails, fish, or other aquatic animals or plants into our lakes, creeks, or rivers
- (2) fish<sub>2</sub>: The changes in aquatic habitat have also affected fisheries in lower valleys and deltas; the absence of nutrient-rich sediments has a detrimental effect on fish productivity.

<sup>3</sup> Polysemy of the word A with the meanings a<sub>i</sub> and a<sub>j</sub> is called regular if, in a given language, there exists at least one other word B with the meanings b<sub>i</sub> and b<sub>j</sub>, which are semantically distinguished from each other in the same way as a<sub>i</sub> and a<sub>j</sub> and if a<sub>i</sub> and b<sub>i</sub>, a<sub>j</sub> and b<sub>j</sub> are non-synonymous. (Apresjan 1974:16)



In contrast, it can also be expected that other cases of regular polysemy do not appear at all in certain domains of knowledge.

### 3.2.2 Alternations

Another common phenomenon affecting lexical items, and especially verbs, are syntactic alternations that introduce meaning distinctions as shown below with *crash*, *pollute* and *compile*. Interestingly, *compile* lends itself to two different alternations in computing.

- (3) ... many programs cannot handle time trouble and many crash.
- (4) The programs can crash PCs on their own, if they conflict with other programs ...
- (5) Pesticides pollute waterways and can harm animals and other plants.
- (6) We are destroying the earth by polluting the atmosphere with toxic emissions.
- (7) The above code compiles properly.
- (8) ... the GNU software and libraries compile and run the kernel.
- (9) A programmer types programming statements and then "compiles" them with this compiler.

Cases of inchoative/causative alternations (illustrated by (3) and (4) and by (7) and (8)) are usually recognized as introducing polysemy, as they correspond to an important syntactic distinction (intransitive vs. transitive). However, other cases are less unanimously considered as polysemous occurrences of lexical items. These latter cases include agent/instrument alternations (as in (5) and (6) and in (8) and (9)) and agent/location locations.

### 3.2.3 Microsenses

Other semantic modulations affecting lexical items are more difficult to characterize than the cases listed in the previous sections. Terminology literature has referred to these phenomena as *multidimensionality*, which is defined as a phenomenon whereby different perspectives are taken on what could be considered a single concept. León Araúz & Reimerink (2010) discuss the example of "sand" that can be defined as "a kind of sediment located in the sea, rivers or soil layers." However, looking at contexts in which *sand* appears, the authors note that the term could be associated with other concepts. In geology, for instance, although "sand" is still defined as a kind of sediment, it is further characterized according to grain size, and is viewed as a part of larger natural entities, such as valleys, deserts, etc. In another domain, that authors call the *coastal domain*, "sand" is also a part of larger natural entities, but these are restricted to coastal ones, such as beaches, and sand barriers. In addition, "sand" is viewed as something involved in natural processes, such as waves, and storms. And the list goes on as other differences are identified in coastal defense and water treatment. Each of these areas seem to trigger different conceptualizations of the concept "sand" and would require that new definitions be written for each of them. However, it is difficult to see on what grounds these conceptualizations are identified and at what point the distinctions should stop.

Cruse (1995) offers a different explanation under the label *microsense* that seems to cover some of the phenomena terminologists consider as manifestations of multidimensionality. In contrast with "standard" polysemy, microsenses are not completely incompatible since they can be linked to the same superordinate. However, they remain mutually exclusive since they can hold paradigmatic relations with different sets of lexical units. With the example *knife*, Croft & Cruse (2004: 127) explain that although *knife* usually denotes a kind of instrument (not completely incompatible), it can be linked, first, to *cutlery*, *fork*, *spoon* and, secondly, to *weapon*, *gun*, etc. (among other readings).

The following examples with the verb *introduce* shows how these subtle differences can occur between the common language reading and a domain-specific reading of a lexical item.

- (10) ... introduce changes directly into the text.
- (11) ... political resistance to introducing an endangered species to unoccupied habitat.

Although *introduce* carries the general meaning of "placing something somewhere", it appears in a specific lexical paradigm when considered from the perspective of endangered species. It is linked to terms such as *reintroduce*, *introduction*, *colonize*, *inhabit*. In contrast, the more general meaning would trigger associations such as *insert*, *insertion*, *delete* and *remove*.

Other cases, which we are concerned with in this article, concern distinctions that would be relevant from the point of view of specialized domains. Consider the verb *hunt* in the following sentences, both extracted from a corpus on endangered species. Does the verb have two distinct meanings with respect to this topic?

- (12) Predatory birds include the snowy owls that hunt waterbirds and lemmings.
- (13) A limited number of licenses to hunt game animals are sold.

In both sentences, *hunt* designates an activity that consists in "pursuing a living organism". However, a hunting situation associated with animals would be linked to feeding and survival and to terms such as *to prey*, *predation*, *predator*, whereas the hunting situation associated with human beings would trigger associations with *hunter*, *poach*, *poacher*, *capture*, etc. and the fact that it can be a threat to the survival of species.



#### 4 Handling and Representing Polysemy in Terminological Resources

For cases of polysemy listed above, meaning distinctions can be made with relational evidence used in lexical semantics (such as synonymy, near synonymy, opposition, or other kinds of paradigmatic relations (Cruse 1986)). For instance, the two meanings of *fish* mentioned earlier can be differentiated based on two different sets of lexical units as shown below. (The appendix summarizes meaning distinctions for lexical items that were mentioned in this article and lists lexical units that were used to validate distinctions.)

- (14) fish<sub>1</sub> as a species: hypernyms: *species*; *vertebrate*; co-hyponyms: *mammal*, *bird*; types of fish: *freshwater* ~, *cartilaginous* ~; meronyms: *fin*, *scale*  
 (15) fish<sub>2</sub> as a resource: hypernym: *resource*; typical place: *fishery*; holonym: *stock*; typical activity that f. can undergo: *to fish*, *to capture*; meronym: *meat*

Once these distinctions are made, different methods can be used in resources to: 1. explain separate meanings; 2. represent how some of these meanings are connected. In most domain-specific resources, different meanings are described in individual entries; in others, they are listed in a single entry. Usually, no real attempt is made to show how some senses are linked in a way that could be helpful for users. In the following sections, we first explain why connecting meanings is not always possible in domain-specific resources. Terminologist must thus make a distinction between: 1. polysemous items whose meanings are only remotely connected within the domain; 2. polysemous items whose meanings are closely related. We then suggest methods for highlighting both similarities and differences between meanings in both situations.

##### 4.1 Why it can be difficult to link different meanings in a domain-specific resource

General language resources use different methods to account for meaning distinctions and the way different meanings are connected, from hierarchical alphanumeric systems to more sophisticated mechanisms that consist in checking cohesiveness between definitions (Barque 2008).

In domain-specific resources, the use of these devices can be hindered by the fact that some meanings of polysemous items are only remotely linked. An example taken from the field of the environment will illustrate this problem. The adjective *green* in this domain conveys two different meanings: *green*<sub>1</sub> “covered with vegetation” (a *green neighbourhood*) and *green*<sub>2</sub> “that has a low impact on the environment” (*green vehicle*). Table 3 shows the definitions given for *green* in the Oxford English Dictionary (OED 2020) that correspond to the domain-specific senses we just mentioned.

Oxford English Dictionary (2020)	I With reference to colour.	
	2. Of a colour intermediate between blue and yellow in the spectrum; of the colour of grass, foliage, an emerald, etc.	
	2a	Covered with or abundant in foliage or vegetation; verdant; (of a tree) in leaf. Also in extended use.
	III In extended uses.	
	13b	Of a product, service, etc.: designed, produced, or operating in a way that minimizes harm to the natural environment.

Table 3: Environmental meanings of *green* recorded in the OED (2020)

In the OED, the two environmental-relevant meanings of *green* appear in the broader spectrum of all the meanings that *green* can convey. (The OED makes over 30 distinctions for this adjective.) A hierarchical alphanumeric system accounts for how the senses recorded in the dictionary are organized. Looking at the gap between the two environmental senses of *green* as recorded in the OED, even if there is a remote metaphorical connection between the “covered with vegetation” and the “that has a low impact on the environment” meanings, it would be difficult to account for it in a domain specific environmental resource without considering other senses that the adjective carries outside the domain of the environment. The hierarchical alphanumeric system used in the OED is informative only to the extent that we have access to the entire structure. This also applies to systems for checking definitional content (Barque 2008). In our environmental resource, there would be a gap that only a reference to an external resource could fix as shown in Figure 1.

We thus suggest alternative tools to make differences and similarities between senses more explicit in domain-specific resources (Sections 4.2 to 4.4).



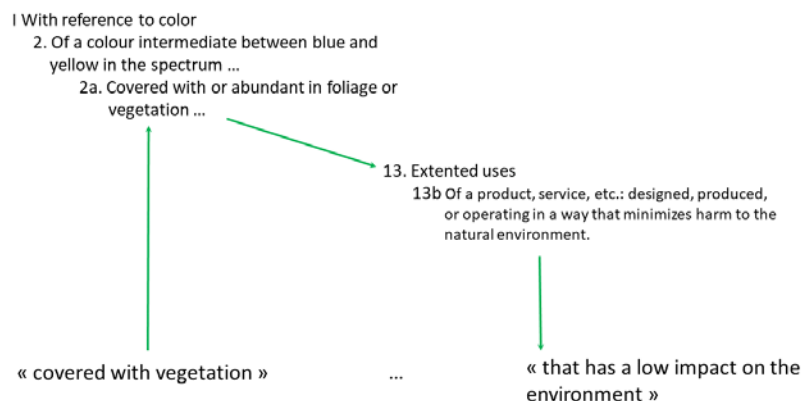


Figure 1: Remotely connected environmental meanings

## 4.2 Lexical relations

One classic method to make meaning distinctions explicit consists in displaying the lexical relations in which each meaning is involved (as shown in Figure 2 for the two environmental meanings of *environment*<sup>4</sup>). The meaning relations are those that validate meaning distinctions. Figure 2 displays the lexical sets graphically but other textual displays can also be used.

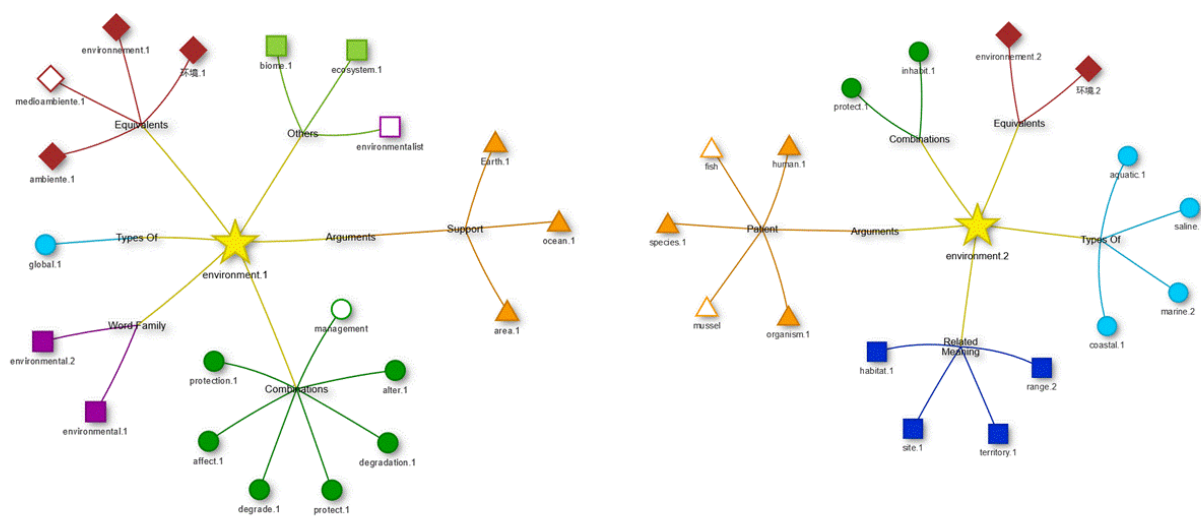
Figure 2: Lexical relations shared by two meanings of *environment* (NeoVisual 2020)

Figure 2 shows that the first meaning of *environment* informally explained as: “a global set of meteorological, biological conditions ...” is connected to units such as *ecosystem*, *biome*, *environmental*, *global*, etc. When it designates “a place where species carry out activities”, *environment* is linked to *site*, *territory*, *habitat*, *range*, *coastal*, *marine*, etc. Figure 2 also shows that the arguments of each meaning are realized with different terms (*environment*<sub>1</sub>: *ocean*, *Earth*, *area*; *environment*<sub>2</sub>: *organism*, *species*, *fish*, etc.) (the next section says more about the argument structure and how it can be used to represent semantic distinctions).

<sup>4</sup> Graphs in this figure also show equivalents in other language. It should be mentioned that, although they appear in the graph, equivalents are not used to support meaning distinctions since equivalents can be polysemous themselves.



### 4.3 Labelling of arguments

The different meanings of polysemous predicative units can be also be represented in terminological resources with an explicit and consistent labelling of arguments. Hanks & Pustejovsky (2005) suggest labelling arguments with types and roles. A similar method labeling arguments of predicative units with semantic roles and typical terms as shown below with the verbs *compile* and *recover*.

- (16) compile<sub>1a</sub>: program<sub>[Patient]</sub> compiles (Computing)
- (17) compile<sub>1b</sub>: compiler<sub>[Instrument]</sub> compiles program<sub>[Patient]</sub> (Computing)
- (18) compile<sub>1c</sub>: programmer<sub>[Agent]</sub> compiles program<sub>[Patient]</sub> with compiler<sub>[Instrument]</sub> (Computing)
  
- (19) recover<sub>1</sub>: user<sub>[Agent]</sub> recovers data<sub>[Patient]</sub> (Computing)
- (20) recover<sub>1</sub>: municipality<sub>[Agent]</sub> recovers material<sub>[Patient]</sub> (Environment)
- (21) recover<sub>2a</sub>: species<sub>[Patient]</sub> recovers (Environment)
- (22) recover<sub>2b</sub>: human<sub>[Agent]</sub> recovers species<sub>[Patient]</sub> (Environment)

In these examples, semantic roles appear between brackets. Typical terms are units that should be representative of the types of arguments that can fulfil an argument position. As can be seen with *compile*, *program* was chosen as the typical term that can realize the Patient (the argument that undergoes the process of compiling). It appears consistently in all three arguments structures of *compile*, albeit in different positions. The same applies to *compiler* labelled as an instrument which appears in the argument structures of two entries for *compile*. The consistent labelling can also be made explicit in definitions, as explained in San Martín & L'Homme (2014).

This method can be used not only for alternations, but for other meanings distinctions as shown below with *select* in computing.

- (23) user<sub>[Agent]</sub> selects option<sub>[Patient]</sub> (Computing)
- (24) user<sub>[Agent]</sub> selects string<sub>[Patient]</sub> (Computing)

### 4.4 Assignment to semantic frames

The identification of lexical relations and the explicit labelling of arguments can be exploited in a third method with the aim of providing a broader perspective on meaning distinctions in specialized fields. This method consists in modeling semantic frames (as understood in Frame Semantics, Fillmore 1982).<sup>5</sup> We cannot give here the full methodological details of how semantic frames are identified based on the terminology used in specialized fields of knowledge (for details, see L'Homme et al. 2020). Rather, we illustrate how the modelling of frames can highlight meaning distinctions in a way that complements the two first methods examined in Sections 4.2 and 4.3. We will use the examples given earlier for the verb *hunt* considered from the perspective of the environment.

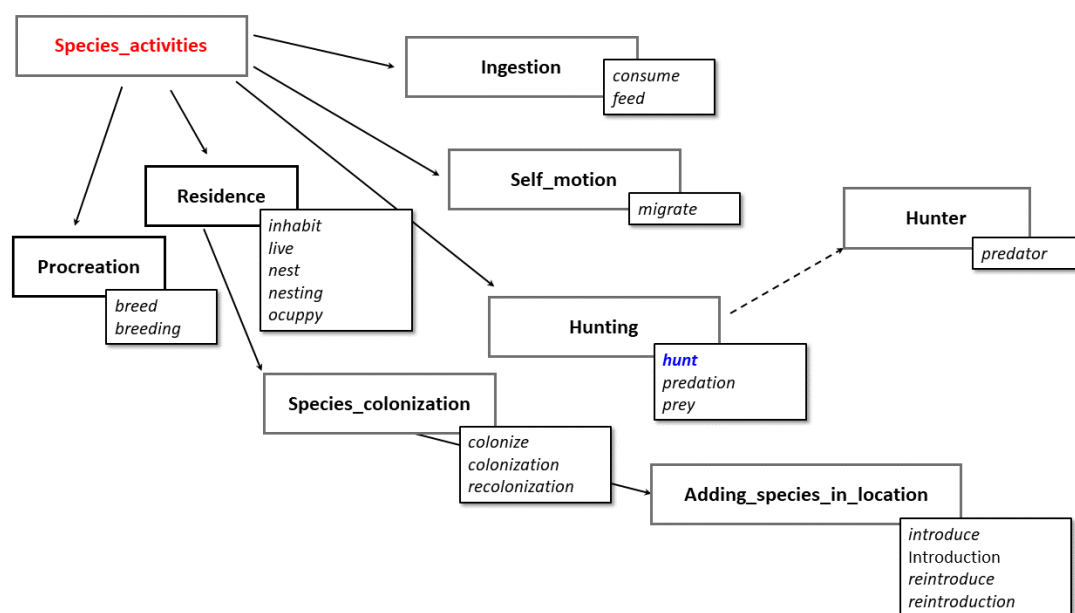


Figure 3: Frame Hunting, lexical content and related frames (based on Framed DiCoEnviro 2020)

<sup>5</sup> Frame Semantics (FS) assumes that the meanings of lexical units (LUs) are construed against a background of experience, beliefs or practices that are based at least partly on social and cultural institutions. Our understanding of lexical units involves a larger background, a broader situation that comprises participants and presuppositions.



Figure 3 shows how the frame Hunting that captures a situation whereby meat eaters chase other animals to feed appears within the broader context of species activities (only part of these activities, such as Procreation, Self\_motion, are reproduced in the figure). The terminological content of frames is also presented.

Figure 4 depicts part of the activities carried by human beings, again from the perspective of the environment (such as Manufacturing and Using\_resources). When comparing Figures 2 and 3, it can be seen that the frame Human\_hunting contains terms that differ from the ones listed in the Hunting frame (associated with species). More importantly, the broader context in which the Human\_hunting frame appears differs drastically from the one surrounding Hunting. Human\_hunting is one of the uses of natural resources that humans carry out, while Hunting appears within a set of activities that species need to do to live, reproduce and survive.

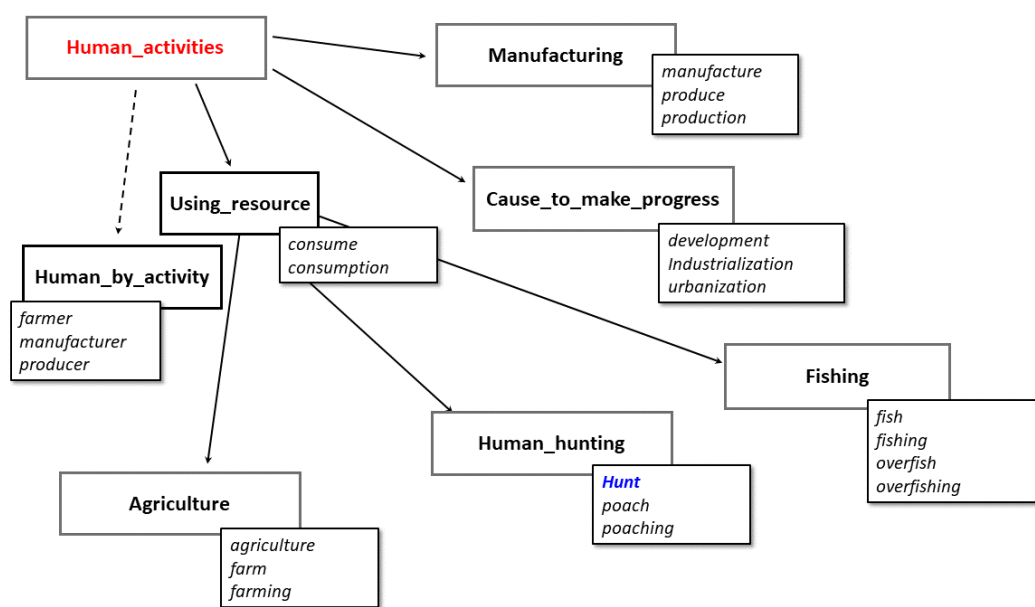


Figure 4: Frame Human\_hunting, lexical content and related frames (based on Framed DiCoEnviro 2020)

## 5 Conclusion

In this paper, we presented different cases of polysemy that can be found within specialized domains, i.e. regular polysemy, alternations, microsenses. When considered from the perspective of specialized domains, it is to be expected that polysemy is reduced when compared to situations that lexicographers need to manage. However, it is still prevalent in many domains and terminologists need to find ways to represent it adequately in domain-specific resources. It can also be surmised that specific phenomena such as microsenses are more common when considering the senses of lexical items from the perspectives of special subject fields. The latter phenomenon should be investigated more thoroughly to better characterize them and define precise criteria in order to determine when distinctions are truly needed.

We also described methods to represent different meanings in domain-specific resources, i.e. list lexical relations (or present them in a graph), label argument structures explicitly, and provide a broader perspective with semantic frames. Use of lexicographic systems, such as hierarchical alphanumeric systems and checking definitional content, is not always possible in domain-specific resources, since some meanings are only remotely connected. The methods we suggest can be used for both closely and remotely linked meanings. The first two methods were implemented in two domain-specific resources: the first one (DiCoInfo 2020) contains terms in the field of computing, the second (DiCoEnviro 2020) records environment terms. The third method was used for terms in the environment (Framed DiCoEnviro 2020). The use of labels in argument structure can also be implemented in definitions.

In this paper, the focus was placed on meaning distinctions from the point of view of specific subject fields. A further extension of this work would be to find ways to better model the interconnections between meanings across domains and across “general” language and specialized areas of knowledge.



## 6 References

- Aldestein, A., & Cabré, M.T. (2002). The Specificity of Units with Specialized Meaning: Polysemy as an Explanatory Factor. In *D.E.L.T.A.*, 18, pp. 1-25.
- Apresjan, J. (1974). Regular Polysemy. In *Linguistics*, 12(142), pp. 5-32.
- Barque, L. (2008). *Description et formalisation de la polysémie régulière du français*, Thèse présentée à l'Université Paris 7 Denis Diderot, Paris, France.
- Béjoint, H. & Thoiron, P. (2000). *Le sens en terminologie*. Lyon: Presses universitaires de Lyon.
- Cooper, M. (2005). A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts. In *Computational Linguistics*, 32(2), pp. 227-248.
- Croft, W. & Cruse, D.A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Cruse, D.A. (1986). *Lexical Semantics*. Cambridge: Cambridge University Press.
- Cruse, D.A. (1995). Polysemy and Related Phenomena from a Cognitive Linguistics Viewpoint. In P. Saint-Dizier & E. Viegas (eds.) *Computational Lexical Semantics*. Cambridge: Cambridge University Press, pp. 33-49.
- DiCoEnviro. *Dictionnaire fondamental de l'environnement* (2020). Accessed at: [http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search\\_enviro.cgi](http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi) [20/07/2020].
- DiCoInfo. *Dictionnaire fondamental de l'informatique et de l'internet* (2020). Accessed at: <http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi> [20/07/2020].
- European Union Terminology (IATE). Accessed at: <https://iate.europa.eu> [07/04/2020].
- Felber, H. (1987). *Manuel de terminologie*, Unesco: Infoterm.
- Fillmore, C.F. (1982). Frame Semantics. In The Linguistic Society of Korea (ed.) *Linguistics in the Morning Calm*. Seoul: Hanshin, pp. 111-137.
- Framed DicoEnviro. Accessed at: <http://olst.ling.umontreal.ca/dicoenviro/framed/index.php>. [24/07/2020].
- FrameNet. Accessed at: <https://framenet.icsi.berkeley.edu/fndrupal/> [24/04/2020].
- Le grand dictionnaire terminologique (2020). Accessed at: <http://www.granddictionnaire.com/>. [07/04/2020].
- Hanks, P., & Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, 10(2), pp. 63-82.
- Kocourek, R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Niemeyer: Oscar Brandstetter.
- L'Homme, M.C., & Polguère, A. (2008). Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. In F. Maniez, & P. Dury (eds.) *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*. Lyon: Presses de l'Université de Lyon, pp. 191-206.
- L'Homme, M.C., Robichaud, B. & Subirats, C. (2020). Building Multilingual Specialized Resources Based on FrameNet: Application to the Field of the Environment. In T. Torrent, C.F. Baker, O. Czulo, K. Ohara & M. R. L. Petruck (eds.) *International FrameNet Workshop 2020. Towards a Global, Multilingual FrameNet. Proceedings*, Workshop of the Language Resources and Evaluation, LREC 2020, pp. 94-102.
- León Araúz, P., & Reimerink, A. (2010). Knowledge Extraction and Multidimensionality in the Environmental Domain. In *Proceedings of the Terminology and Knowledge Engineering (TKE) Conference*. Dublin: Dublin City University. Accessed at: <http://lexicon.ugr.es/pdf/leonreimerink2010.pdf> [18/02/2020].
- Merriam-Webster Online Dictionary. Accessed at: <https://www.merriam-webster.com/> [03/04/2020].
- Meyer, I., Zaluski, V. & Mackintosh, K. (1997). Metaphorical Internet Terms: A Conceptual and Structural Analysis. In *Terminology*, 4(1), pp. 1-33.
- NeoVisual (2020). Accessed at: <http://olst.ling.umontreal.ca/dicoenviro/neovisual/> [20/07/2020].
- Oxford English Dictionary. Accessed at <https://www.oed.com/> [16/04/2020].
- Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C., Baker, C. & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. Accessed at: [https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the\\_book](https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=the_book) [27/01/2017]
- Sager, J.C. (1990). *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.
- San Martín, A., & L'Homme, M.-C. (2014). Definition Patterns for Predicative Terms in Specialized Dictionaries. In *Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland.
- Termium. Accessed at: <https://www.btb.termiumplus.gc.ca/> [07/04/2020].



## Appendix

Term	Domain	Paradigmatic relations
<i>compile</i> <sub>1a</sub>	Computing	<i>Run</i>
<i>compile</i> <sub>1b</sub>	Computing	<i>Compiler</i>
<i>compile</i> <sub>1c</sub>	Computing	<i>Translate</i>
<i>crash</i> <sub>1a</sub>	Computing	<i>fail, terminate</i>
<i>crash</i> <sub>1b</sub>	Computing	<i>shut down, run, start</i>
<i>email</i> <sub>1</sub>	Computing	<i>mailbox, inbox</i>
<i>email</i> <sub>2</sub>	Computing	<i>message, hoax, post</i>
<i>email</i> <sub>3</sub>	Computing	<i>program, application</i>
<i>environment</i> <sub>1</sub>	Environment	<i>ecosystem, biosphere</i>
<i>environment</i> <sub>2</sub>	Environment	<i>habitat, territory, site</i>
<i>extinct</i> <sub>1</sub>	Environment	<i>extant, surviving</i>
<i>extinct</i> <sub>2</sub>	Environment	<i>Active</i>
<i>fish</i> <sub>1</sub>	Environment	<i>species, vertebrate, freshwater, cartilaginous</i>
<i>fish</i> <sub>2</sub>	Environment	<i>fishery, stock, resource</i>
<i>introduce</i> <sub>1</sub>	Environment	<i>insert, place, remove</i>
<i>introduce</i> <sub>2</sub>	Environment	<i>reintroduce, introduction, colonize, inhabit</i>
<i>hunt</i> <sub>1</sub>	Environment	<i>predator, prey</i>
<i>hunt</i> <sub>2</sub>	Environment	<i>poach, hunter, capture</i>
<i>pollute</i> <sub>1a</sub>	Environment	<i>acidify, contaminate</i>
<i>pollute</i> <sub>1b</sub>	Environment	<i>spill, contaminate, depollute, polluter</i>
<i>pollution</i> <sub>1</sub>	Environment	<i>contamination, acidification, polluter</i>
<i>pollution</i> <sub>2</sub>	Environment	<i>pollutant, contaminant, toxic, gaseous</i>
<i>recover</i> <sub>1</sub>	Computing	<i>restore, corrupt, damage</i>
<i>recover</i> <sub>1</sub>	Environment	<i>recycle, dispose, eliminate</i>
<i>recover</i> <sub>2a</sub>	Environment	<i>recovery, survive</i>
<i>recover</i> <sub>2b</sub>	Environment	<i>reintroduce, restore, decimate</i>
<i>sea</i> <sub>1</sub>	Environment	<i>ocean</i> <sub>1</sub> , <i>land, at ~</i>
<i>sea</i> <sub>2</sub>	Environment	<i>lake, river, Black, Mediterranean</i>
<i>select</i> <sub>1</sub>	Computing	<i>choose, deselect</i>
<i>select</i> <sub>2</sub>	Computing	<i>activate, highlight</i>
<i>server</i> <sub>1</sub>	Computing	<i>computer, remote, to network</i>
<i>server</i> <sub>2</sub>	Computing	<i>application, client, email</i>





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Papers**

**Lexicography for Special Needs**







# Sign Language Corpora and Dictionaries: a Multidimensional Challenge

Vacalopoulou A.

*ILSP-Institute for Language and Speech Processing / Athena RC*

## Abstract

This paper is an analysis of the main challenges in developing sign language resources such as corpora and dictionaries. Although difficulties in data collection and processing are common with those in similar projects for vocal languages, there are extra complications that seem to be unique to the creation of resources for sign languages. These, more language specific, problems could be categorised under three general headings: (a) linguistic obstacles, (b) financial obstacles, and (c) social obstacles. Most of the challenges in studying and describing any sign language spring from the nature of these languages themselves, which is why this nature is briefly described. Instead of dealing with the typical two-dimensional, linear representation of the linguistic message, researchers have to cope with a more complex and dynamic medium involving elements including hand position and movement, eye gaze, facial expression as well as head and body movement. All these, among others, make the acquisition and processing of signed material more expensive and time-consuming. Finally, the activity of building and exploiting sign language resources can also be held back by social factors, including choice of informants, communication barriers and prejudice.

**Keywords:** sign language lexicography; multimodal lexicography; sign language corpora; sign language resources; Greek Sign Language.

## 1 Introduction

As anyone who has ever contributed to the making of any dictionary knows, general and more detailed typological issues dictate both the content and the form of lexicographic products. As dictionary typology is among the fundamentals that guide a lexicographer's work, standard works on lexicography never fail to dedicate short or more extensive descriptions of it and how it affects dictionary writing (Zgusta 1971: 198-221; Béjoint 2000: 32-41; Hartmann 2001: 57-74; Atkins & Rundell 2008: 24-43). Although most lexicographers tend to specialise in a particular type of dictionary, they sometimes find themselves involved in very diverse projects. These can be commercial or academic, print or electronic, offline or online, purely linguistic or more encyclopaedic, monolingual or bilingual and multilingual, diachronic or synchronic, general or specialised, intended for decoding or encoding, short glossaries or multi-volume works, written for native speakers or targeting learners of the language. Conscientious lexicographers who are faced with a new type of project tend to research different aspects of the anticipated product and its end users to adjust their craft accordingly. Nothing, however, can fully prepare a lexicographer for the challenges of compiling, for the first time, a reference work involving sign language.

Based on their academic background and/or experience, most dictionary compilers would tend to assume that this is yet another bilingual project and would try to approach it in such a way. In many ways, any bilingual reference work can be more perplexing than a monolingual one simply because of the need to study more than one language at the same time (Lew 2013: 289). As a result, bilingual lexicography involves not only the extra element of comparison but also the collaboration between at least two native speakers of different languages. Nevertheless, one discovers that awareness of classic pitfalls of bilingual dictionaries in theory and practice is not enough to provide solutions to the problems that occur in sign language lexicography.

To a great extent, this is due to historical reasons as sign language lexicography is a relatively new discipline worldwide (Schermer 2006: 321; McKee & Vale 2017: 6-7), which leaves several aspects yet unstudied. However, as others have shown (Zwitserslood 2010: 444-445) a great part of the challenge lies in the nature of sign languages themselves. The aim of this paper is to list and categorise different challenges involved in the design and creation of sign language resources based on twenty years of professional involvement in the field as well as on testimonies by researchers with similar experience in the hope that researchers who are about to embark on similar ventures gain some perspective on the subject. In the following section, a brief description of the characteristics of sign languages is provided in relation to those of spoken languages so that the reader can have an overview of the nature of sign languages. Next, an account of the different challenges involved in the development of sign language resources is given; these challenges are classified here under three general headings: linguistic, financial, and social issues. The paper closes with a recapitulation of the points mentioned.

## 2 Spoken and Sign Languages: Similarities and Differences

There have been quite a few misconceptions regarding the nature of sign languages. Based on stereotyped notions, people often tend to expect any language system to be similar to the one (or the ones) they themselves recognise and use. As a result, the mere fact that utterances in sign language are not formed through speech but through signing has led some to think that sign language is less of a language (Armstrong & Karchmer 2002; Zwitserslood 2010: 444; Wilcox & Occhino 2016: 1) more similar to fabricated language (Zwitserslood 2010: 444; Wilcox & Occhino 2016: 2; Vale 2017: 14), a visual

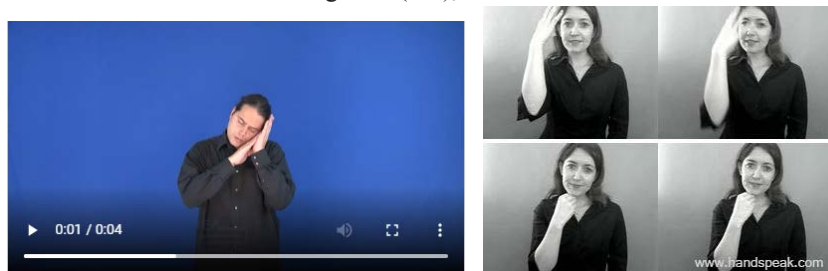


interpretation of spoken language (Wilcox & Occhino 2016: 2; Vale 2017: 14, Yule 2020: 230), or simply mime (Zwitserslood 2010: 444; Vale 2017: 14, 18). Another popular myth that still seems quite prevalent is that only one sign language exists and is shared by deaf people universally (Crystal 1992; Wilcox & Occhino 2016: 3; Vale 2017: 14). Although all these ideas of sign language are still common fallacies among the general public, sign language linguistics has been describing and researching very real and distinct sign languages since the mid-1960s. At the same time, sign language lexicography emerged, the first notable example of which was the celebrated *Dictionary of American Sign Language on Linguistic Principles* (Stokoe, Casterline & Croneberg, 1965). Although there had already been attempts to create glossaries and dictionaries to serve the needs of specific communities, the ground-breaking element of this reference work was that, for the first time, it arranged a sign language through a phonological system instead of alphabetised glosses (McKee & Vale 2017: 6). Eventually, research has established the linguistic status of different sign languages offering new insights to the ways in which they develop and operate in use. Such findings help lexicographers – among other sign language researchers – to decipher the true nature of sign languages and challenge stereotypical views of sign languages by understanding what makes spoken and signed languages similar and what makes them different. The fundamental similarity among all signed languages is that they are natural, that is, they are spontaneously created by members of a community to serve their communicative needs as opposed to “the artificially constructed systems used to expound a conceptual area (e.g. ‘formal’, ‘logical’, ‘computer’ languages) or to facilitate communication (e.g. Esperanto)” (Crystal 2008: 265). Consequently, every sign language is the creation of a specific deaf community, the members of which also share a common culture and are native signers of that language; as any other similar system – it can also be learnt by non-members of that community. Each of these languages is distinct in terms of lexical, morphological, syntactic and semantic aspects (Mayberry & Squires 2006: 291; Wilcox & Occhino 2016: 3), which do not directly correspond to with those of other spoken or signed languages.



Figure 1: The sign for *umbrella* in GSL (NOEMA+).

On the other hand, there is a vital difference between spoken and signed languages, which relates to their modality, i.e. the fact that they are visual-gestural as opposed to oral-aural (Zwitserslood 2010: 457; McKee & Vale 2017: 2). In other words, instead of being linear, the structure of the language is multidimensional as it is produced, perceived and understood in space. Signs (which give these types of languages their name) are the building blocks of communication in the sense that they usually convey the intended meaning. Signs, however, cannot be taken to have a one-to-one correspondence to words as, instead of sounds and syllables, they consist of different elements, which often carry some meaning themselves (Johnston & Schembri 1999: 117-118): (a) handshapes, i.e. the specific shape formed by one or both hands, (b) hand position, i.e. where hands are located, e.g. in front of the body or next to it, (c) hand movement, i.e. the way in which the hands move, (d) hand orientation, i.e. the direction in which the hands are placed, e.g. fingers facing the body, (e) non-manual elements in the face and other body parts apart from the hands, e.g. facial expressions or head tilting. Figure 1 shows a video still of the sign for *ομπρέλα umbrella* in Greek Sign Language (GSL), exemplifying some of these components. The sign starts by both hands shaped in fists (handshape) facing inwards (orientation) and placed in front of the body (position), followed by the top hand moving upwards (movement) as if opening an umbrella. An example of a non-manual element can be seen in Figure 2 (left), where the head tilts towards the hands to represent *sleep*



in GSL.

Figure 2: The sign for *sleep* in GSL (left, NOEMA+) and in ASL (HandSpeak).

A direct consequence of the visual-gestural modality is that the articulation of the abovementioned elements happens not only sequentially but sometimes also simultaneously, marking another difference between spoken and signed languages (Sandler 2006: 336). A second obvious element springing from this modality is the fact that a lot of signs seem to be



characterised by iconicity in demonstrating meaning, as shown in the GSL example of *umbrella* in Figure 1. In fact, as Taub (2001) has argued, it is this iconicity that has misled some into thinking that sign languages are universal in nature rather than separate arbitrary systems (as cited in Sandler 2006: 336). In fact, there is a lot of arbitrariness in sign languages not only concerning signs that represent things that cannot be demonstrated in such a way (such as abstract concepts) but in those corresponding to concrete things as well. This can be demonstrated by the fact that the same concrete thing can be represented by very different signs across sign languages. An example would be the representation of *sleep* in two different languages, GSL and American Sign Language (ASL) shown in Figure 2. In the video still, *sleep* is represented in GSL by putting one hand on top of the other, holding them both next to one side of the head and then closing the eyes and bending the head onto the hands as if they were a pillow. On the other hand, in the printable version in ASL *sleep* is signed by opening one hand in front of the face with the palm facing the face, then moving the hand towards the chin while joining the fingers together and touching them with the thumb.

### 3 Issues for Sign Language Lexicography

After this general description of the nature of sign languages, which springs from their modality, an attempt is made to analyse and categorise a series of the challenges involved in sign language lexicography partially drawing on a research group's experience in various Greek sign language lexicography projects. These projects include the development of relevant resources, such as the capturing of Greek Sign Language material in video, the annotation of the respective corpora as well as the design and development of various GSL dictionaries (Efthimiou et al. 2004; Efthimiou et al. 2017; Efthimiou et al. 2018; Vacalopoulou, Efthimiou & Vasilaki 2018; Vacalopoulou et al. 2018). The analysis will only refer to general challenges deriving from the multimodal nature of sign languages rather than more technical difficulties relating to the lexicographic treatment of specific signs of parts of signs, how detailed the information for each sign should be, how lemmas are selected or organised, etc.<sup>1</sup>

#### 3.1 Linguistic Issues

In order to represent any sign language and process it computationally, several transcription systems have been developed depending on project-specific use and needs. Among the most popular conventions internationally is the representation of signs using word glosses; in cases when one word is not enough to describe the respective sign, more words are included, typically joined together by hyphens. Glossing, which is a very helpful technique for alphabetising sign lists, has been widely used in the description of sign languages in terms of their morphology, syntax, and discourse (Miller 2006: 353) and it remains a standard way of sign transcription to this day.

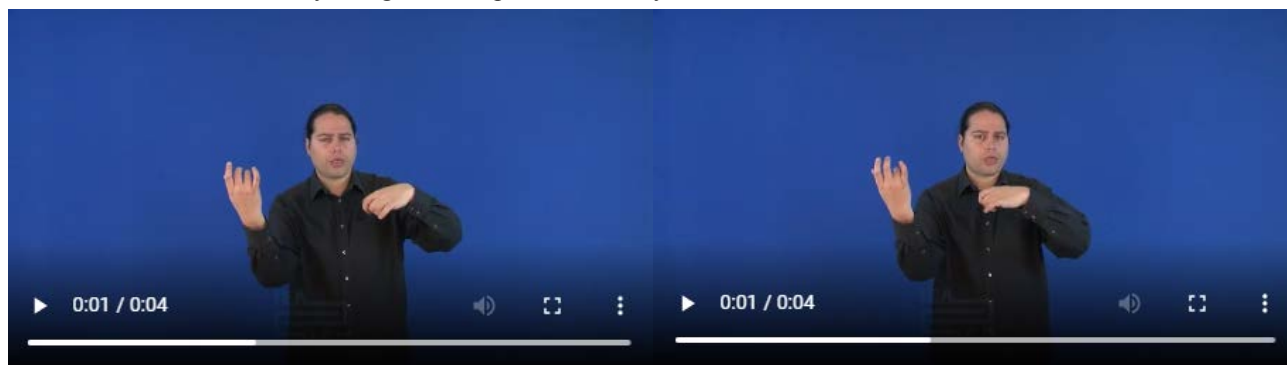


Figure 3: Video stills of the sign for *persuade* in GSL (NOEMA+).

This type of representation, however, poses several challenges, among which the fact that not all glosses include the full range of meanings carried by a sign. For instance, relevant research in Australian Sign Language has shown that the glosses OPEN-WINDOW and SHUT-WINDOW can be misleading as the same signs are generically used to represent the opening and shutting of any flat-surfaced thing (Johnston 2001: 251). Another problem results from cases of ambiguity as to what exact glossing should be used when the meaning of the sign is not straightforward; in other words, when the referent is not easily translatable to the respective oral language (Mesch & Wallin 2008: 135). In such occasions, the equivalent of a paraphrase is used in the gloss, which comprises more than one words. An example would be one of the signs for the verb *to persuade* in the specialised sense of “charming or humouring someone in order to convince them about something”. The GSL sign for this (Figure 3) is formed by playing an invisible fiddle while pushing out one's slightly open lips hinting to the charm element behind this action, which is depicted in the gloss “CONVINCE-DIPLOMACY”. In this case of polysemy, the same sign is, therefore, used to denote both *persuade* and *diplomacy*.

In addition, there are several spatial verbs – including verbs of movement and location – that tend to be represented in more than one translational equivalent in sign languages due to the fact that the type of movement linked to each of them is visually different depending on the context. As a result, there will be different signs in a language for the verb *to close* depending on what closes or what is being closed: “close the window”, “close one's eyes”, “close the curtains”, “close the

<sup>1</sup> Attempting to further expand on these or on the description of specific projects would exceed the scope of this study. For an overview of most of these issues, as well as an account of current sign language lexicographic practices, see Zwitserlood 2010.



shop”, etc. As explained in Morgan & Woll (2007: 1161), in such cases, “information is provided obligatorily about the location of a referent, where it moves from and to, how fast it moves, and what semantic class it belongs to”. The way in which these (and other) elements are shown in most sign languages is through the linguistic device of classifiers, which are generic handshapes that denote a specific group of concepts. These are added to signs in order to signify, for instance, the shape of an object, a change of posture, the direction of a movement, or the speed in which this movement takes place.



As a result, some concepts are represented in sign languages through a combination of (at least one) handshape plus a classifier, which specialises the meaning of the sign. An example in GSL would be *desk*, which is formed combining the sign for *writing* plus a classifier showing a flat horizontal surface (Figure 4). Being combinatory items, classifiers are not usually given lemma status in sign language dictionaries (Ivanova 2010: 127).



Figure 4: Video stills of the sign for *desk* in GSL (NOEMA+).

In an attempt to record all the different aspects mentioned above and, most notably, the fact that the modality of sign languages allows for simultaneous occurrence of several different elements, it has become clear that glosses may be helpful though not enough. In fact, glosses do not reveal information about sign languages per se but rather they connect signs to the respective lexical units of an oral language. This is why various systems have been developed for the notation of sign language phonology, their selection depending on the needs of each particular project (Miller 2006: 353). One of the most widely used ones, also employed in our GSL resource development projects, is HamNoSys (Figure 5), which was built by the Academy of Sciences in Hamburg and can be used to transcribe any sign language (Prillwitz et al. 1989).

Lemma	HamNoSys
ΤΡΕΧΩ	τρεχω
ΜΑΛΩΝΩ	μαλωνω
ΚΑΤΗΓΟΡΩ	κατηγορω
ΦΙΛΩ	φιλω
ΕΣΕΙΣ	εσεις

Figure 5: GSL transcription using *HamNoSys*.

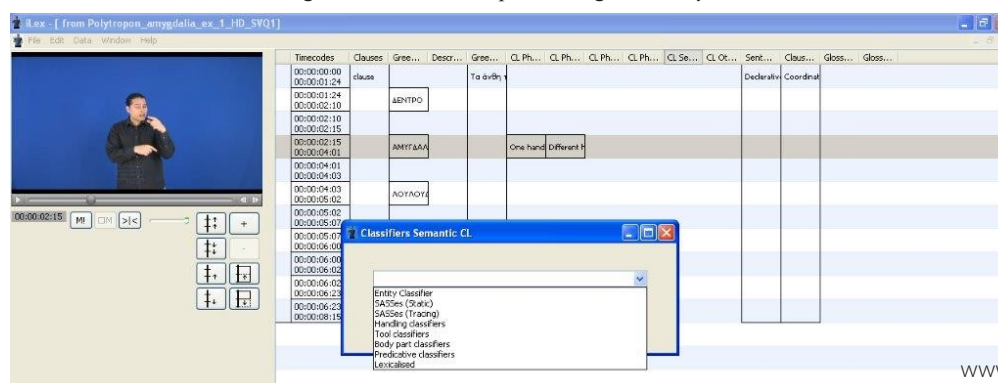




Figure 6: Classifier annotation by semantic function in *iLex*.

Notation systems are typically used in combination with specialised software to facilitate the computational recording of the phonological aspects of sign languages. An example is shown in Figure 6, where a sentence in GSL is broken down in parts for annotation in the *iLex* software for transcribing and linking signed corpora and dictionaries (Hanke & Storz 2008). Nevertheless, as extensive and detailed as most of these systems are, they are not always successful in fully grasping signs, as sign boundaries are seldom clear-cut in real utterances when there is considerable overlapping (Herrmann 2008: 69). A central issue here is the time parameter, that is, the fact that, apart from sequential articulation, there is also simultaneous articulation. The problem of segmentation, of course, is not unique to sign language processing, as boundaries between words are also often difficult to set in continuous speech (Creer & Thompson 2004; Himmelmann 2006). Indeed, signed and spoken language share a lot of similarities, such as being largely spontaneous and informal. This is why reliance on signed corpora for the description of a sign language could be compared in terms of difficulty to only having spoken corpora to describe a spoken language. Unit segmentation, as many more of the challenges in signed language linguistics, has been linked to the relatively new emergence of this scientific domain (Álvarez Sánchez, Báez Montero & Fernández 2008: 10), a perspective that brings hope for future solutions through relevant research.

### 3.2 Financial Issues

All sign languages are minority languages. A key obstacle in researching minority languages as well as collecting and creating resources is that they are not always financially supported from official organizations as the respective audience is not expected to be large enough (Ivanova 2010: 125; Vale 2017: 3). As a result, the number of available resources for most sign languages is still limited; this is particularly true for signed corpora that involve a considerable investment in terms of both money and time in order to be suitably collected and annotated (Crasborn & Zwitterlood 2008: 49; Naert et al. 2018: 139; Bragg et al. 2019: 19, 25). Such an investment mainly includes two types of expenses: those relating to the acquisition of signed material and those involving its processing.

Costs relevant to material acquisition are connected with the process of obtaining the appropriate equipment for the recording and processing of sign language utterances. This is usually done in video recordings and typically includes various types of cameras and other sensors. In order to fully grasp the multidimensional nature of sign messages, researchers tend to use high-resolution, multiple and/or depth cameras, a choice that increases the overall cost. When other types of sensors are used, body suits that capture motion will also need to be bought (Jedlička, Krňoul & Železný 2006: 102; Kanis & Krňoul 2008: 88; Bragg et al. 2019: 24). In addition, the recording of authentic videos is very often facilitated by sign language interpreters, another costly addition to the overall budget. The high cost of collecting sign language material has led some researchers to more inexpensive options such as using existing material from online sources or involving outside users in crowdsourcing platforms. However cheap, these possibilities do not come without drawbacks including quality control issues and the lack of appropriate annotation (Bragg et al. 2019: 24). Apart from acquisition, there are financial issues to consider in terms of training annotators on the use of relevant technologies, a process which usually is also time-consuming (op. cit: 21). Part of these costs could be decreased, however, if a standard system for annotating sign languages was to be adopted (Bragg et al. 2019: 25); indeed, the necessity for having standardised any collection of data intended for lexicographic use has been acknowledged by professional lexicographers (Atkins & Rundell 2008: 84). Reduction of costs is also one of the reasons why research has been aiming at the direction of automating the entire annotation process as much as possible (Meurant 2016).

### 3.3 Social Issues

It is only in recent years that sign languages, though not (yet) all of them, have been granted language status (Kristoffersen & Troelsgård 2012: 294). GSL, for that matter, was officially recognised as a language in 2000 (Timmermans 2005: 104) but had not been granted equal status with Modern Greek until 2017 (HFD 2017). Consequently, and given the various types of prejudice mentioned earlier, it comes as no surprise that research in the field of sign languages is a newcomer in linguistics. The social parameter has, therefore, been the main constraint for the shortage of signed resources. Indeed, this marginalisation has made deaf communities of the world more or less sceptical towards endeavours initiated by hearing people or organizations. Jones (2002: 56), for instance, reports examples of prejudice against professionals in the wider field of deafness who are not deaf themselves as mentioned by Lane back in the early 1990s. Given this tendency, reluctance to participate in such projects is not a rare phenomenon in sign language research.

As already mentioned, sign language lexicographic projects are more complex than any ordinary bilingual project. First, although not always the case, it is considered best practice among lexicographers to involve native speakers of both languages in the compilation of bilingual dictionaries (Atkins & Rundell 2008: 102; Stamper 2012). The activity of building and exploiting sign language resources can sometimes be held back by social factors as, in the context of GSL, linguists who are also native signers are scarce and there are hardly any lexicographers around sharing the same background. This is a reality for most sign languages, the users of which are linguistic minority groups within much broader communities. As a result, sign language lexicography, has been following, along with minority language lexicography, an inevitable tradition of resources compiled and processed by non-native users (Chelliah & de Reuse 2011: 56; Cristinoi & Nemo 2013). Thus, sign language lexicographers unavoidably rely on native informants to ensure that their attempt to describe the language is accurate and up to date, as no pre-existing sign language corpora are readily available. Selecting the right informants for each project can be a complex procedure involving a series of different criteria (Langer et al 2018: 492), some of which are listed in Álvarez Sánchez, Báez Montero & Fernández (2008: 10):



Social background: “place and date of birth, age of deafness occurrence, deafness degree, deaf/hearing family, job of closest family members”; Education: “degree and type of studies, special/ordinary school, use/absence of SL in school”; “Linguistic skills”: in the research sign language, oral language, lip-reading, written language. In addition, any balanced selection of informants would include both men and women from various age groups. Given the fact that the population in question belong to a linguistic and cultural minority, it is evident that a well-adjusted selection of informants is a very demanding task. If this is seen in combination with the bias against hearing professionals, it is evident that the task borders on the impossible.

Furthermore, it is considered good practice that every item intended for inclusion in either a corpus (when this is not a spontaneous one) or a dictionary be reviewed by more than one informants so as to ensure that the actual meaning of the utterance is established. This process, however, is not without challenges, as no native signer consensus is established for a large number of issues in most sign languages (Johnston 2008: 82; Chen Pichler et al. 2016: 31). In fact, diversity among native signers is significant and relates to various factors such as “ethnicity, geographic region, age, gender, education, language proficiency, hearing status, etc.” (Bragg et al. 2019: 18). The fact that there are added parameters (such as hearing status) influencing diversity in sign languages combined with the scarcity of relevant research make the prospect of reaching consensus even more distant.

As if forming a balanced set of informants and trying to reach consensus among them is not enough of a challenge already, it has been noted that lack of formal teaching of sign languages to native signers may result to lack of linguistic conscience among the group (Álvarez Sánchez, Báez Montero & Fernández 2008: 11). Whatever the reason, practice has shown that it is often complicated to present informants with a set of glosses and ask them to represent them in sign. Indeed, in several occasions, we have found that abstracting the actual meaning or use of specific GSL lexical items can be quite difficult for informants who tend to concentrate on the glosses or words presented to them instead. This misleading one-to-one correspondence has often led informants to claim that several items “do not exist” in GSL, only to discover – along with researchers – that they very much exist, when informants are prompted to use them in context in actual GSL conversation. This one of the (several) reasons why there is an increasing tendency for more authentic signed resources as well as for the inclusion of authentic examples in sign language dictionaries (Langer et al 2018; Mesch & Schönström 2018: 121).

## 4 Recapitulation

Much like the nature of sign languages itself, the challenge of creating signed resources is a multidimensional one. For the adventurous linguists and lexicographers who get involved in related projects, this means that three types of issues will occasionally get in the way: linguistic, financial, and social. This paper attempted to describe and classify most of them in a way that is meaningful to researchers (about to be) involved in the design and creation of sign language resources. In the near future, technical issues such as standardising annotation systems and further automating the transcription process are expected to significantly lower the now high cost for acquiring and processing signed data. The availability of more authentic signed material will hopefully result not only in more accurate representations of these languages but in some that are more generally embraced by the deaf communities. As this field of research grows, most challenges relating to its recent emergence will no doubt start becoming milder and easier to meet.

## 5 References

- Álvarez Sánchez, P., Báez Montero, I.C., & Fernández Soneira, A. (2008). Linguistic, sociological and technical difficulties in the development of a Spanish Sign Language (LSE) corpus. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 9-12.
- Armstrong, D., Karchmer, M. (2002). William C. Stokoe and the study of signed languages. In D. Armstrong, M. Karchmer, & J. Van Cleve (eds.), *The Study of Signed Languages: Essays in honor of William C. Stokoe*, Washington, D.C., Gallaudet University Press, pp. xi–xix.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Béjoint, H. (2000). *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.
- Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., Morris, M. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. The 21st International ACM SIGACCESS Conference on Computers and Accessibility, Oct 2019, Pittsburgh, pp. 16-31.
- Chelliah, S.L., de Reuse, W.J. (2011). *Handbook of descriptive linguistic fieldwork*. Dordrecht: Springer.
- Crasborn, O., Zwitterlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 44-49.
- Creer, S., Thompson, P. (2004). Processing spoken language data: the BASE experience. In LREC 2004 Workshop Proceedings on Compiling and Processing Spoken Language Corpora. Lisboa, 24 May 2004, pp. 20-27.
- Cristinoi, A. Nemo, F. (2013). Challenges in endangered language lexicography. In *Lexicography and Dictionaries in the Information Age: Selected Papers from the 8th ASIALEX International Conference*, Bali, 20-22 August 2013. Denpasar: Airlangga University Press, pp. 126-132.
- Crystal, D. (1992). Sign Language. In T. McArthur (ed.), *The Oxford Companion to the English Language*, Oxford: Oxford University Press, pp. 935.
- Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Malden: Blackwell.



- Efthimiou, E., Fotinea, S.E., Kakoulidis, P., Goulas, T., Dimou, A.L., Vacalopoulou, A. (2017). Sign Search and Sign Synthesis Made Easy to End User: The Paradigm of Building a SL Oriented Interface for Accessing and Managing Educational content sign search and sign synthesis made easy to end user: the paradigm of building a SL oriented interface for accessing and managing educational content. In HCI 2017 Proceedings: 11th International Conference on Human-Computer Interaction. Vancouver, 9-14 July 2017, pp. 14-26.
- Efthimiou, E., Fotinea, Vacalopoulou, A., Goulas, T., Vasilaki, K., Dimou, A.L. (2018). Sign language technologies in view of future internet accessibility services. In PETRA '18 Proceedings: 11th Pervasive Technologies Related to Assistive Environments Conference. Corfu, 26-29 June 2018, pp. 495-501.
- Efthimiou, E., Vacalopoulou, A., Fotinea, S.E., Stainhaouer, G. (2004). Multipurpose design and creation of GSL dictionaries. In LREC 2004 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: From Sign Writing to Image Processing. Lisbon, 30 May 2008, pp. 51-58.
- HandSpeak: American Sign Language Online*. Accessed at: <http://www.handspeak.com/> [27/04/2020].
- Hanke, T., Storz, J. (2008). "iLex – A database tool for integrating sign language corpus linguistics and sign language lexicography". In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 64-67.
- Hartmann, R.R.K. (2001). *Teaching and Researching Lexicography*. Harlow: Longman-Pearson.
- Herrmann, A. (2008). Sign language corpora and the problems with ELAN and the ECHO annotation conventions. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 68-73.
- HFD (Hellenic Federation of the Deaf). (2017, September 7). *Recognition of the Greek Sign Language as Equal to the Greek Language* [Press release]. Accessed at <https://bda.org.uk/wp-content/uploads/2017/09/906-2017-PRESS-RELEASE-Recognition-of-the-Greek-Sign-Language-as-Equal-to-the-Greek-Language.pdf> [30/04/2020].
- Himmelman, N.P. (2006). The challenges of segmenting spoken language. In J. Gippert, N.P. Himmelman, & U. Mosel (eds.) *Essentials of Language Documentation*. Berlin: De Gruyter Mouton, pp. 253-274.
- Ivanova, N. (2010). The Icelandic sign language dictionary project: some theoretical issues. In LREC 2010 Workshop Proceedings: 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Valletta, 1 May 2010, pp. 125-128.
- Jedlička, P., Krňoul, Z. & Železný, M. Methods for recognizing interesting events within sign language motion capture data. In 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. Portorož, 28 May 2016, pp. 101-104.
- Johnston, T. (2001). Nouns and verbs in Australian sign language: an open and shut case? *Journal of Deaf Studies and Deaf Education*, 6(4), pp. 235-257.
- Johnston, T. (2008). Corpus linguistics and signed languages: no lemmata, no corpus. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 82-87.
- Johnston, T., Schembri, A. (1999). On defining Lexeme in a signed language. *Sign Language and Linguistics*, 2(2), 115-185.
- Jones, M.A. (2002). Deafness as culture: a psychosocial perspective. In *Disability Studies Quarterly* 22(2), pp. 51-60.
- Kanis, J. Krňoul, Z. (2008). Interactive HamNoSys notation editor for signed speech annotation. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 88-93.
- Kristoffersen, J.H., Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-318.
- Langer, G., Müller, A., Wähl, S., & Bleicken, J. Authentic examples in a corpus-based sign language dictionary - why and how. In XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 17-21 July 2018, pp. 483-497.
- Lew, R. (2013). Identifying, ordering and defining senses. In J. Howard (ed.). *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 284-302.
- Mayberry, R.I., Squires, B. (2006). Sign Language: Acquisition. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (2nd ed.), 291-296. Amsterdam: Elsevier. Accessed at: <https://www.sciencedirect.com/science/article/pii/B0080448542008543> [27/04/2020].
- McKee, R., Vale, M. (2017). Sign Language Lexicography. In P. Hanks, G.-M. de Schryver (eds.), *International Handbook of Modern Lexis and Lexicography*. Accessed at: <https://www.semanticscholar.org/paper/Sign-language-lexicography-McKee-Vale/e3d1d3a158ec1fd1652fa6462129616fd1baf86a> [27/04/2020].
- Mesch, J., Wallin, L. (2008). Use of sign language materials in teaching. In LREC 2008 Workshop Proceedings: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Marrakech, 1 June 2008, pp. 134-137.
- Mesch, J., Schönström, K. From design and collection to annotation of a learner corpus of sign language. In LREC 2018 Workshop Proceedings: 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Miyazaki, 12 May 2018, pp. 121-126.
- Meurant, L., Cleve, A., & Crasborn, O. (2016). Using sign language corpora as bilingual corpora for data mining: contrastive linguistics and computer-assisted annotation. In 7th Workshop on the Representation and Processing of



- Sign Languages: Corpus Mining. Portorož, 28 May 2016, pp. 159-166.
- Miller, C. (2006). Sign Language: Transcription, Notation, and Writing. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (2nd ed.), 328-338. Amsterdam: Elsevier. Accessed at: <https://www.sciencedirect.com/science/article/pii/B008044854200242X> [27/04/2020].
- Morgan, G., Woll, B. (2007). Understanding sign language classifiers through a polycomponential approach. *Lingua*, 117: 1159-1168.
- Naert, L., Reverdy, C., Larboulette, C., & Gibet, S. Per channel automatic annotation of sign language motion capture data. In LREC 2018 Workshop Proceedings: 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Miyazaki, 12 May 2018, pp. 139-146.
- NOEMA+. Online Greek Sign Language Dictionary. Accessed at: <http://sign.ilsp.gr/signilsp-site/index.php/el/noima/> [27/04/2020].
- Chen Pichler, D., Hochgesang, J., Simons, D., & Lillo-Martin, D. Community input on re-consenting for data sharing. In 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. Portorož, 28 May 2016, pp. 29-34.
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T., & Henning, J. (1989). *HamNoSys. Version 2.0. Hamburg Notation System for Sign Language: An Introductory Guide*. Hamburg: Signum Verlag.
- Sandler, W. (2006). Sign Language: Overview. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (2nd ed.), 328-338. Amsterdam: Elsevier. Accessed at: <https://www.sciencedirect.com/science/article/pii/B008044854200239X> [27/04/2020].
- Stamper, K. (2012, May 9). A Letter to a Prospective Lexicographer. *harm-less drudg-ery*. Accessed at: <https://korystamper.wordpress.com/> [27/04/2020].
- Schermer, T. (2006). Sign Language: Lexicography. In K. Brown (ed.) *Encyclopedia of Language and Linguistics* (2nd ed.), 321-324. Amsterdam: Elsevier. Accessed at: <https://www.sciencedirect.com/science/article/pii/B0080448542002315> [27/04/2020].
- Timmermans, N. (2005). *The status of sign languages in Europe*. Report drawn in co-operation with the Committee on the Rehabilitation and Integration of People with disabilities. Strasbourg: Council of Europe.
- Vacalopoulou, A., Efthimiou, E., Fotinea, S.E., Goulas, T., & Dimou, A.N. (2018). Making online educational content accessible in Greek sign language. In EDULEARN2018 Proceedings: 10th International Conference on Education and New Learning Technologies. Palma, 2-4 July 2018, pp. 7305-7310.
- Vacalopoulou, A., Efthimiou, E., & Vasilaki, K. (2018). Multimodal corpus lexicography: compiling a corpus-based bilingual Modern Greek—Greek Sign Language dictionary. In XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 17-21 July 2018, pp. 17-21.
- Vale, M. (2017). Folk definitions as a model for sign language dictionary definitions: A user-focused case study of the Online Dictionary of New Zealand Sign Language. PhD thesis. Victoria University of Wellington, Wellington, New Zealand.
- Yule, G. (2020). *The Study of Language* (7th ed.). Cambridge: Cambridge University Press.
- Wilcox, S., Occhino, C. Historical Change in Signed Languages. (2016). In *Oxford Handbooks Online*. Accessed at: <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935345.001.0001/oxfordhb-9780199935345-e-24> [27/04/2020].
- Zgusta, L. (1971). *Manual of Lexicography*. Prague: Academia.
- Zwitsersloot, I. (2010). Sign language lexicography in the early 21st century and a recently published dictionary of Sign Language of the Netherlands. In *International Journal of Lexicography*, 23(4), pp. 443-476.

## Acknowledgements

I acknowledge support of this work by the project “Computational Sciences and Technologies for Data, Content and Interaction” (MIS 5002437) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). I am also indebted to my colleague, Kiki Vasilaki, for her helpful insights in GSL.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**

**Lexicography and Language Technologies**







# Learning dictionary skills from Greek EFL coursebooks: How likely?

Dalpanagioti Th.

*Aristotle University of Thessaloniki, Greece*

## Abstract

This paper presents a review of the dictionary-oriented material which is included in the EFL coursebooks used in Greek state secondary education. The aim of the study is to determine whether, to what extent and what kind of dictionary skills can be developed through the mainstream coursebooks under review. To this end, based on the relevant literature, a checklist of dictionary skills is first designed to serve as an evaluation tool for examining the coursebooks. Findings reveal an overall limited number of dictionary-oriented exercises, their random distribution across proficiency levels and the underrepresentation of basic receptive and productive dictionary skills. Therefore, since the coursebooks reviewed are not characterized by a thorough or informed treatment of dictionary skills, we may conclude that learning dictionary skills from coursebooks is rather unlikely. To address this gap, we offer suggestions as to how teaching materials can be modified or enriched with a view to developing learners' dictionary-using competence in a systematic way.

**Keywords:** dictionary skills; learners' dictionaries; coursebook evaluation; TEFL

## 1 Introduction

Dictionaries are generally recognized as useful learning tools; dictionary use figures in taxonomies of vocabulary learning strategies, it ranks high in users' preferences, and, in combination with inferencing, it considerably enhances vocabulary retention (Nation 2013; Schmitt 1997, 2000; Scholfield 1997). Recent advances in pedagogical lexicography seek to render dictionaries more accessible and user-friendly learning tools (Heubeger 2016; Rundell 1998); however, users cannot take full advantage of the wealth of information and facilities provided due to their poor dictionary skills (Chi 2013; Lew & Galas 2008). It is thus common practice for user-perspective researchers to emphasize the need for training users (native speakers, language learners, or translators) in how to make effective use of dictionaries (print or electronic, monolingual or bilingual) (Chi 2013: 182; Lew 2013:16; Pastor & Alcina 2010: 308). Although studies have focused on identifying users' needs and attitudes regarding dictionary use, practical aspects of dictionary use teaching such as "syllabuses for teaching users at various proficiency levels, teaching methodologies, materials and assessment" have not yet been explored (Chi 2013: 183). It is in this respect that this paper aims to make a contribution.

Focusing on the context of foreign language teaching in Greece, this study attempts to explore the role of dictionary skills in mainstream teaching materials. More precisely, we investigate whether dictionary pedagogy is integrated into the EFL coursebooks used in Greek state secondary education (Junior and Senior High School). The tool used for the evaluation of the coursebooks is a checklist designed to provide a detailed but concise overview of the treatment of dictionary skills. Section 2 sets the background of the survey study by presenting the checklist which acts (a) as a tool for reviewing the selected coursebooks in section 3 and (b) as a framework for proposing amendments in section 4.

## 2 Background of the survey

The first step in the review process is to decide on a classification of dictionary skills against which teaching materials are to be examined. Studies of dictionary skills often draw on Nesi's (1999) comprehensive taxonomy which is structured in terms of the stages in the consultation process (i.e. before study, before dictionary consultation, locating entry information, interpreting entry information, recording entry information, and understanding lexicographical issues). This classification framework has been variously used for specifying electronic dictionary skills (Lew 2013), for assessing the dictionary-oriented contents of textbooks (Molenda & Kiermasz 2013), or for relating dictionary skills to CEFR proficiency levels (Campoy-Cubillo 2015). Taking a different perspective, Nation (2013: 419-423) views dictionary use in relation to the task that prompts the consultation act and separates the dictionary skills needed for receptive use (i.e. reading, listening, L2-L1 translation) from those needed for productive use (i.e. writing, speaking, L1-L2 translation).

In light of the similarities and differences between the above specifications of dictionary skills, we



propose a checklist which foregrounds Nation's (2013) distinction between receptive and productive dictionary use, while also capturing the details of Nesi's (1999) taxonomy. This checklist, which is presented in Table 1, serves as a monitoring and evaluation tool for examining whether, and to what extent, Greek EFL coursebooks help Junior and Senior High School students develop dictionary skills. According to the national curriculum for teaching modern foreign languages in the Greek state school (<http://www.pi-schools.gr/programs/depps>), it is expected that, after the completion of secondary education, students will be able to use print and electronic dictionaries as sources of information (for receptive and productive tasks) and as a language learning strategy in general. However, more specific dictionary skills are not included in the competences expected to be developed.

	Steps	Skills
<b>General dictionary awareness</b>	Selecting a dictionary	1- Knowing what types of dictionary exist
		2- Knowing what kinds of information are found in dictionaries
<b>Receptive dictionary use</b>	Getting information from the context	3- Deciding on the part of speech of the look-up item
		4- Deciding on the form of the look-up item
		5- Guessing the general meaning of the look-up item
		6- Deciding whether consultation is necessary
	Finding the dictionary entry	7- Understanding the structure of the dictionary
		8- Understanding alphabetization and cross-referencing in print dictionaries
		9- Understanding the use of wildcards and hyperlinking in electronic dictionaries
		10- Interpreting the dictionary symbols for the different parts of speech
	Choosing the right sub-entry	11- Distinguishing the component parts of the entry
		12- Interpreting the L2 definitions or the L2-L1 translations
		13- Finding word groups (collocations, multi-word expressions)
	Relating look-up information to the context	14- Adapting the meaning found in the dictionary to the context of the word
		15- Evaluating the success of the search/ Recording entry information
<b>Productive dictionary use</b>	Finding the wanted word form	16- Finding an equivalent in a bilingual (L1-L2) dictionary
		17- Finding synonyms/opposites/word families/related words in a monolingual L2 dictionary or thesaurus
	Checking constraints on the use of the word	18- Interpreting restrictive labels concerning register, frequency, etc.
		19- Interpreting information concerning idiomatic and figurative use
	Working out the grammar and collocations of the word	20- Interpreting grammatical coding schemes and abbreviations
		21- Interpreting information about collocations
		22- Deriving information from examples
	Checking the form of the word before using it	23- Finding information about the spelling of words
		24- Interpreting IPA and pronunciation information

Table 1: A checklist of dictionary skills.



### 3 Critical review of coursebooks

Against this background, we proceed to identify and classify the dictionary-oriented materials included in seven EFL student's books (ranging from A1+ to B2+ CEFR level) and five workbooks (accompanying Junior High School student's books). All these textbooks are used in mainstream High Schools and are accessible in electronic form through the Digital School Project platform (<http://ebooks.edu.gr>). This platform not only includes .pdf versions of the print textbooks but also enriched versions with multimedia content such as “the audio of the listening comprehension tasks, suggested answers and models, additional references such as illustrations and word definitions, games, quizzes, videos and documentaries” (Mitsikopoulou 2014: 413). The digitally enriched resources are also examined with a view to identifying materials that aim at developing dictionary skills.

To provide an accurate picture of the attention that dictionary skills receive, Table 2 displays the number of dictionary-oriented exercises included in each one of the school books under review. What is striking is that the overall number of dictionary-related exercises is rather limited and that there is no clear pattern in their distribution across proficiency levels. Using the checklist of dictionary skills proposed above, we can present a more detailed picture of the specific dictionary skills represented in the teaching materials. Table 3 thus shows how many times each dictionary skill is targeted by the coursebooks.<sup>1</sup> A close look at Table 3 reveals that the skills related to productive dictionary use are more underrepresented than the skills related to receptive dictionary use. It is noteworthy that the skills most frequently involved in the exercises are “deciding whether consultation is necessary”, “interpreting the L2 definitions or the L2-L1 translations” and “finding word families in a monolingual L2 dictionary”; on the contrary, skills related to context –both getting information from context (reception) and incorporating information into context (production)– are notably absent (see skills 3-5 and 18-22 at Table 3).

Although the ability to use a dictionary is included in self-assessment sections of some student's books, no steps seem to be taken to help learners gradually develop this vocabulary learning strategy. Digital enrichment could be a unique opportunity to fill this gap and promote electronic dictionary skills; yet, only L2 glossaries have until now been added at the beginning of each unit without any awareness raising activities. Lastly, in examining coursebooks which address students of different grades, we would expect dictionary skills –similarly to other skills– to be graded along the continuum of proficiency levels; yet that is not the case. To sum up, the study reveals that dictionary skills are not treated in a thorough or systematic manner in the EFL teaching materials reviewed.

CEFR levels	EFL school books under review	Number of dictionary-oriented exercises
A1+	1st Grade of Junior High School: Student's Book (Beginners)	9
	1st Grade of Junior High School: Workbook (Beginners)	3
A2	1st Grade of Junior High School: Student's Book (Advanced)	9
	1st Grade of Junior High School: Workbook (Advanced)	1
A2	2nd Grade of Junior High School: Student's Book (Beginners)	2
	2nd Grade of Junior High School: Workbook (Beginners)	1
B1	2nd Grade of Junior High School: Student's Book (Advanced)	1
	2nd Grade of Junior High School: Workbook (Advanced)	6
B1+	3rd Grade of Junior High School: Student's Book	1
	3rd Grade of Junior High School: Workbook	1
B2	1st Grade of Senior High School: Student's Book	3
B2+/C1-	2nd & 3rd Grade of Senior High School: Student's Book	2

Table 2: The total number of dictionary-oriented exercises in the school books under review.

<sup>1</sup> The total number of tokens exceeds the number of exercises because some exercises target more than one dictionary skills.



Dictionary skills	1st Grade of Junior High School			2nd Grade of Junior High School			3rd Grade of Junior High School (B1+)		1st Grade of Senior High School (B2)	2nd & 3rd Grade of Senior High School (B2+/C1-)
	Beginners (A1+)		Advanced (A2)	Beginners (A2)		Advanced (B1)	Student's Workbook Book		Student's Book	Student's Book
	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Workbook Book	Student's Book	Student's Book
<b>General dictionary awareness</b>										
1- Knowing what types of dictionary exist	2	2	-	-	-	-	-	-	2	1
2- Knowing what kinds of information are found in dictionaries	2	2	-	-	-	-	-	-	-	-
<b>Receptive dictionary use</b>										
3- Deciding on the part of speech of the look-up item	-	-	-	-	-	-	-	-	-	-
4- Deciding on the form of the look-up item	-	-	-	-	-	-	-	-	-	-
5- Guessing the general meaning of the look-up item	-	1	-	-	-	-	-	-	-	-
6- Deciding whether consultation is necessary	1	2	-	-	1	3	-	1	2	1
7- Understanding the structure of the dictionary	-	-	-	-	-	-	-	-	-	-
8- Understanding alphabetization and cross-referencing in print dictionaries	1	-	-	-	-	-	-	-	-	-
9- Understanding the use of wildcards and hyperlinking in electronic dictionaries	-	-	-	-	-	-	-	-	1	-
10- Interpreting the dictionary symbols for the different parts of speech	-	-	-	-	-	-	-	-	-	-
11- Distinguishing the component parts of the entry	-	1	-	-	-	1	-	-	-	-
12- Interpreting the L2 definitions or the L2-L1 translations	3	1	3	2	-	3	1	1	3	1



Dictionary skills	1st Grade of Junior High School				2nd Grade of Junior High School				3rd Grade of Junior High School (B1+)		1st Grade of Senior High School (B2)	2nd & 3rd Grade of Senior High School (B2+/C1-)
	Beginners (A1+)		Advanced (A2)		Beginners (A2)		Advanced (B1)		Student's Book	Workbook	Student's Book	Student's Book
	Student's Book	Workbook	Student's Book	Workbook	Student's Book	Workbook	Student's Book	Workbook				
13- Finding word groups (collocations, multi-word expressions)	1	-	1	-	-	1	-	1	-	-	-	-
14- Adapting the meaning found in the dictionary to the context of the word	1	-	2	-	1	-	-	2	-	-	-	-
15- Evaluating the success of the search/ Recording entry information	3	-	1	-	-	-	-	-	-	-	1	-
<b>Productive dictionary use</b>												
16- Finding an equivalent in a bilingual (L1-L2) dictionary	2	1	1	-	-	-	-	-	-	-	-	-
17- Finding synonyms/opposites/word families/related words in a monolingual L2 dictionary or thesaurus	2	-	3	-	-	-	-	2	-	-	-	-
18- Interpreting restrictive labels concerning register, frequency, etc.	-	-	-	-	-	-	-	-	-	-	-	-
19- Interpreting information concerning idiomatic and figurative use	-	-	-	-	-	1	-	-	-	-	-	-
20- Interpreting grammatical coding schemes and abbreviations	-	-	-	-	-	-	-	-	-	-	-	-
21- Interpreting information about collocations	-	-	-	1	-	-	-	-	-	-	-	-
22- Deriving information from examples	-	-	-	1	-	1	-	-	-	-	-	-
23- Finding information about the spelling of words	-	-	-	-	-	-	-	-	-	-	-	-
24- Interpreting IPA and pronunciation information	1	-	-	-	-	-	-	-	-	-	-	-

Table 3: The number of tokens (instances) assigned to each dictionary skill.



#### 4 Suggestions for developing dictionary skills through coursebooks

Addressing the limitations highlighted above, this section provides some directions towards an informed inclusion of dictionary-related materials in EFL coursebooks. Firstly, coursebooks should guide learners to use different dictionary types in light of Campoy-Cubillo's (2015) proposal for defining and grading dictionary skills according to CEFR proficiency levels. Random references to different dictionary types in textbooks for beginners or vague references to "a dictionary" should be replaced by a clear progression from a bilingual dictionary (A1, A2) to a simplified monolingual L2 glossary (B1) to a monolingual L2 dictionary for advanced learners (B2-C2).

Secondly, learners need to be trained in using dictionaries as sources of information for both receptive and productive tasks by contextualizing their look-ups. In the case of receptive dictionary use, it is important that learners are involved in guessing meaning from context before looking up an unknown item; a combination of inferring (requiring depth of processing) and dictionary use (making sure the information retained is correct) would best promote vocabulary retention. In the case of productive dictionary use, learners should receive considerable practice in finding not only the desired L2 form but also its typical context (lexicogrammatical, pragmatic, etc.). Two examples are provided in the Appendix to demonstrate how coursebook exercises can be modified with a view to developing receptive and productive dictionary skills respectively. The implication is that dictionary-using competence is enhanced when options are removed from matching exercises and (intermediate) learners are motivated to search for the missing information (meaning or collocation) in a monolingual learners' dictionary (*Longman Dictionary of Contemporary English Online*).

Lastly, learners should progress from using dictionaries as sources of information to using them as discovery learning tools. For example, digitally enriched (upper) intermediate coursebooks could include links to resources (such as thesauri and corpus-based examples banks) incorporated in electronic dictionaries, and motivate learners to explore them in relation to receptive and productive tasks. In this way, dictionary use can promote learner autonomy.

#### 5 Conclusion

The aim of this paper has been to present an overview of the treatment of dictionary skills in the EFL coursebooks used in Greek state secondary education. The survey of 7 student's books, their digitally enriched versions and 4 workbooks clearly shows a dearth of dictionary-oriented exercises, their uneven distribution across proficiency levels and the underrepresentation of basic receptive and productive dictionary skills. Taking account of the prominent role of coursebooks in the EFL classrooms of Greek state schools, the implication of this study is that students do not receive explicit or adequate training in dictionary use. In an attempt to fill this gap, we have offered suggestions for systematically enhancing dictionary skills across different proficiency levels to meet learners' reception and production needs. Modifying or enriching existing teaching materials in light of these suggestions would make a contribution towards developing a dictionary culture within the mainstream education system.

#### 6 References

- Campoy-Cubillo, M. C. (2015). Assessing dictionary skills. In *Lexicography ASIALEX*, 2(1), pp. 1-23.
- Chi, A. (2013). Researching pedagogical lexicography. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London & New York: Bloomsbury Academic, pp. 165-187.
- Heuberger, R. (2016). Learners' dictionaries: History and development; Current issues. In Ph. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press, pp. 25-43.
- Lew, R. (2013). Online dictionary skills. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: Thinking outside the paper*. Proceedings of the eLex 2013, 17-19 October 2013, Tallinn, Estonia Ljubljana: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 16-31.
- Lew, R., K. Galas. (2008). Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English. In E. Bernal and J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 1273-1285.
- Longman Dictionary of Contemporary English Online*. Accessed at: <https://www.ldoceonline.com> [15/04/2020].
- Mitsikopoulou, B. (2014). Digital enrichment of EFL textbooks. In A. Psaltou Joyce, E. Agathopoulou, M. Mattheoudakis (eds.) *Cross-curricular Approaches to Language Education*. Newcastle: Cambridge Scholars Publishing, pp. 404-430.
- Molenda, M., Kiermasz, Z. (2013). Dictionary skills in advanced learners' coursebooks — Materials survey. In O. Majchrzak (ed.) *PsychoLingwistyczne Eksploracje Językowe*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, pp. 183-202.
- Nation, I. S. P. (2013). *Learning Vocabulary in Another Language*. 2nd edition. Cambridge: Cambridge University Press.
- Nesi, H. (1999). The specification of dictionary reference skills in higher education. In R. Hartmann (ed.) *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Subproject 9: Dictionaries*. Berlin: Free University Berlin, pp. 53-66.
- Pastor, V., Alcina, A. (2010). Search techniques in electronic dictionaries: A classification for translators. In *International Journal of Lexicography*, 23(3), pp. 307-354.
- Rundell, M. (1998). Recent trends in English pedagogical lexicography. In *International Journal of Lexicography*, 11(4), pp. 315-342.
- Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt, M. McCarthy (eds.) *Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, pp. 199-227.
- Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.



Scholfield, P. (1997). Vocabulary reference works in foreign language learning. In N. Schmitt, M. McCarthy (eds.) *Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press, pp. 279-302.

## Appendix

Match the words below in the text (1-8) with their definitions (a-h).

1. robust	a. financial support
2. insight	b. reach a point of highest development
3. refinement	c. identify with great accuracy or precision
4. scrutiny	d. the process of making a substance pure
5. agency	e. strong and reliable
6. funding	f. organisation representing a group of people
7. culminate	g. clear understanding of a complicated problem
8. pinpoint	h. detailed examination to get more information

Find these words in the text (page 51). Guess their meaning and check your answers in the dictionary.

(1st Grade of Senior High School – Student's book)

Scientists will now have to show their work will not only produce physiological insights but will also generate statistically **robust** data. If not, they will lose their funding.

**ro·bust** /rəˈbʌst, ˈrəʊbʌst \$ rəˈbʌst, ˈrou-/ **adjective**

- a robust person is strong and healthy  
a robust man of six feet four  
see thesaurus at **healthy**
- a robust system, organization etc is strong and not likely to have problems  
The formerly robust economy has begun to weaken.
- a robust object is strong and not likely to break **SYN sturdy**  
a robust metal cabinet  
see thesaurus at **strong**

Figure 1: Modifying a sample matching exercise to promote receptive dictionary use.

Complete each sentence by matching the appropriate adjective a-e to each noun in sentences 1-5.

a) ridiculous b) blonde c) ornate d) aristocratic e) sudden

- My aunt had a ..... desire to dye her hair black.
- In Ancient Egypt the ladies wore ..... make-up on their faces.
- Some kids wear the most ..... colour T-shirts.
- Ifigenia has decided to get some ..... highlights in her hair.
- In the Byzantine Empire, the ..... ladies wore purple dresses and chlamys.

(3rd Grade of Junior High School – Workbook)

Complete each sentence and check your answers in the dictionary.

**COLLOCATIONS**

**ADJECTIVES**

**great/strong**  
His one great desire in life was to own a Mercedes.  
The desire was too strong to resist.

**overwhelming** (=so strong that it takes control of you)  
He felt an overwhelming desire for a cigarette.

**deep/fierce** (=very great)  
The people of the village had a deep desire for revenge.

**a genuine/real desire**  
All her life she had a genuine desire to help the poor.

**a natural desire**  
Kids have a natural desire to find out about new things.

**a burning desire** (=an extremely strong desire)  
She had a burning desire to pack her case and leave.

**an insatiable desire** (=a desire that cannot be satisfied)  
She had an insatiable desire for publicity.

Figure 2: Modifying a sample matching exercise to promote productive dictionary use.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**

**Lexicography and Corpus Linguistics**







# The Greek Children Spoken Language Corpus (GCSL Corpus) / Το Προφορικό Σώμα Κειμένων Ελληνόφωνων Παιδιών (ΠΣΚΕΠ): Παρουσίαση, Εφαρμογές και Προοπτικές

Motsiou E.

University of Thessaly, Greece

## Περίληψη

Με την εξέλιξη της τεχνολογίας και τη δυνατότητα δημιουργίας σωμάτων κειμένων που περιέχουν μεγάλο όγκο δεδομένων η μελέτη της γλώσσας διευκολύνθηκε σημαντικά. Ειδικά όσον αφορά την έρευνα του αυθόρμητου λόγου, τα προφορικά σώματα κειμένων προσφέρουν μια ποικιλία και αριθμό δεδομένων που θα ήταν δύσκολο να αποκτηθεί με κάποια άλλη μέθοδο συλλογής. Το *Προφορικό Σώμα Κειμένων Ελληνόφωνων Παιδιών* (ΠΣΚΕΠ/ GCSL Corpus) έχει σαν σκοπό να διευκολύνει τη μελέτη για τη γλωσσική ανάπτυξη δίνοντας σε φοιτητές και ερευνητές ελεύθερη πρόσβαση σε υλικό για την εξελισσόμενη γλώσσα παιδιών προσχολικής ηλικίας των οποίων πρώτη γλώσσα είναι η Ελληνική. Το Σώμα αποτελείται από ψηφιακά αρχεία με συνομιλίες ενηλίκων και παιδιών, τα οποία εμπλουτίζονται σταδιακά, και προσφέρει τη δυνατότητα στο χρήστη να επιλέξει συνομιλίες χρησιμοποιώντας παράλληλα έναν συνδυασμό κριτηρίων αναζήτησης.

**Λέξεις-κλειδιά:** προσχολική ηλικία, αυθόρμητος λόγος, προφορικά σώματα κειμένων

## 1 Εισαγωγή: Η Μελέτη του Παιδικού Λόγου και τα Σώματα Κειμένων

Η μελέτη της εξέλιξης της παιδικής γλώσσας και η περιγραφή των ιδιαίτερων χαρακτηριστικών της είναι ένα γνωστικό πεδίο που έχει απασχολήσει ερευνητές διάφορων επιστημονικών κλάδων συναφών με την ανάπτυξη του παιδιού. Μια τέτοιου είδους έρευνα απαιτεί την καταγραφή και την ανάλυση μεγάλου όγκου δεδομένων από την παιδική γλώσσα. Με την ανάπτυξη της τεχνολογίας, η συλλογή υλικού από την παρατήρηση/ καταγραφή του αυθόρμητου παιδικού λόγου και από τις πειραματικές μεθόδους διευκολύνθηκε σημαντικά. Σήμερα, η εκτεταμένη αξιοποίηση των τεχνολογιών στην ανάλυση της γλώσσας και τη δημιουργία βάσεων δεδομένων έχει εξασφαλίσει τη δυνατότητα καταγραφής ενός εκτεταμένου και ποικίλου υλικού από τον παιδικό λόγο, την κωδικοποίησή του και την αυτόματη ανάλυση μεταβλητών, ώστε η ανάπτυξη της επιστήμης της παιδικής γλώσσας να διαθέτει τεράστιες προοπτικές για το μέλλον.

Παρότι ο αυθόρμητος λόγος είναι κεντρικό κομμάτι για τη μελέτη της γλώσσας και αναγνωρίζεται διεθνώς ότι τα προφορικά σώματα κειμένων είναι απαραίτητο εργαλείο για την έρευνα, τα δεδομένα από τον προφορικό λόγο είναι γενικά ελλιπή, εν μέρει εξαιτίας των δυσκολιών που έχουν να κάνουν με τη συλλογή τους (Pavlidou 2012). Το ίδιο ισχύει και για τον παιδικό λόγο, για τον οποίο ο αριθμός των βάσεων δεδομένων αλλά και η προσβασιμότητα είναι αρκετά περιορισμένα. Η πιο γνωστή διεθνής βάση δεδομένων παιδικού λόγου αναπτύχθηκε στην Αμερική το 1984 και ονομάστηκε CHILDES (MacWhinney 2000). Η βάση αυτή περιέχει δεδομένα και από την ελληνική γλώσσα, ωστόσο είναι μικρής σχετικά έκτασης.

Είναι φανερό ότι μια συλλογή γλωσσικού υλικού από παιδιά που καταγράφει σύγχρονα και διαρκώς αυξανόμενα δεδομένα, επιτρέποντας την ανάλυση και την εξαγωγή συμπερασμάτων για την αναπτυσσόμενη γλώσσα διευκολύνει και προωθεί τόσο τη μελέτη της φυσιολογικής ανάπτυξης της παιδικής γλώσσας όσο και της αποκλίνουσας, αλλά και τροφοδοτεί με κρίσιμα πορίσματα το χώρο της εκπαίδευσης και διδακτικής της γλώσσας (Behrens 2008; Diessel 2009). Το *Προφορικό Σώμα Κειμένων Ελληνόφωνων Παιδιών* (ΠΣΚΕΠ) έχει σαν σκοπό να διευκολύνει την έρευνα για τη γλωσσική ανάπτυξη δίνοντας σε φοιτητές και ερευνητές πρόσβαση σε υλικό για τη γλώσσα μικρών παιδιών με μητρική την Ελληνική. Το υλικό είναι μια συλλογή συνομιλιών ενηλίκων και παιδιών προσχολικής ηλικίας, συνεπώς σε ένα μεγάλο μέρος αποτελεί πηγή αυθόρμητου παιδικού λόγου.

## 2 Το Προφορικό Σώμα Κειμένων Ελληνόφωνων Παιδιών (ΠΣΚΕΠ)

Η ανάπτυξη ενός προφορικού σώματος κειμένων τυπικά ακολουθεί τις εξής διαδικασίες: διερεύνηση/ επέκταση πηγών, μεταγραφή, καθαρισμός-αποθήκευση, τυποποίηση, κωδικοποίηση και επισήμειωση (Γούτσος & Φραγκάκη 2015, Wynne 2005 κ.ά.). Η πηγές, οι διαδικασίες συλλογής και μεταγραφής των δεδομένων, καθώς και τα βασικά χαρακτηριστικά του ΠΣΚΕΠ περιγράφονται αμέσως παρακάτω.

### 2.1 Τα Δεδομένα

Το υλικό του ΠΣΚΕΠ αποτελείται από αυθεντικές συνομιλίες ενηλίκων και παιδιών προσχολικής ηλικίας. Ο συνολικός αριθμός των παιδιών, των οποίων ο λόγος καταγράφεται, είναι μέχρι σήμερα 91, 49 κορίτσια και 42 αγόρια, με ΜΟ ηλικίας τα 4 έτη. Όλα τα παιδιά είναι τυπικά αναπτυσσόμενα, μονόγλωσσα ή δίγλωσσα, που ζουν σε αστικές, ημιαστικές ή αγροτικές περιοχές (για παράδειγμα, Βόλος, Καβάλα, Αθήνα, Αγία Λάρισα, Αλόνησος, Νεοχώρι Καρδίτσας κτλ.).



## 2.2 Διαδικασία Συλλογής Υλικού

Το υλικό συγκεντρώνεται, από το 2015 και εξής, με τη βοήθεια των φοιτητών του Παιδαγωγικού Τμήματος Προσχολικής Εκπαίδευσης του Πανεπιστημίου Θεσσαλίας, στα πλαίσια του ετήσιου μαθήματος «Ανάπτυξη του λόγου στο παιδί». Ύστερα από κατάλληλη εκπαίδευση, οι φοιτητές/ φοιτήτριες μαγνητοφωνούν συνομιλίες με παιδιά (δικές τους ή άλλων), έχοντας εξασφαλίσει τη συγκατάθεση των γονέων και έχοντας πάρει όλα τα απαραίτητα μέτρα προστασίας προσωπικών δεδομένων. Από τους φοιτητές πραγματοποιείται και μια πρώτη απομαγνητοφώνηση του υλικού.

## 2.3 Μεταγραφή Δεδομένων και Προετοιμασία της Βάσης του ΠΣΚΕΠ

Η κύρια φάση της μεταγραφής των μαγνητοφωνημένων δεδομένων συνεχίστηκε μετά από τη σύσταση ομάδας έργου<sup>1</sup> με κατάλληλα εκπαιδευμένους συνεργάτες, οι οποίοι έχουν αναλάβει τον έλεγχο και τις διορθώσεις της πρώτης καταγραφής, την επιμέλεια και την προετοιμασία του υλικού για την εισαγωγή του στη βάση. Ακολουθεί μια τρίτη φάση αντιπαραβολής, ελέγχου και διόρθωσης από την κύρια ερευνήτρια. Κάθε συνομιλία αποθηκεύεται σε μορφή word doc (εικόνα 1) μαζί με τα μεταδεδομένα για την «ταυτότητα» κάθε συνομιλίας (πίνακας 1).

<b>E:</b> Δε σου αρέσει η τσιχλόφουσα; <b>E;</b>
<b>N:</b> οοο
<b>E:</b> Δε σ' αρέσει; Θυμάσαι που κάναμε μπάνιο στη θάλασσα;
<b>N:</b> ne
<b>E:</b> Εε, φορές μπρατσάκια; Για πες μου, εδώ φορές μπρατσάκια; Ή ξέρεις να κολυμπάς μόνη σου;
<b>N:</b> pal'a xron'a ee... evaza bratsak'a ala tora de vazo
<b>E:</b> Βατραχοπέδιλα βάζεις;
<b>N:</b> ο ((αρνητική απάντηση)) ee... m bori k'e ne
<b>E:</b> Μπορεί και ναι; Α, δηλαδή ξέρεις να κολυμπάς και λίγο ε; Κάνεις βουτιές;
<b>N:</b> ne
<b>E:</b> Πολύ μεγάλες κάνεις βαθιά;
<b>N:</b> na o nonos mu me petai sti thalasa TOSO psila ((δείχνει πόσο ψηλά την πετάει ο νονός της))

Εικόνα 1: Απομαγνητοφωνημένη & μεταγεγραμμένη συνομιλία

<b>Αριθμός συνομιλίας:</b> (σειρά συνομιλίας στο Σώμα)
<b>Συλλογή υλικού:</b> Κωδικός ερευνητή/τριας που συλλέγει το υλικό
<b>Απομαγνητοφώνηση:</b> κωδικός ερευνητή/τριας που απομαγνητοφωνεί το υλικό
<b>Συμμετέχοντες:</b> αναφορά συμμετεχόντων στη συνομιλία
<b>Φύλο νηπίου:</b> κορίτσι/ αγόρι
<b>Ηλικία νηπίου:</b> έτη και μήνες
<b>Χώρος:</b> χώρος διεξαγωγής συνομιλίας, π.χ. σπίτι, νηπιαγωγείο κτλ
<b>Τόπος διαμονής:</b> γεωγραφικός χώρος διαμονής νηπίου
<b>Πλαίσιο:</b> περιγραφή πλαισίου επικοινωνίας κατά τη διάρκεια της συνομιλίας, π.χ. παιχνίδι, αφήγηση κτλ
<b>Ημερομηνία καταγραφής:</b> Η/Μ/Ε
<b>Διάρκεια:</b> χρονική διάρκεια συνομιλίας (ΜΟ 14'30'')
<b>Κωδικός:</b> κωδικός καταχώρησης συνομιλίας ( <b>βλ.εικόνα 2</b> )
<b>Σύμβολα συμμετεχόντων:</b> Ε (ερευνητής/τρια), Ν (νήπιο), Μ (μητέρα), Π (πατέρας), Σ (συμμετέχων/ουσα: φιλικά ή λοιπά συγγενικά πρόσωπα νηπίου ή ερευνητών ή γονέων)

Πίνακας 1: Καρτέλα στοιχείων συνομιλιών (μεταδεδομένα)

<sup>1</sup> Η ομάδα έργου που δημιούργησε το ΠΣΚΕΠ αποτελείται από τους:

Ελένη Μότσιοι, Επίκουρη καθηγήτρια Παιδαγωγικού Τμήματος Προσχολικής Εκπαίδευσης του Πανεπιστημίου Θεσσαλίας, επιστημονικά υπεύθυνη

Θάνο Λίτσο, Μεταπτυχιακό φοιτητή του Τμήματος Αγγλικής Γλώσσας και Φιλολογίας του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, υπεύθυνο για την μεταγραφή και επιμέλεια των αρχείων του Σώματος

Γιάννη Κυριαζή, Ηλεκτρονικό Μηχανικό ΤΕ στο Παιδαγωγικό Τμήμα Προσχολικής Εκπαίδευσης του Πανεπιστημίου Θεσσαλίας, υπεύθυνο για τη δημιουργία της ηλεκτρονικής βάσης δεδομένων και ιστοσελίδας του ΠΣΚΕΠ.



Ο κωδικός κάθε συνομιλίας αποτελείται από τα εξής στοιχεία:



Εικόνα 2: Ανάλυση κωδικού συνομιλιών

Ο κωδικός εξειδικεύει επιπλέον ανάμεσα σε μονόγλωσσο ή δίγλωσσο παιδί (ML/ BL) και είδος γεωγραφικής περιοχής διαμονής του νηπίου (U (urban)/ SU (semi-urban)/ R (rural)/ NC(non categorized), στοιχεία που μπορούν να χρησιμοποιηθούν και ως κριτήρια επιλογής στο κουμπί Αναζήτηση.

Ο λόγος του ερευνητή/-τριας ή άλλων συμμετεχόντων καταγράφεται ορθογραφικά. Κατά την κωδικοποίηση του παιδικού λόγου επιδίωξη ήταν να αποδοθούν και φωνητικές ιδιαιτερότητες/ αποκλίσεις. Έτσι, για την καταγραφή του, χρησιμοποιήθηκε κατά βάση το Διεθνές Φωνητικό Αλφάβητο, με ορισμένες προσαρμογές για διευκόλυνση της ανάγνωσης, στην κατηγορία των ουρανικών: συγκεκριμένα, τα ουρανικά συμβολίζονται με [k', g', x', n', l'] (όμως κρατείται το σύμβολο [j] για το ουρανικό τριβόμενο ηχηρό).

Κατά την απομαγνητοφώνηση και μεταγραφή χρησιμοποιήθηκαν οι ελάχιστες δυνατές συμβάσεις/ σύμβολα, προκειμένου να διασφαλιστεί η κατά το δυνατόν ευχερής ανάγνωση της συνομιλίας και να περιοριστούν οι μη γλωσσικές πληροφορίες (Sinclair 1991). Τα σύμβολα που χρησιμοποιούνται είναι τα εξής:

- Οι παρατηρήσεις του ερευνητή που επεξηγεί/ σχολιάζει την κατάσταση ή το λόγο του παιδιού μπαίνουν σε διπλή παρένθεση: (( ))
- Εάν κάποιος φθόγγος, λέξη, φράση κτλ δεν ακούγεται καλά τοποθετείται σε μονή παρένθεση/ μια άδεια παρένθεση στη θέση του ακατανόητου εκφωνήματος: ( )
- Κεφαλαία: δηλώνουν αύξηση της έντασης της φωνής
- Αποσιωπητικά: δηλώνουν παύσεις στο λόγο
- Διπλά σύμβολα: δηλώνουν διάρκεια στην εκφώνηση

Το ΠΣΚΕΠ είναι ένα απλό, ειδικό και ανοιχτό σώμα κειμένων, που περιέχει προς το παρόν 84 συνομιλίες (128.100 λέξεις). Ως Σύστημα Διαχείρισης Περιεχομένου (CMS) έχει χρησιμοποιηθεί το ανοικτό και ελεύθερο λογισμικό WordPress. Η ιστοσελίδα έχει εγκατασταθεί σε μία πλατφόρμα LAMP (Linux/Apache/MySQL/PHP) στο χώρο του Πανεπιστημίου Θεσσαλίας.

## 2.4 Εμφάνιση, Δομή και Χαρακτηριστικά του ΠΣΚΕΠ

Πληκτρολογώντας στον φυλλομετρητή τη διεύθυνση <http://gcslece.uth.gr> ο χρήστης βρίσκεται στην αρχική σελίδα του ΠΣΚΕΠ:

Εικόνα 3: Το περιβάλλον του ΠΣΚΕΠ



Στο περιβάλλον αυτό ο χρήστης έχει τη δυνατότητα της επιλογής των παρακάτω κουμπιών:

- To Corpus*: περιέχει πληροφορίες για το Σώμα και την ερευνητική ομάδα
- Περιήγηση*: ο χρήστης μπορεί να δει σε έναν συγκεντρωτικό πίνακα όλες τις καταγεγραμμένες συνομιλίες, και να επιλέξει να δει ορισμένες πατώντας τον κωδικό κάθε μιας:

Περιήγηση													
Σημείωση: Η περιήγηση γίνεται με βάση τον κωδικό της συνομιλίας.													
Α/Α	ΣΥΛΛΟΓΗ ΥΛΙΚΟΥ	ΑΠΟΜΑΓΝΗΤΟΦΩΝΗΣΗ ΥΛΙΚΟΥ	ΣΥΜΜΕΤΕΧΟΝΤΕΣ	ΦΥΛΟ 1	ΗΛΙΚΙΑ 1	ΦΥΛΟ 2	ΗΛΙΚΙΑ 2	ΧΩΡΟΣ	ΤΟΠΟΣ ΔΙΑΜΟΝΗΣ	ΠΛΑΙΣΙΟ	ΗΜΕΡΟΜΗΝΙΑ	ΔΙΑΡΚΕΙΑ	ΚΩΔΙΚΟΣ 1
1	Ερευνήτρια Ευ. Σκ.	Ερευνήτρια Ευ. Σκ.	Νήπιο, Ερευνήτρια	Κορίτσι	3 ετών και 6 μηνών			Σπίτι νηπίου, αυτοκίνητο	Βόλος	Ελεύθερη παρατήρηση, συζήτηση	December 17, 2014	10' 56"	1f3.6MLU2014
2	Ερευνήτρια Μα. Μπ.	Ερευνήτρια Μα. Μπ.	Νήπιο, Αδερφή νηπίου, Ερευνήτρια	Αγόρι	5 ετών και 5 μηνών			Σπίτι νηπίου	Βόλος	Ελεύθερη συζήτηση	January 11, 2015	5' 02"	2m5.5MLU2015
3	Ερευνήτρια Βα. Γε.	Ερευνήτρια Βα. Γε.	Νήπιο, Ερευνήτρια	Αγόρι	4 ετών και 6 μηνών			Σπίτι νηπίου	Αγιά Λαρίσας	Συζήτηση	April 13, 2015	10' 35"	3m4.6MLU2015
4	Ερευνήτρια Αι. Κα.	Ερευνήτρια Αι. Κα.	Νήπιο, Ερευνήτρια	Αγόρι	5 ετών και 3 μηνών			Σπίτι νηπίου	Καβάλα	Παιχνίδι	April 17, 2015	10' 52"	4m5.3MLU2015
5	Ερευνήτρια Μα. Πα.	Ερευνήτρια Μα. Πα.	Νήπιο, Ερευνήτρια	Κορίτσι	5 ετών και 2 μηνών			Σπίτι νηπίου, στο παιδικό δωμάτιο	Σέρρες	Παιχνίδι, συζήτηση	January 3, 2015	4' 55"	5f5.2MLU2015
6	Ερευνήτρια 1 Αο. Κο.	Ερευνήτρια 2 Αβ. Στ.	Νήπιο, Ερευνήτρια	Κορίτσι	4 ετών και 2 μηνών			Σπίτι νηπίου	-	Συζήτηση	December 1, 2006	-	6f4.2MLU2016
7	Ερευνήτρια Γκ. Θω.	Ερευνήτρια Γκ. Θω.	Νήπιο, Ερευνήτρια	Κορίτσι	4 ετών και 5 μηνών			Σπίτι νηπίου	Λάρισα	Συζήτηση και άντληση πληροφοριών μέσα από το παιχνίδι (ζωγραφική, επιτραπέζιο παιχνίδι και	June 11, 2015	17'05"	7f4.5MLU2015

Εικόνα 4: Περιήγηση – συγκεντρωτικός πίνακας συνομιλιών

Επιλέγοντας έναν από τους κωδικούς ο χρήστης μπορεί να δει ολόκληρη τη συνομιλία που αντιστοιχεί σε αυτόν:

Αριθμός συνομιλίας:	15	E: Ποια είναι τα αγαπημένα σου παραμύθια;
Συλλογή υλικού:	Ερευνήτρια 1 Πν. Να.	N: Απ' όλα ποιο σ'αρέσει πιο πολύ;
Απομαγνητοφώνηση:	Ερευνήτρια 2 Σο. Ηλ.	N: mhm apo afta ru exo spiti mu?
Συμμετέχοντες:	Νήπιο, Ερευνήτρια 1	E: Από αυτά που έχεις σπίτι σου, από αυτά που έχετε εδώ, ποιο σ'αρέσει πιο πολύ;
Φύλο νηπίου:	Κορίτσι	N: a....., a....., afta ru maresun p'xo poli ine
Ηλικία νηπίου:	5 ετών	E: Φέρτο μαζί σου.
Χώρος:	Τάξη νηπιαγωγείου	N: Oiko mu ine.
Τόπος διαμονής:	Βόλος	E: Δικό σου είναι;
Πλαίσιο:	Κατευθυνόμενη συζήτηση	N: ne.
Ημερομηνία καταγραφής:	30-Ιουν-1905	E: Για έλα εδώ να κάτσουμε.
Διάρκεια:	10'	N: ojavase to na dis.
Κωδικός:	15f5MLU2008	E: Ξέρεις τι, δε θα το διαβάσω εγώ, θα μου πεις εσύ τι λέει. Θέλεις, Να μου πεις περίπου τι λέει, Πώς λέγεται;
Σύμβολα συμμετεχόντων:	N, E	N: ammmmmmm.... i anak'iklosi.
		E: Και η φίλη της.
		N: ne k'e i anak'iklosi k'e i fili tis.
		E: Θέλεις, όταν το ανοίγουμε, να μου δείχνεις τι λέει.
		N: oki, den ksero na ojavazo.
		E: Φυσικά δεν ξέρεις να διαβάζεις, αλλά να μου δείχνεις περίπου την ιστορία. Τι δείχνει εδώ; Γιατί σ'αρέσει αυτό το παραμύθι;
		N: jati ..... jati aftos o dukaiaktis e iiii, tha... se liyo epioti adjasan tha ksana jinun opos prin.
		E: Αλήθεια; Θα γίνουν όπως πριν; Για δείξε μου πού γίνονται όπως πριν;
		N: pu?
		E: Πού γίνονται όπως πριν. Πιο κοντά.
		N: ne, ala na sas po kali?
		E: Ναι ό,τι θέλεις.
		N: eeee, afto ix'e kolon'a mesa.
		E: Α, είχε κολόνια, εεε; Αυτό τι είχε μέσα;
		N: k'e adjase afto k'e to petaksa k'e afto ix'e. Ili itane vazo me lulu dja xoris tet'xo.
		E: Αααα, και αυτό τώρα τι είναι;

Εικόνα 5: Περιβάλλον συνομιλιών

**Σημείωση:** μπορεί να γίνει αναζήτηση συγκεκριμένων λέξεων στην επιλεγμένη σελίδα μέσω φυλλομετρητή (browser), πατώντας Ctrl+F στο πληκτρολόγιο και εισάγοντας το γλωσσικό στοιχείο: η λέξη "μαρκάρεται" ανάλογα και με τα κριτήρια που προσφέρονται και ανάλογα με τις φορές που συναντάται στο κείμενο της συγκεκριμένης σελίδας (βλ. Εικόνα 6).



Αριθμός συντομίας:	16
Σύλλαγγή υμνοειδ:	Ερευνητήρια 1 Πν. Ηα.
Απομακρυντοποίηση:	Ερευνητήρια 2 Σο. Ηλ.
Συμμετέχοντες:	Νήπιο, Ερευνητήρια 1
Φύλο νηπιού:	Κορίτσι
Ηλικία νηπιού:	5 ετών
Χώρα:	Τόζη νηπιολαλείου
Τόπος διαμονής:	Βόλος
Γλωσσά:	Κατευθυνόμενη συζήτηση
Ημερομηνία καταγραφής:	30-Ιουν-1905
Διάρκεια:	10'
Κωδικός:	16F5MLU2008
Σύμβολα συμμετεχόντων:	N, E

Ε: Για είναι το αγαπημένο σου παραμύθι;

N:

Ε: Αν' όλα ποιο σ'αρέσει πιο πολύ;

N: mnen apo afta pu exo spiti mu?

Ε: Από αυτά που έχεις σπίτι σου, από αυτά που έχεις εδώ, ποιο σ'αρέσει πιο πολύ;

N: a.... a...., afta pu maresun gri'o poli ine

Ε: Φέρο μαζί σου.

N: oiko mu ine.

Ε: Δικό σου είναι;

N: ne.

Ε: Για έλα εδώ να κάτσουμε.

N: θnawase io na ois.

Ε: Ξέρεις τι, Δε θα το διαβάσω εγώ, θα μου πεις εσύ τι λέει. Θέλεις. Να μου πεις περίπου τι λέει. Πώς λέγεται;


N: απεπιπεπη.... i anak'iklosi.

Ε: Και η φίλη της;

N: ne k'e i anak'iklosi k'e i fili tis.

Ε: Θέλεις, όταν το αναλογιζόμαστε, να μου δείχνεις τι λέει;

N: οκί, den ksero na θnawazo.

Ε: φυσικά δεν ξέρεις να διαβάζεις, αλλά να μου δείχνεις περίπου  ιστορία. Τι δείχνει εδώ; Γιατί σ'αρέσει αυτό το παραμύθι;

N: jati....., jati aftos o bukaikiatis e iia, θα... se ligo epidi adjasan tha ksana jnun opou prin.

Ε: Αλήθεια. Θα γίνουν όπως πριν; Για δείξε μου πού γίνονται όπως πριν;

N: pu?

Ε: Πού γίνονται όπως πριν; Νο κοινό.

N: ne, aia na sas po kat?

Ε: Ηαι ό τι θέλεις.

N: eeee, afto i'e kolo'n'a mesa.

Ε: Α, είχε κολόνα, ecc. Αυτό τι είχε μέσα;

N: K'e adjase afto k'e to petaksa k'e afto i'e, ii itane vazto me lulu'da xoris te'b'o.

Ε: Αααα, και αυτό τώρα τι είναι;

N: iilne... ta axrista prajmata pu den, pu eeehun.... K'e mn'a bo, K'e adjasan mn'a boja k'e to pi'ani i boja.

Ε: Ηαι, και αυτό είναι παραμύθι; Ξέρεις.

Εικόνα 6: Αναζήτηση γλωσσικών στοιχείων στον φυλλομετρητή

iii. *Αναζήτηση συνομιλιών*: ο χρήστης μπορεί να επιλέξει συνομιλίες χρησιμοποιώντας έναν συνδυασμό κριτηρίων αναζήτησης (φύλο, ηλικία παιδιού, τόπος διαμονής, μονόγλωσσο/ δίγλωσσο παιδί κ.ά.)

ΑΠΟΜΑΓΝΗΤΟΦΩΝΗΣΗ ΥΛΙΚΟΥ

ΦΥΛΟ ΝΗΠΙΟΥ 1

ΗΛΙΚΙΑ ΝΗΠΙΟΥ 1

ΦΥΛΟ ΝΗΠΙΟΥ 2

ΗΛΙΚΙΑ ΝΗΠΙΟΥ 2

ΧΩΡΟΣ

ΤΟΠΟΣ ΔΙΑΜΟΝΗΣ

ΗΜΕΡΟΜΗΝΙΑ

ΔΙΑΡΚΕΙΑ

ΚΩΔΙΚΟΣ 1

ΚΩΔΙΚΟΣ 2

Show 10 entries

Search:

ΑΠΟΜΑΓΝΗΤΟΦΩΝΗΣΗ ΥΛΙΚΟΥ	ΦΥΛΟ ΝΗΠΙΟΥ 1	ΗΛΙΚΙΑ ΝΗΠΙΟΥ 1	ΦΥΛΟ ΝΗΠΙΟΥ 2	ΗΛΙΚΙΑ ΝΗΠΙΟΥ 2	ΧΩΡΟΣ	ΤΟΠΟΣ ΔΙΑΜΟΝΗΣ	ΗΜΕΡΟΜΗΝΙΑ	ΔΙΑΡΚΕΙΑ	ΚΩΔΙΚΟΣ 1	ΚΩΔΙΚΟΣ 2
Ερευνήτρια Ευ. Σκ.	Κορίτσι	3 ετών και 6 μηνών			Σπίτι νηπίου, αυτοκίνητο	Βόλος	December 17, 2014	10' 56"	1f3.6	MLU2014
Ερευνήτρια Μα. Μπ.	Αγόρι	5 ετών και 5 μηνών			Σπίτι νηπίου	Βόλος	January 11, 2015	5' 02"	2m5.5	MLU2015
Ερευνήτρια Βα. Γε.	Αγόρι	4 ετών και 6 μηνών			Σπίτι νηπίου	Αγιά Λάρισας	April 13, 2015	10' 35"	3m4.6	MLR2015

Εικόνα 7: Επιλογή συνομιλιών με κριτήρια αναζήτησης

iv. *Χρήσιμοι σύνδεσμοι*: σύνδεσμοι για προφορικά και γραπτά σώματα κειμένων, παιδικού και ενήλικου λόγου, από την ελληνική αλλά και άλλες γλώσσες.

*Διαθεσιμότητα:* το ΠΣΚΕΠ προσφέρει ελεύθερη πρόσβαση στα δεδομένα του στη διεύθυνση <http://gcs1.ece.uth.gr>.

### 3 Εφαρμογές, Προοπτικές και Περιορισμοί

Ο αρχικός στόχος δημιουργίας του ΠΣΕΠ ήταν η διευκόλυνση της ποιοτικής έρευνας του παιδικού λόγου· φυσικά, ένα μέρος του Σώματος μπορεί να αξιοποιηθεί και ποσοτικά. Η δυνατότητα που δίνεται είναι να μελετηθεί ο λόγος των νηπίων είτε από την άποψη των συνομιλιακών ικανοτήτων, είτε από την άποψη της συνοχής και συνεκτικότητας, εκεί όπου ο λόγος είναι συνεχής ή/ και περιλαμβάνει μικρές αφηγήσεις και περιγραφές· είναι επίσης δυνατή η χρήση των δεδομένων για διαγλωσσικές συγκρίσεις. Πέρα από τη μελέτη των συνομιλιών, δίνεται η δυνατότητα να απομονώσει κανείς συγκεκριμένα γλωσσικά φαινόμενα, αλλά και να τα μελετήσει σε σχέση με ορισμένες μεταβλητές (όπως φύλο ή ηλικία). Ιδιαίτερη σημασία έχουν τα δεδομένα από τον αυθόρμητο λόγο για τη μελέτη φαινομένων όπως οι παιδικές αποκλίσεις και οι νεολογισμοί (Μότσιου, 2016). Ένα αυθεντικό γλωσσικό υλικό μπορεί να αξιοποιηθεί τόσο ερευνητικά, και μάλιστα για διάφορες ειδικότητες και σκοπούς (γλωσσολόγοι, ψυχολόγοι, λογοθεραπευτές κτλ) όσο και παιδαγωγικά, ως αυθεντικό γλωσσικό υλικό/ πηγή δεδομένων στη διδασκαλία.

Η διεύρυνση των δυνατοτήτων της αναζήτησης και η επισημείωση των δεδομένων είναι μια προέκταση του έργου που θα διευκόλυνε σημαντικά την αναζήτηση μεμονωμένων γλωσσικών στοιχείων: τόσο η εξειδικευμένη αναζήτηση όσο και ο συνεχής εμπλουτισμός του ΠΣΚΕΠ αποτελούν μελλοντικές προοπτικές, ωστόσο οι στόχοι αυτοί αναπροσαρμόζονται πάντα σε σχέση τη διαθεσιμότητα των ερευνητικών προγραμμάτων και πόρων.



#### 4 Βιβλιογραφικές Αναφορές

- Behrens, H. (2008). *Corpora in Language Acquisition Research: History, Methods, Perspectives*. Amsterdam/Philadelphia: John Benjamins.
- CHILDES (Child Language Exchange System). Διαθέσιμο στο: <https://chilides.talkbank.org/> [25/03/2020]
- Diessel, H. 2009. Corpus linguistics and first language acquisition. In A. Lüdeling & M. Kytö (eds) *Corpus Linguistics. An International Handbook*. Volume 2. Berlin/New York: Walter de Gruyter, 1197-1212.
- Goutsos, D. & Fragaki, G. (2015). *Introduction to Corpus Linguistics* [Γούτσος, Δ. & Φραγκάκη, Γ., *Εισαγωγή στη Γλωσσολογία Σωμάτων Κειμένων*]. Athens: SEAB. Διαθέσιμο στο: <http://repository.kallipos.gr/handle/11419/1932> [25/03/2020]
- MacWhinney, B. (2000). *The CHILDES project: tools for analysing talk*. 3<sup>rd</sup> ed., Mahwah, NJ: Lawrence Erlbaum Associates.
- Μότσιου, Ε. (2017). Νεολογικές κατασκευές στον αυθόρμητο λόγο ρωσόφωνων και ελληνόφωνων παιδιών: μια πρώτη προσέγγιση. *Γλωσσολογία* 25, 19-31. Διαθέσιμο στο: <http://glossologia.phil.uoa.gr/node/136> [25/03/2020]
- Pavlidou, Th.-S. (2012). The Corpus of Spoken Greek: goals, challenges, perspectives. *LREC Proceedings, Workshop 18* (Best Practices for Speech Corpora in Linguistic Research), 23-28.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Wynne, M. (ed.) (2005). *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. Διαθέσιμο στο [http://icar.cnrs.fr/ecole\\_thematique/contaci/documents/Baude/wynne.pdf](http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf) [25/03/2020]

#### 5 Ευχαριστίες

Ιδιαίτερες ευχαριστίες οφείλονται στους πρωτοετείς φοιτητές και φοιτήτριες του Παιδαγωγικού Τμήματος Προσχολικής Εκπαίδευσης του Πανεπιστημίου Θεσσαλίας που συνέλεξαν σε πρώτη φάση το υλικό και συνεχίζουν να συμβάλλουν στη μελλοντική διεύρυνση του ΠΣΚΕΠ.

Το έργο υλοποιήθηκε με χρηματοδότηση της Επιτροπής Ερευνών (Ειδικός Λογαριασμός Κονδυλίων Έρευνας) του Πανεπιστημίου Θεσσαλίας κατά το ακαδημαϊκό έτος 2019-20.





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**

**Lexicography for Specialised Languages,  
Terminology and Terminography**







# Audio Recordings in a Specialized Dictionary: A Bilingual Translation and Phrase Dictionary of Medical Terms

Sviķe S.<sup>1</sup>, Šķirmante K.<sup>2</sup>

<sup>1</sup> Ventspils University of Applied Sciences, Latvia

<sup>2</sup> Ventspils University of Applied Sciences, Latvia

## Abstract

The present climate of insufficient funding is having an impact on the development of dictionaries such that new projects would benefit from employing as their source already-existing language material, which could then be made available for publication in contemporary electronic forms. This study reviews the design and development of a bilingual translation and phrase dictionary of medical terms in the form of mobile application “English-Latvian-English Phrasebook and Dictionary of Medical Terms” (called MED). This electronic dictionary is the result of a collaborative effort from researchers from two Latvian higher education institutions, namely Riga Stradins University (RSU) and the faculties of Translation Studies (FTS) and Information Technologies (FIT) of Ventspils University of Applied Sciences (VeUAS). The dictionary presents in a systematic manner the Latvian and English language terminology found in the study materials from RSU’s specialty study courses. The collected terminology was thoroughly reviewed for relevance and supplemented with additional terms during the development of the dictionary. The need for such a dictionary was verified through a survey carried out before the implementation of the project. The successful development of the dictionary has benefitted considerably from VeUAS researchers’ prior experience in the development of electronic dictionaries (in the form of mobile applications) and the expertise of RSU’s medical specialists. As well as describing the functionality of the dictionary, this study describes the database model used in its development and provides an insight into the execution of the project. Additionally, it offers a detailed description of the creation and implementation of a particularly salient feature of the mobile application, namely audio recordings of terms and phrases.

**Keywords:** Audio Recordings, Specialized Dictionary, Mobile Application, Medical Terms.

## 1 Overview of the electronic dictionary’s macro and microstructure

In this section, only a brief overview of the structure of the dictionary is presented. The present description is meant to provide a general context for the ensuing discussion of the creation and development of the application’s audio recordings and the related technical solutions analysed in this article. For the design of the dictionary’s macrostructure, the requests and recommendations obtained from responses to a custom-made survey were taken into consideration. The results of the survey were presented at the international scientific conference “The Word: Aspects of Research”, organized by Liepaja University on November 28 and 29, 2019 in Liepaja.

As shown in Figure 1, the macrostructure of the dictionary consists of the following sections, found in the form of a menu in the mobile application:

- Home. Contains a term search page.
- Info about MED. Contains informative texts about the project and the project executors, as well as a description of the mobile application.
- Fields of Medicine. Shows various available subfields of medicine, each containing relevant terms from among those compiled in the dictionary.
- List of sources of Information. Provides detailed information about all the literature sources used.
- Educational games. Provides access to games and interactive exercises for learning medical terms.
- Review of Latvian Grammar. Offers an overview of Latvian grammar with examples.
- Language Change Menu. Allows the user to switch the application’s interface language between Latvian and English with the help of a “toggle” button.



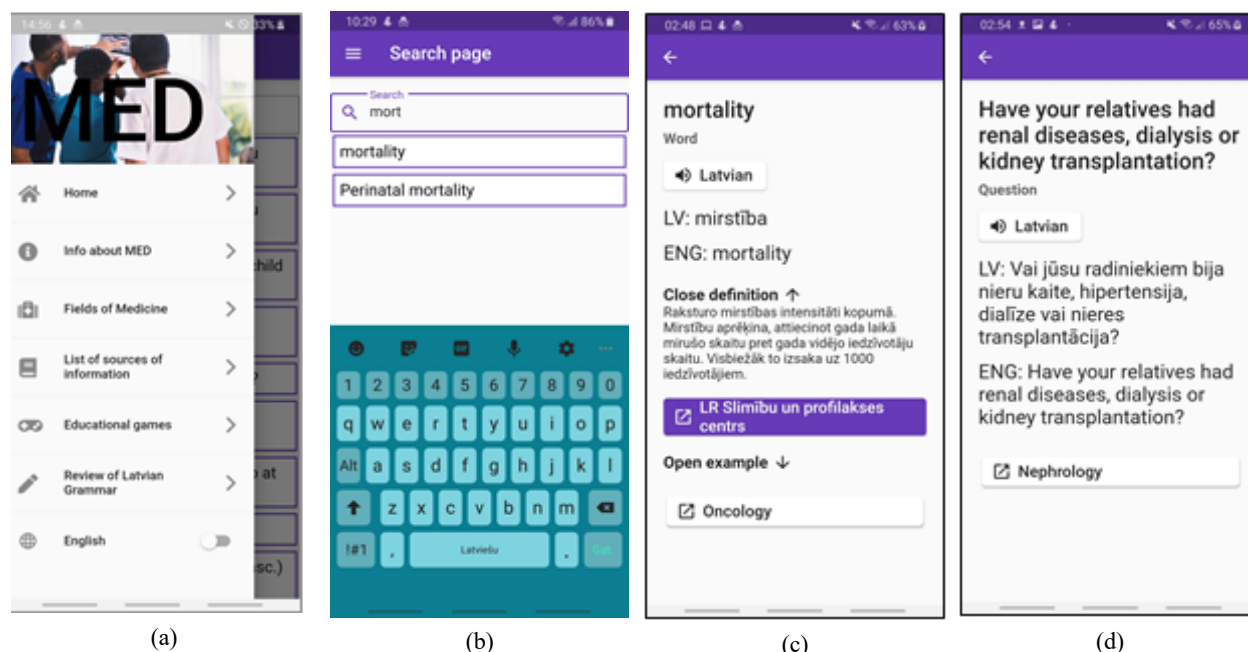


Figure 1: Screenshots of the mobile application. (a) Main menu; (b) Search view; (c) An example of the translation results for the search term "mortality"; (d) An example of the translation results for a question.

The main view (see Figure 1 (b)) is intended for entering medical terms and searching for their equivalents, in the English-to-Latvian or Latvian-to-English combinations. The view has an input field where users can enter a search term using the keypad of their smart device. An additional function incorporated into the application in order to improve user convenience is the option of searching by entering only a part of the word or term. That is, while entering a part of the search term into the input field, all dictionary entries containing said part of the term are shown. After the user chooses from the suggestions displayed, the system searches only for the selected term and the final results appear in the main view. After selecting the corresponding record, the application finds the given term in the database and returns the translation (equivalent), with the option of playing an audio recording with its pronunciation.

The MED dictionary includes three types of entries: English and Latvian terms, phrases, and questions. The need for such a dictionary, containing the features above described, was verified through a survey conducted among medical students and future doctors, and carried out before the commencement of the project. The survey revealed that set phrases and questions are very necessary and useful for communication between doctors and patients. This aspect encouraged developers to include them in the dictionary, together with their corresponding audio recordings. The core section of the dictionary also includes 200 terminological units from the medical field, selected with a focus on terms that might present particular difficulties in translation. The definitions of these terms include hyperlinks to additional information sources, as well as to contextual examples of their use in medical texts.

## 2 Practical work

The electronic dictionary as a mobile application (Android version and iOS version) was created using Google Flutter Framework<sup>1</sup>. The development of the mobile application's Android version was the initial priority. After the testing phase of the application using multiple Android mobile phones and emulators, additional development and configuration tasks were carried out in order to compile the mobile application version for iOS. The Android version of the mobile application was developed on the basis of Android API 19 and using Android Studio. In the testing of this version, various testing emulators (from Android 4.4 to Android 10) and physical smart devices (Xiaomi Mi Note 3 and Samsung Galaxy A7) were used. The iOS version of the mobile application was configured, recompiled and rebuilt using Xcode 11, and iPhone 6 (with iOS version 8.0) to iPhone 11 (with iOS version 13) were used as testing emulators. Development related to mobile application publishing in Apple Store (application version of iOS) and Google Play (application version of Android) is underway. Application development phases were carried out using the Agile Scrum<sup>2</sup> method.

For data storing, SQLite database technology was used for both versions of the applications. SQLite database technology provides data storage in a local database, taking into account the specifics of the dictionary and the interest shown by survey respondents to use applications without the need of a Wi-Fi or mobile internet connection.

Google spreadsheets were initially used as the working tables where researchers stored their research results. This working environment had also been used in previous projects on electronic dictionaries implemented by VeUAS

<sup>1</sup> Google Flutter Framework, more information here: <https://flutter.dev/>

<sup>2</sup> Agile Scrum development method, more information here: <https://www.scrum.org/resources/what-is-scrum>



(Rudziša, Sviķe, Štekerhofa 2019: 379–391; Sviķe, Stalažs 2019: 418–429; Sviķe, Šķirmante 2019: 1–17). The terms, compiled from both research and education institutions, were arranged in Google spreadsheets tables. To convert a dictionary from a Google spreadsheets format to a database format, new software was developed using JAVA programming language and the external library Apache POI (Java API for Microsoft Documents)<sup>3</sup>, designed to manage Microsoft Word and Excel documents using the JAVA application. The software can autonomously create the database model and its tables, including terms. The database model, shown in Figure 2, is based on the document structure of the electronic dictionary.

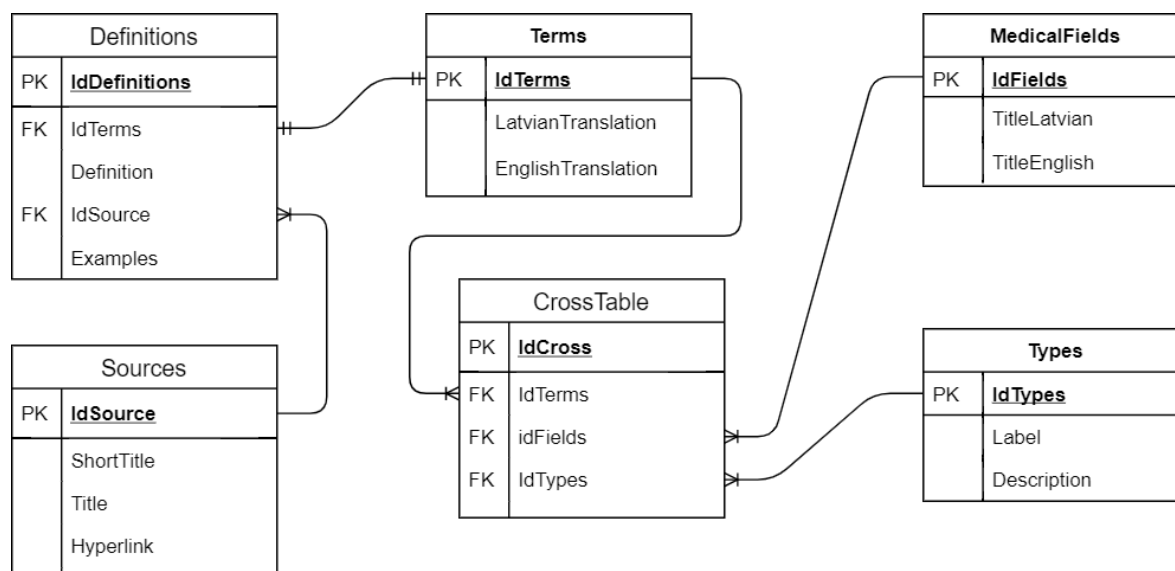


Figure 2: Application database model with relations

Recorded audio files were stored in the mobile application's resources, where each dictionary term or phrase is stored together with its own audio file, with a name corresponding to the identifier of the specific term assigned by the database. For example, the term "blood-vessel" is stored in the database table "Terms" with the IdTerms number 305, and its corresponding audio file title is "305.mp3". The application contains a total of 3785 terms, phrases and questions related to the medical field which are stored with their associated audio files.

### 3 Recording and processing of the dictionary's audio files

The Sanako Lab 100 software was used to record and process the dictionary's audio files. RecordPad Sound Recorder and Audacity were used for re-recording any faulty recordings. For this project, a group of 4th-year students from VeUAS' Faculty of Translation Studies made the audio recordings as part of their compulsory scientific practice. Students were already familiar with the Sanako Lab 100 recording system as the software had been used in their interpretation laboratory during interpretation classes, working in teams or pairs (where one student manages the recordings from the teacher's computer with the Sanako Lab software).

To simplify the recording process of the 3785 audio files, all dictionary terms were grouped in sets of 10, and all the terms of each set were recorded in a single file, with a specific amount of time assigned to each term (5, 10 or 20 seconds). For example, the first term was pronounced at second 0, the second term at second 5, the third term at second 10, and so on, obtaining in this way silent intervals between terms. The recording was done using the established random access method (Behymer 1974) with a fixed-length application. Each term had an individual database identifier (ID) assigned, and the recordings were named according to a numeral scheme indicating the IDs of the first and last terms recorded in the file: e.g. 230\_239.mp3. For the automated processing of all audio files, a Python script was developed. The script split each file into 10 sections, each section being then processed separately and stored in a new audio file with the term identifier as the file name. Therefore, from the source file 230\_239.mp3 the single-term files named 230.mp3, 231.mp3..., 239.mp3 were obtained. The Python script was then used to process each section of the term file by first removing the silent sections created when using a random access method to extract single terms, and subsequently decreasing the bitrate to 64k to reduce file size. A typical 5-second recording was around 215KB in size before processing, but reduced to around 10KB after processing, without incurring in any loss in the quality of the recording. The final total size for all of the application's audio files is 48.9MB, with an average file size of 13.98KB for each term, phrase or question.

While audio recordings from different speakers are considered an advantage, especially in the case of learning dictionaries (Garrett 2019: 201), the individual voice qualities of each speaker, speed of speech and lagging were the

<sup>3</sup> Apache POI Project, more information here: <https://poi.apache.org/>



cause of some difficulties during the processing the recordings. A significant issue was the adaptation of file processing methods to each speaker to remove silent parts from the recorded audio files. For example, the intensity of sound pronunciation differed among speakers (notably in the case of letters 's', 'k' and 'p'), and therefore it became necessary to process first the sound intensity of each speaker for various sounds, and only then single-term audio files could be automatically obtained.

#### 4 Used technologies and techniques

The technologies and techniques used in the development of the application are shown in Figure 3. Application development processes and workflows are described using arrows.

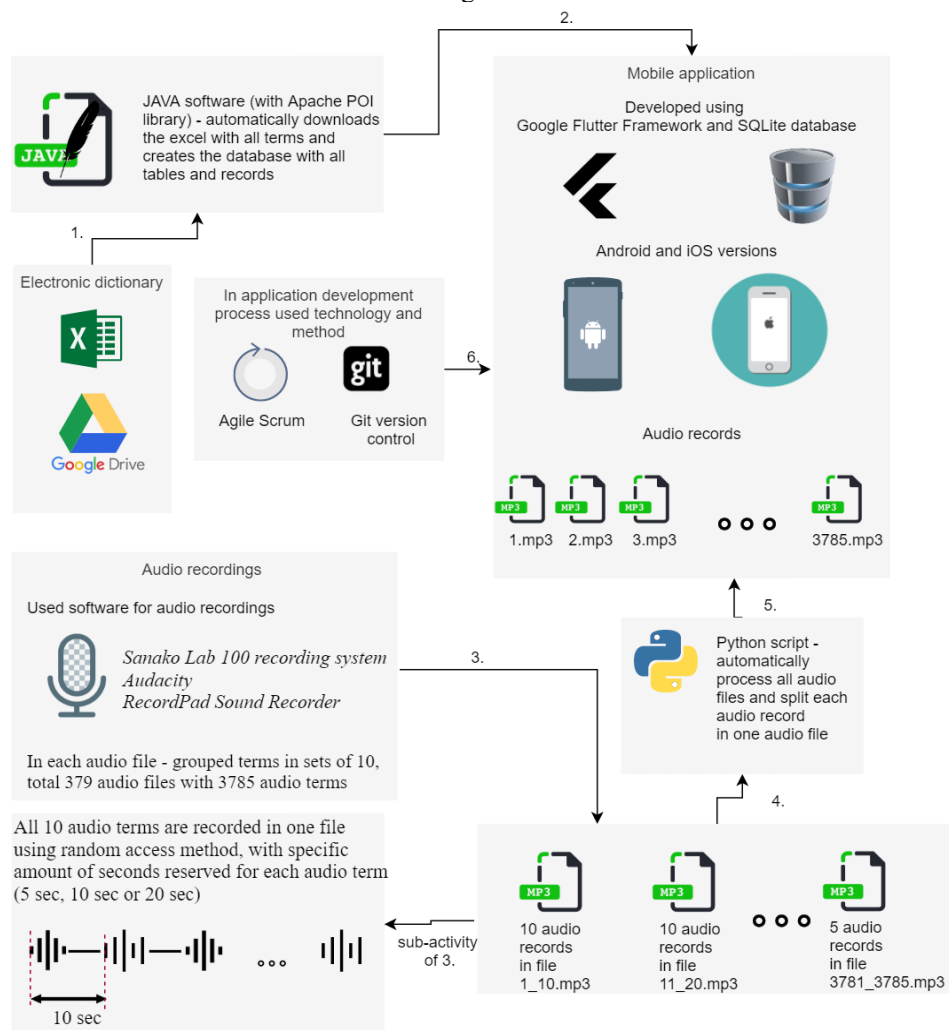


Figure 3: Used technologies and techniques

An automatic download from Google drive was obtained using JAVA software, which was then used to process data from the spreadsheet tables and create an SQLite-type database containing the databases' tables, relations, Primary Keys (PK), Foreign Keys (FK), as well as data. The Database model is shown in Figure 2. When a SQLite database file is created, it is added into the application project. The same database has been used for both versions of the application (Android and iOS).

Application development and audio recording were tasks carried in parallel. To simplify the recording process for the speakers, ten (10) terms were stored as a single audio file. All recorded files were stored in shared directories (one directory per speaker), as the recording process spanned across several sessions and it was necessary to control the progress of the recordings. Using shared directories made it easy to manage the re-recording faulty recordings. As illustrated in Figure 3, each audio file consisted of an audio wave which included recorded pauses with lower sound intensity so they could later be processed and split automatically. Each speakers' sound intensities for different pronunciations were analysed and taken into account before the automated processing of audio files.

During the application development process, a custom Python script was written to carry out an automated splitting of the recorded audio files into single-term ones. Each audio file title corresponds to the term's primary key in the database, thus providing faster and more efficient audio file searches in the application during playback. After file processing, the resulting audio files were stored in the application resource directory and not in the database, since retrieving a term's audio file from a resource directory is faster than its retrieving and decoding from a database.



The Agile Scrum method was chosen to organize the development process more efficiently. Each development phase consisted of 2-week stages called ‘sprints’, with specific results obtained at the end of each ‘sprint’. To share application source code between programmers and to manage source code versions during different phases of the development process, the version control and GitHub software development platform were used.

## 5 Conclusion and Future Work

This research forms part of a project concerning the conception and development of a bilingual translation and phrase dictionary of medical terms, as well as the creation of a working model of the dictionary in the form of a mobile application. Aiming to maximize the dictionary’s usefulness, there exists the possibility of adding new terms, phrases and even sections in the future. The research work here described may be continued in the future in order to expand its theoretical scope, improve the dictionary’s current list of entries, and modernize the mobile application according to newfound technological possibilities and users’ needs. This study offers a detailed description of the process for the incorporation of audio recordings to the application and characterises the technologies used in the dictionary. The conclusions and suggestions of the present research could be particularly useful in the development of future translation and phrase dictionaries of this type, especially considering the dynamic “shift from p-lexicography to e-lexicography” (Tarp 2012: 107–119).

At the end of the project, it has been concluded that only one voice per gender should be selected to be used in the audio files of the dictionary, as this would provide consistency throughout recordings. For future recordings, each speaker’s sound intensities for different letters could be analysed using neural networks. Research in input data corrections algorithms using artificial neural networks in mobile applications is currently underway and soon it might be possible to implement relevant research results in the developed mobile application.

The data material incorporated into the dictionary was divided and compiled according to the areas of responsibility of the working group members of both research institutions. It should be taken into account that when working in large groups it is crucial to monitor completed work, plan regular group meetings, establish future steps, identify problematic issues and implement adjustments. Since the dictionary project was executed in about a year, it was necessary to continuously check whether all steps of the working process were being executed. After the completion of the project, it has been concluded that the work entrusted to the students must be checked thoroughly.

## References

- Behymer, J.A., Ogilvie, R.A., Merten, A.G. (1974). Analysis of indexed sequential and direct access file organizations. In *SIGFIDET '74: Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, access and control*, pp. 389–417.
- Garrett, A. (2019). Online Dictionaries for Language Revitalization. In L. Hinton, L. Huss, & G. Roche (eds.) *The Routledge handbook of language revitalization*. Abington-on-Thames: Routledge, pp. 197–206.
- Rudziša, V., Sviķe, S., Štekerhofa, S. (2019). Juridisko pamatterminu glosārijs līgumtiesībās Latvijā izdoto nozarvārdnīcu kontekstā. In G. Smiltņiece, L. Lauze (eds.) *Vārds un tā pētīšanas aspekti*. 23 (1/2), Liepāja: LiePA, pp. 379–391.
- Sviķe, S., Stalažs, A. (2019). “Jaunās botāniskās vārdnīcas” mikrostruktūra: tradicionālais, mainīgais un inovatīvais. In G. Smiltņiece, L. Lauze (eds.) *Vārds un tā pētīšanas aspekti*. 23 (1/2), Liepāja: LiePA, pp. 418–429.
- Sviķe, S., Šķirmante, K. (2019). Practice of Smart LSP Lexicography: The Case of a New Botanical Dictionary with Latvian as a Basic Language. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. 1-3 October 2019, Sintra, Portugal, Brno: Lexical Computing CZ, s.r.o. pp. 1–17.
- Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 107–118.

## Acknowledgements

This research has been funded by the Latvian Council of Science, project “Smart complex of information systems of specialized biology lexis for the research and preservation of linguistic diversity”, No. lzp-2020/1-0179









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**

**Historical and Scholarly Lexicography and Etymology**







# Paper Quotation Slips to the Electronic Dictionary of the 17th- and 18th-Century Polish - Digital Index and its Integration with the Dictionary

Bilińska-Brynk J.<sup>1</sup>, Rodek E.<sup>2</sup>

<sup>1</sup> University of Warsaw

<sup>2</sup> Polish Academy of Sciences

## Abstract

The paper presents the results of experimental paper quotation slips' tagging that was conducted to investigate the possibility of electronic indexing of scanned paper quotation slips constituting a citation archive (a card-index) for the *Dictionary of the 17th- and 18th half of the 18th-Century Polish* (e-SXVII <https://sxvii.pl>).

The paper citation archive consists of more than 3 million paper quotation slips posing an exemplification of ca. 116,000 of words, which means 86,000 dictionary entries – all of them placed in 836 boxes. There is the need for integration of the archive and the lexicographic panel in order to accelerate the lexicographic work and eliminate human-related mistakes. The test allowed the authors to determine the project priorities, main methodological problems and to decide on future project proceedings. The presented case study may be interesting for other lexicographic teams facing the same problems and looking for an efficient, cheap and quick solution to the problem of using such an abundance of available data.

**Keywords:** quotation slips, historical dictionary, indexing, card-index, citations archive

## 1 The need for lexicographic tools modernisation

The Electronic Dictionary of the 17th- and 18th-Century Polish (e-SXVII <https://sxvii.pl>) records vocabulary that comes from Polish baroque and the early Age of Enlightenment, which is known as the Middle Polish language period (cf. Bronikowska et al. 2020). The dictionary history goes back to 1954, when *Polish Language of 17th- and 18th-Century Research Group* was established at the Institute of Polish Language Polish Academy of Sciences and dictionary data excerption started (Majdak 2012 [21/01/2020]; 2018: 177-178; Siekierska & Sokołowska 1999: IV [19/03/2020]). Citations archive containing lexical data with its context usage was preserved on paper quotation slips, filed and stored in boxes in alphabetical order<sup>1</sup>.

Initially, a full excerption was held, i.e. the whole lexical resources from 422 marked sources from the selected age were analysed; however, at the beginning of the 1990s, when there were already 83 thousand of words recorded in the paper catalogue, the decision was taken about starting a non-full excerption (selective, called reasoned) (Siekierska & Sokołowska 1999: VI [19/03/2020]). Finally, in the paper citation archive (KXVII) there are stored 2.8 million entry cards that confirm the existence of 116 thousand words (about 86 thousand entries) attested in 275 texts (various volumes) that functioned in 1601-1750 (Siekierska 1998: 84-86). Apart from that, there also originated a paper catalogue and an index of proper names (ca. 11 thousand entries), a paper catalogue and an index of foreign words (ca. 7 thousand entries), a paper catalogue of entries for Michał Abraham Troc's<sup>2</sup> dictionaries. All of them are now rarely used by lexicographers mainly because of the lack of electronic editions.

Since the moment of huge transformation of the dictionary form and its development into an electronic one (Bronikowska et al. 2020; Gruszczyński 2005) there have also been essential changes in its production process. There emerged the need for developing tools enabling online work but basing on foregoing sources. Therefore, the whole electronic index of the card catalogue that is now essential for dictionary entry editors was prepared. Unfortunately, the index is only a list of headwords recorded in the catalogue; hence, it contains only the basic information about the existence of the word, but does not make it possible to quickly reach the quotation slips and verify their contents.

The current analogue form of the card catalogue makes it gradually useless, and uncomfortable – to say the least – both for the usage and preparation, as well as archaic. It is possible to use either paper card catalogue stored in the dictionary editorial office or its digitised version available as scans in DjVu format on the Digital Repository of Sciences Institutes server (RCIN; <https://rcin.org.pl/dlibra/publication/20029>). The scans are ordered by cards distribution in boxes, i.e. mainly in alphabetical order (with exception of phonetic variant forms filed together with the main entry), not numbered and only divided into sections corresponding to the boxes volume, i.e. containing ca. 3,5 thousand quotation slips. Searching through such a collection is quite troublesome. Moreover, it is essential to remember that there is the need for maintenance of the server containing the scanned catalogue, which is not a property of the Polish Language Institute Polish Academy of

<sup>1</sup> More on the history of e-SXVII (Gruszczyński 2005; Majdak 2018; Siekierska & Sokołowska 1999 [19/03/2020]).

<sup>2</sup> M.A. Troc was the author of three dictionaries: *Nouveau dictionnaire françois, allemand et polonois*, v.1 (1744) and v.2 (1747), Lipsk, and *Nowy dykcyonarz to iest mownik polsko-francusko-niemiecki*, v. 3 (1764), Lipsk. All of them have been incorporated into the source canon for e-SXVII.



Sciences, which is the owner of the KXVII catalogue<sup>3</sup>.

Furthermore, making KXVII available in a new digital form is becoming more and more necessary as there are no more djview plugins in web browsers in the form they used to be in the past and this makes opening scans stored as DjVu files more difficult, non-user friendly and troublesome or even almost impossible for some users.

When the Electronic Corpus of the 17th- and 18th-Century Polish Texts (to 1772) (KorBa; <https://korba.edu.pl/>) was launched, it became another significant citation source for dictionary entries and, as it is accessible online, it developed into the main source of material exemplification for the entries<sup>4</sup> because the editors do not use now the paper KXVII as frequently as before.

It must not be, however, forgotten that there are quotations in the paper catalogue from the sources that are available neither in KorBa nor in online digital libraries, nor as paper books in the e-SXVII editorial office. It is a major impediment for the editors as they can observe a given word only in the quotation slip without being able to check the broader word context. Frequently, those are the only records of the words or variant forms of the basic words. Moreover, KorBa does not collect many texts that constituted the basis of the first dictionary material excerption and are recorded in KXVII<sup>5</sup>.

In conclusion, it has to be admitted that there is a major need to balance the meaning of both citation data sources – the KXVII paper catalogue and the KorBa corpus – and transform an analogue citation archive into a functional tool for the e-SXVII editors.

## 2 Analogue quotation slips in modern lexicography

There are multiple approaches to the retrodigitisation of the dictionaries. There is a possibility of scanning the dictionary, performing OCR, data encoding and data enrichment (Kallas J. et al 2020: 26). In the e-SXVII, which is a digitally born dictionary, we aim to integrate the dictionary with the data collected in the paper form. Therefore, we need to scan it and encode the data.

The idea of digitising and making paper quotation slips available to the users is not new. There have already been projects involving such actions, e.g. *Dictionary of Old Norse Prose* (ONP) (cf. Johannsson 2019). However, there was another approach to the issue of citations and their results and goals are different from the goals set in KXVII project.

In case of a *Dictionary of Old Norse Prose*, it was decided that the citation archive would be scanned for the needs of the integration with a dictionary so the editors started with finishing the headword list they already had and then scanned the quotation slips. This part of work has already been done when it comes to KXVII, but it was not aimed from the very beginning at making the scans a part of the dictionary itself so the scans were neither stored nor named in a proper form for the processing that is now necessary. After making the headword list and scanning the quotation slips in the ONP project, the editors processed the quotation slips multiple times and used database to store them, including tagging the scans. The process took ca. 3 years (Johannsson 2019: 254). Then the dictionary assistants keyed in relevant citations into the database. “The edition and the citations are linked by the sigla in the database (...)” (Johannsson 2019: 255). At the end of the process the editors had completely processed scans and information stored in quotation slips that they would further use. However, for the needs of our project, where there is a shortage of financial sources and time, there is a need to shorten the process of making quotation slips handier for the editors and useful for the users. Therefore, we would like to make only some of the work similar to that done in the ONP project and we need our own database and interface to integrate it with our dictionary interface.

## 3 The experimental indexing

In order to transform card-index KXVII into a form that would make working with it comfortable and efficient, there was designed a project of indexing quotation slips and incorporating them into the respective entries in e-SXVII in both editor and user interfaces. Therefore, five randomly chosen card boxes were indexed in aim to optimise the future indexing of the catalogue. These were card boxes containing headwords starting with the letter from the middle of the alphabet (ca. 17,5 thousand paper quotation slips which means 0,625% of the collection). It was done by means of djview4poliqarp software (Bień 2016 [16/03/2020]). The software was initially developed to enable searching digitised texts in the DJVu format and was later supplemented with an indexing function. In the assignment we used the latter functionality. Although it turned out that the software was not efficient and suitable enough for the planned work, the test allowed us to estimate the time and cost of preparing a new index to the citation archive and also to determine the project priorities, main methodological problems and to decide on future project proceedings. Please compare the figure 1 with the example of a scanned paper quotation slip and a test index in the djview4poliqarp software. On the left there is a scan and on the right there is an index where the first and the last scan of the scans’ scope containing examples of the headword are tagged. After the word there is a number informing whether it is a first scan or the last one and how many scans are in the scope and the brackets inform whether it is a beginning of the scope (left square bracket) or the end (right square bracket). Therefore eg. *marszczyć się 0020]* (English to *pucker; cockle*) means that there are 20 scans in the scope and it is the last scan.

<sup>3</sup> RCIN was established as a scientific institutes consortium with EU and Polish government financial sources, but it is now supported by institutes themselves (<https://rcin.org.pl/dlibra/text?id=Projekty> [14/07/2020]). Therefore, there is neither stability nor guarantee that the collected resources will be available in the future as it depends on the current economic and political situation.

<sup>4</sup> The first KorBa edition was published in 2013-2018 and since 2019 the second edition has been prepared. It is now being expanded (it will contain 25 million tokens) and the time horizon was moved until 1800, which is 50 years longer than in the card index (<https://korba.edu.pl/overview> [14/07/2020]).

<sup>5</sup> In the second edition of KorBa there will be a large amount of texts that are important lexicographic sources, but which were not incorporated into the first edition of the corpus.



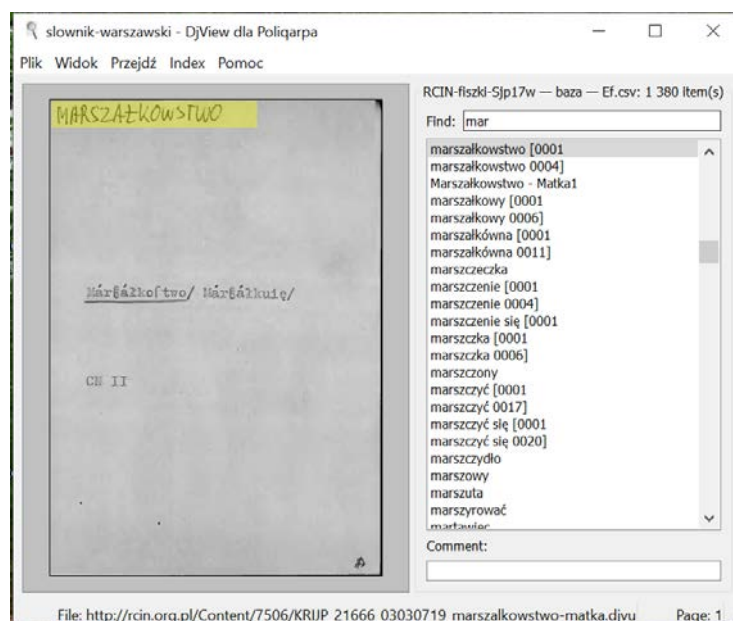


Figure 1. A screenshot of the djview4poliqarp software, scanned paper quotation slip *marszałkstwo* (English *marshalcy*) and the test card index

The software allows, among other things, to tag scans (in our case scanned quotation slips) with e.g. the entry name that this card relates to. As there are multiple quotation slips for every headword (sometimes even hundreds), it was decided that only the first scan and the last scan were tagged, which made it possible to divide the scans into groups relating to one word and number them. We would like the software to number automatically also the rest of the scans within the scope which would allow the editor to orientate themselves in the selection place. It should be possible thanks to the card meter and special tags.

#### 4 The results

Thanks to the experimental tagging there were determined the priorities of the future project:

- Maximum comfort for the index user, which de facto means collecting as much data in one place as possible.
- Annotating person's work optimisation, i.e.:
  - minimising the time needed to annotate a scan;
  - avoiding situations where an annotator would need to make their own decisions;
  - minimising the possibility of errors.

It sometimes happens that the quotation on the slip does not apply to the headword and illustrates a different word from the one written on the card. However, the headword is shown in the previous index in the editor interface and, therefore, the new index should not be ordered newly and the tagger should not be given the permission to encode the proper headwords. Only the entry editor should verify the quotation slip content while working on the entry. In this way the person tagging the archive does not need to have any special scientific skills to perform the task, which would make it possible to outsource tagging and thereby relieve the e-SXVII editorial board work.

The main problems that occurred was how to tag quotation slips in two specific situations:

- 1) Presentation of two headwords on one quotation slip that disturb the alphabetical order. Phonetic and morphological variants that are illustrated in the quotation were additionally lemmatised (cards are dually lemmatised) and incorporated into the card scope without preserving the alphabetical order. Therefore, there is no possibility to prepare a linear index, e.g. within the scope of the cards containing quotations for the entry *maska* (English *a mask*), one can find also not ordered cards with the variant form *maszka* and also cards with two headwords signed as *maska*, *maszka*. Compare examples in the figure 2.



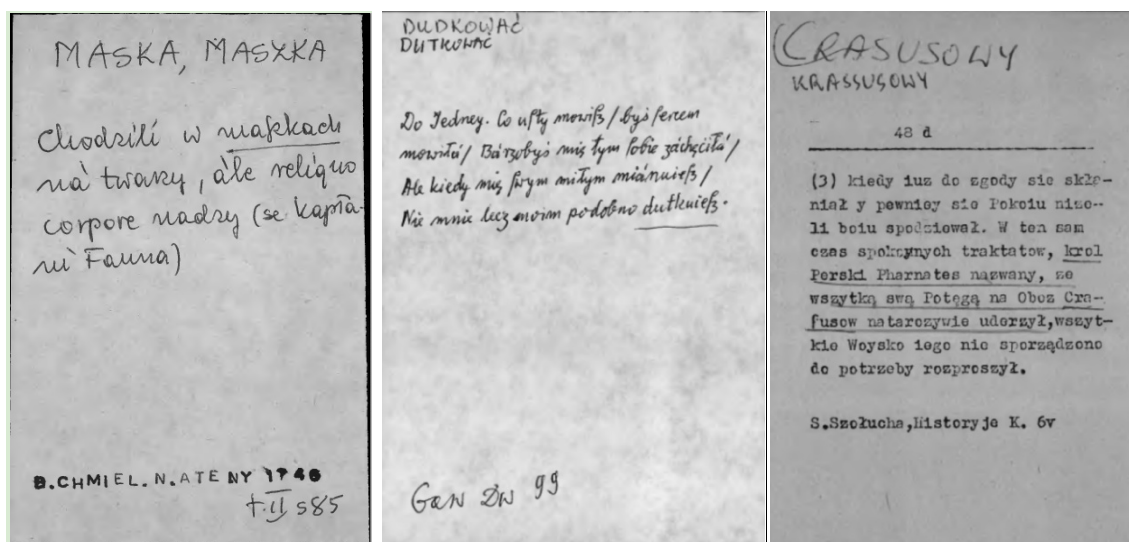


Figure 2. Examples of paper quotation slips with two headwords

## 2) Accidentally misplaced individual quotation slips.

Both problems are fixable by the modification of the software search engine query syntax, e.g. by enabling searching for any character string instead of conducting an alphabetical search as is being done now. Then there would be no searching problem with cards tagged 'X a. Y' ('X or Y', where X and Y mean headwords).

It also seems impossible to accelerate the work using the OCR software as the headwords are handwritten and there have been multiple quotation slips editors. Moreover, the vocabulary is historical both when considering the forms themselves and spelling (e.g. *dedukcyja*, English *a deduction*, now spelled as *dedukcja*) and there also sometimes occur quotation slips that contain postprepositional phrases as their headwords (e.g. *po frantowsku*<sup>6</sup>) that are now lemmatised in a different way. Furthermore, the tagging person could tag the quotation slips according to the current e-SXVII headwords preparation guidelines, e.g. passive adjectival participles (*frasowany*, English *worried*) are lemmatised as infinitives (*frasować*, English *to worry*) and only when there is no record of the personal verb form should they be lemmatised as participle forms. Thus, in the new electronic index there would sometimes be a different headword from the headword written on a paper quotation slip.

The analysis allowed to estimate the time needed for the project and indirectly also its cost. When using the chosen software in its current form with the already designed editorial guidelines, one needs ca. 3 hours to tag one box of quotation slips, which means 2,508 hours necessary to perform the task (836 boxes) by one person, which results in ca. 16 months of one full-time working person. However, it is considered impossible to work efficiently 8 hours per day doing such a task; hence, the numbers would certainly increase.

The above-described possible changes to the djview4poliqarp software could improve and accelerate the tagger's job and also eliminate the main query problem. It would considerably decrease costs of indexing the archive and speed it up. The estimated time per box would be two hours, i.e. ca. 10 months of one-man job.

However, the software is standalone, which means that one has to install it on a computer and then, if there are many people working on the same project, which would be expected in the described case, there is a need to merge their work afterwards (in CSV files). It leads to other possible errors and even to data loss. In this case it would also be more demanding and time consuming to manage the project as well as to control contractors' work than while using an online application programme. It seems that developing such software for the needs of the project would be the best solution to the problems described. It would make it possible to do the task remotely, control the tagger(s) in the real time, facilitate assigning material to the contractors and, thanks to the necessary backup option, avoid the danger of possible data loss. It would shorten the estimated indexing time to ca. 1.5 hours per box (ca. 8 months per one person), which is around half of the time needed when using the current software option. It should also be possible to integrate the new index with e-SXVII, which is one of the basic assumptions of the task.

## 5 Integrating KXVII with e-SXVII

We strive to transform KXVII into a tool for e-SXVII editors that will be functional and user-friendly. Presently, the dictionary is in the phase of creating the so-called stub entries, i.e. entries containing at least one recorded form and illustrated in the quotation grammar form (cf. Bronikowska et al. 2020). The editors can use one of the citation sources and more often they opt for KorBa than KXVII. Moreover, KorBa has already been integrated with e-SXVII (both with the

<sup>6</sup> It is a combination of the preposition "PO" with an adjective in Dative in the so-called short historic form. However, as the Polish short adjective forms were already linguistic relicts in the 17th-18th. c. we treat them in e-SXVII as morphological variant forms and record them in dual entries, e.g. BEZPIECZNY, BEZPIECZEN. But in case of adjectives ending in -ski we do not recreate a possible short form, but we only record long ones.



editor's interface and user's interface, cf. Bronikowska et al. 2016 [23/01/2020]; Bronikowska et al. 2020). However, it has to be noted that a new KXVII index integrated with the scanned quotation slips (as distinct from a present index being only an entry list) and its implementation in the editor's interface will be crucial in the next stages of e-SXVII development.

The entry editor should have access to each and every archived quotation in a nick of time and to be able to preview the scanned quotation slips in the entry edition part of the interface. Basic functionalities would be: information about the number of quotation slips relating to one headword, reference to the place of the slip in the collection, marking/tagging scanned quotation slips as already processed. It is crucial to give a possibility to change the order of the cards and grouping them, e.g. depending on the meaning while editing the entry. It could be useful if the editor could also reject wrongly matched, false or blank quotation slips.

KXVII is, on the one hand, a lexicographic tool and should be integrated with the e-SXVII editor's interface, but, on the other hand, being an archive of quotations coming from the works written in 1600-1750, it can be interesting for the researchers working on the language of this period and, therefore, it could also be available for e-SXVII users.

Admittedly, the process of data excerption took 40 years and during this time there were various decisions made when solving some problems and the data is not fully consistent. For instance, there are paper quotation slips with only a headword and reference information<sup>7</sup> of the quotation without the quotation itself. On some slips there are also illegible (or difficult to read) handwritten notes concerning grammar forms or meanings made by numerous editors. The abbreviations used originally have also changed as the growing number of dictionary sources required doing so in a more systematic way. However, lists of abbreviations are already prepared in both forms allowing users to decode the meanings and available to the e-SXVII users. The KXVII scans are already available in the public domain so it is possible to use it now, but it would be quicker if they were incorporated into the e-SXVII entries.

A close communication between quotation slips and dictionary entries, their incorporation into the entries, especially germ ones, would enable showing the whole collected but still not processed material. Presently, at the e-SXVII website there is a button "More quotations in Baroque Corpus", which automatically generates an appropriate query to the corpus and the results are shown on the corpus website in the new web browser tab or window. The integration of the KXVII with the dictionary user interface could be done in a similar way enabling the user to get access to all the available quotations not only those incorporated into an entry.

## 6 Summary

The quickly changing reality, new technologies development and growing users' expectations also provoke changes in lexicographers' work. There is a need to adjust the tools in order to be able to take advantage of the material efficiently and fully.

The integrated paper citation archive would be especially useful both for the e-SXVII editors and its users, but can also be interesting for other linguists researching the language of that time. As it is historic material, there are e.g. untemporised *appellative names* that would be interesting for some researchers.

Developing new indexing software would also enable quick processing of the three smaller abovementioned collections that were produced during the excerption of dictionary sources: the card catalogue and index of proper names, the card catalogue and index of foreign words (7 thousand paper quotation slips) and M.A. Troc's dictionary card catalogue.

We suppose that our approach to the analogue quotation slips may be interesting to other lexicographic teams or even libraries that face the same problems and are looking for an efficient, cheap and quick solution to the issue of using the abundance of available data.

## 7 References

- Bień, J.S. (2016). Elektroniczne indeksy fiszek słownikowych. In *Kwartalnik Językoznawczy*, 2, p. 16-27. Accessed at: <https://doi.org/10.14746/kj.2016.2.2> [http://pmichal-kwartjcz.home.amu.edu.pl/teksty/teksty2016\\_2\\_26/Bien.pdf](http://pmichal-kwartjcz.home.amu.edu.pl/teksty/teksty2016_2_26/Bien.pdf) [16/03/2020].
- Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M. & Woliński, M. (2016). The use of electronic historical dictionary data in corpus design. In *Studies in Polish Linguistics*, vol. 11, issue 2, pp. 47-56. Accessed at: <https://doi.org/10.4467/23005920SPL.16.003.4818> [23/01/2020].
- Bronikowska R., Majdak M., Wieczorek A. & Żółtak M. (2020) The Electronic Dictionary of the 17th- and 18th-century Polish. In (eds) *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion.*, (in the current volume).
- Gruszczyński W. (2005). O przyszłości „Słownika języka polskiego XVII i 1. połowy XVIII wieku”. In *Poradnik Językowy*, 7, pp. 48–61.
- Johannsson, E. (2019). Integrating analog citations into an online dictionary. In C. Navarretta, M. Agirrezabal & B. Maegaard (eds.) *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pp. 250-258.
- Kallas J. et al (2020). D1.1 Lexicographic Practices in Europe: A Survey of User Needs. Accessed at: <https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS-D1.1-Lexicographic-Practices-in-Europe-A-Survey-of-User-Needs.pdf> [26/07/2020].
- Majdak, M. (2012). Słownik języka polskiego XVII i 1. połowy XVIII wieku, Kraków 1996- IJP PAN. In *Poradnik Językowy*, 8, pp. 105-111. Also in M. Bańko, M. Majdak, M. Czeszewski, (eds) *Słowniki dawne i współczesne*.

<sup>7</sup> A siglum: resource name abbreviation and page number.



- Internetowy przewodnik edukacyjny. Accessed at: <http://leksykografia.uw.edu.pl/slowniki/21/slownik-jezyka-polskiego-xvii-i-1-polowy-xviii-wieku-krakow-1996> [21/01/2020].
- Majdak, M. (2018). Elektroniczny słownik języka polskiego XVII i XVIII wieku IJP PAN. In M. Pastuch, M. Siuciak (eds) *Historia języka w XXI wieku. Stan i perspektywy*. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 176-182.
- Siekierska, K. (1998). Słownik języka polskiego XVII i 1. połowy XVIII wieku, historia przedsięwzięcia, założenia teoretyczne, plan prac, prognozy na przyszłość. In *Język Polski*, 1-2, pp. 82-90.
- Siekierska K. & Sokołowska T. (1999). Historia *Słownika*, [w:] Słownik języka polskiego XVII i 1. połowy XVIII wieku. T. 1 z. 1, Kraków, pp. IV-VIII. Accessed at: <https://rcin.org.pl/dlibra/show-content/publication/edition/25137?id=25137> [19.03.2020].
- e-SXVII - Gruszczyński, W. (ed.). Elektroniczny słownik języka polskiego XVII i XVIII wieku/Electronic Dictionary of the 17th- and 18th-century Polish. Accessed at: <https://sxvii.pl/> [19.03.2020].
- KorBa - Elektroniczny Korpus Tekstów Polskich z XVII i XVIII wieku (do 1772 roku)/Electronic Corpus of 17th- and 18th-century Polish Text (up to 1772). Accessed at: <http://www.korba.edu.pl> [19.03.2020].
- KXVII - Kartoteka *Słownika języka polskiego XVII i 1. połowy XVIII wieku*. Accessed at: <https://rcin.org.pl/dlibra/publication/20029> [6/04/2020].
- MED - Middle English Dictionary. Ed. Robert E. Lewis, et al. Ann Arbor: University of Michigan Press, 1952-2001. Online edition in Middle English Compendium. Ed. Frances McSparran, et al. Ann Arbor: University of Michigan Library, 2000-2018. Accessed at: <http://quod.lib.umich.edu/m/middle-english-dictionary/> [16/04/2020].
- RCIN - Repozytorium Cyfrowe Instytutów Naukowych. Accessed at: <https://rcin.org.pl/dlibra/publication/20029> [6/04/2020].



# The Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish - Towards the Open Formula Asset of the Historical Vocabulary

Bronikowska R.<sup>1</sup>, Majdak M.<sup>1</sup>, Wieczorek A.<sup>1</sup>, Żółtak M.<sup>2</sup>

<sup>1</sup> Institute of Polish Language, Polish Academy of Sciences

<sup>2</sup> Austrian Centre for Digital Humanities and Cultural Heritage

renata.bronikowska@ijp.pan.pl, magdalena.majdak@ijp.pan.pl, aleksandra.wieczorek@ijp.pan.pl, mateusz.zoltak@oeaw.ac.at

## Abstract

The paper discusses the *Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish* (abbreviated e-SXVII), an important resource for the study of language, history and culture of the period. After several dozen years of gathering material, conceptual work, and after the publication of five fascicles, the print dictionary project was discontinued in favour of a digital version. The work has since accelerated significantly, although development is still ongoing. The paper focuses on new aspects of the methodology stemming from the open form of the dictionary. The innovation offers significant benefits both for the editors and the users. This formula also allows for e-SXVII to be integrated with other electronic language resources, like corpora and digital libraries – a feature currently under intense development.

**Keywords:** electronic dictionaries; integration of linguistic resources; Middle Polish; historical vocabulary

## 1 Introduction

The 17<sup>th</sup> and 18<sup>th</sup> centuries were an important period in the development of the Polish language, when several notable grammatical categories took shape, while others disappeared; vocabulary borrowed greatly from other languages (including Latin, German, French, Turkish and others), whereas style and syntax continued to evolve. Texts published in that period constitute an important source for research into history, culture, literature, history of science, and other fields, and their complete comprehension requires a dictionary. The *Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish* (*Elektroniczny słownik języka polskiego XVII i XVIII wieku*; hereafter e-SXVII)<sup>1</sup> is the first lexicon in Poland to focus on these two centuries and the first ever purpose-built electronic dictionary of Polish.<sup>2</sup> The decision to develop and publish the dictionary entirely in electronic form had significant impact both on the development work, including many solutions within the subject matter, and on its availability and ease-of-access. This approach has the advantage of enabling integration with other electronic resources for historical and modern Polish, such as corpora and digital libraries.

## 2 Polish Dictionaries Noting Historical Vocabulary

The vocabulary of past periods in the history of the Polish language is presented in several large lexicographic works (printed, then digitised). The lexical layer of the oldest works (until 1500) is registered in the *Dictionary of Old Polish* (*Słownik staropolski*, SStp). The vocabulary of 16<sup>th</sup>-century works is described in the still-unfinished *Dictionary of 16<sup>th</sup>-century Polish* (*Słownik polszczyzny XVI w.*, SXVI). The *Dictionary of Polish Language*, edited by W. Doroszewski (*Słownik języka polskiego*, SJPdOr) in the second half of the 20<sup>th</sup> century, was intended mostly as a dictionary of modern Polish, yet it reaches back to mid-18<sup>th</sup> century. This paper discusses a dictionary intended to bridge the gap and augment the documentation of historical vocabulary with resources for the 17<sup>th</sup> and 18<sup>th</sup> centuries.

## 3 Evolution of Dictionary Development – the Print to Digital Transition<sup>3</sup>

Work on a dictionary of 17<sup>th</sup>- and 18<sup>th</sup>-century Polish<sup>4</sup> started in 1954. Its result was a card index of ca. 2.8 million paper slips listing use cases for ca. 80 thousand lexemes excerpted from 275 texts (hereafter referred to as KXVII).<sup>5</sup> The list of sources is supposed to be as representative as feasible and include not only literary works (in the 17<sup>th</sup> and 18<sup>th</sup> centuries written mostly in verse), but also functional texts, such as administrative files, inventories, specialist manuals, dictionaries, phrasebooks, personal documents, letters, and others. The first volume of the *Dictionary of the Polish*

<sup>1</sup> Links for electronic resources mentioned in the paper can be found in the References section at the end.

<sup>2</sup> This is understood as a dictionary that was meant to appear in an electronic version from the very beginning and as such has developed new, more adequate methodology. Digital forms of print dictionaries of Polish (especially modern Polish) have been available before. Although e-SXVII is a continuation of a traditional dictionary (i.e. one made with print in mind), its concept and form are now so different from the original as to consider it a separate work (more on this subject later on).

<sup>3</sup> This paper focuses on the description of the electronic version of the dictionary of 17<sup>th</sup>- and 18<sup>th</sup>-century Polish. The history of the print version can be found in the works of Siekierska (1998), Majdak (2012, 2018), and others.

<sup>4</sup> Originally the dictionary was intended to cover the 17<sup>th</sup> century and the first half of the 18<sup>th</sup> century.

<sup>5</sup> This archive was digitised in 2011 and is now available online (see KXVII).



*Language of the 17<sup>th</sup> and First Half of the 18<sup>th</sup> Century* (Słownik języka polskiego XVII i pierwszej połowy XVIII wieku, SXVII) appeared in print as five fascicles, published between 1999 and 2004. Predictions at the time indicated that if that pace was maintained, the work would be completed in approximately 100 years (per Gruszczyński 2005: 48). The search for ways to expedite development prompted the 2004 decision to move to a digital form. In light of its evolution, this new version of the dictionary was renamed the *Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish* (Elektroniczny słownik języka polskiego XVII i XVIII wieku, e-SXVII).

The changes made in 2004 have to be considered revolutionary for the era, even if today they appear natural, necessary, and obvious. An internal network connected to the internet was created, the usage of computers as just text editors was abandoned in favour of working within a dictionary editor's platform, and a webpage was developed for the dictionary (cf. Majdak 2018: 178). Moreover, the postulated conversion of as many sources as possible into electronic form also started to become reality, giving rise to the *Electronic Corpus of 17<sup>th</sup>- and 18<sup>th</sup>-century Polish Texts (up to 1772)* (Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 roku), KorBa; see p. 6).

#### 4 The e-SXVII IT System

The dictionary authors resolved to offer e-SXVII to the users in the form of a webpage. In order to simplify the architecture of the IT system, the dictionary editor's interface was also developed in the form of a web application.

As for the general concept, the IT system used in the dictionary is simply a content management system (CMS) with dictionary entries in place of posts. A characteristic feature of the e-SXVII system is very precise modelling of the complex structure of a dictionary entry, both in the dictionary editor's interface and within the relational database storing the dictionary data. This solution not only forces the dictionary editors into a unified system of data entry, but also ensures clean separation between the data and the presentation. This, in turn, permits for quick changes to the way e-SXVII is presented to the user (as only the display template is changed, without having to alter the content of the entries), as well as very precise content searches or easy integration with other computer systems. Since frameworks for content management systems with such complex data structures were still in their infancy in 2004<sup>6</sup>, the e-SXVII CMS was written from scratch in PHP, with Postgresql as the relational data base.

#### 5 Changes to Dictionary Editor Workflow

The present form of the dictionary uses more advanced techniques both in its development and in resolving search queries. The open formula permits constant improvements to the lexicographic description and removes the need for printing errata and supplements, while online presentation removes limitations on illustrative material.

A very important feature for a dictionary editor is the ability to alter existing entries. In particular, this raises a possibility of making changes to the methodological principles of the lexicographic description (where applicable), while preserving the macrostructural integrity of the dictionary.

A good example here are the entries for numerals. Originally, they were separated into several grammatical categories on a largely semantic basis, as is the traditional Polish practice. However, electronic resources and tools, as well as newer print works, are abandoning this classification in favour of one based on the grammatical properties of numerals. Therefore, a decision was recently made to introduce changes to categories assigned to numerals in the dictionary: they are now a single part of speech with additional descriptors listing inflectional categories typical to the specific word (e.g. the numeral *dwa* "two" inflects for case and gender, while *drugi* "second" inflects for case, gender, and number). These changes have affected over twenty extant entries.

#### 6 User Amenities

The open form of the dictionary allows for the publication of material that is still undergoing development. The users also gain new features they can use while searching for entries or their elements. At the same time, they must accept the possibility that the dictionary editors will introduce changes to finished entries or the dictionary structure as a whole.

##### 6.1 Access to Unfinished Entries

The dictionary development team distinguishes three development stages of an entry: "stub" (only contains basic information about the grammar and perhaps a single quotation); "entry in preparation," and "entry fully developed" (approved by editor in chief). The users can access an entry at any point of its development. The entry view displays information about its status and date of the most recent modification.

##### 6.2 Active Elements of the Entry

Some elements of an entry function as hyperlinks to other entries, allowing the user to easily find dictionary information supplementing what is shown in the entry they are viewing. For example, some definitions include words that are currently archaic and no longer understood, and so they feature links to definitions of those words (e.g. the definition of the word *arkabuźnik* "heavy cavalryman armed with an → arquebus" includes a link to the entry for *arkabuz* "arquebus").

##### 6.3 User-friendly Entry View

An entry appears in a truncated version by default, only displaying the entry name, abbreviated grammatical notes and phonetic variants (if any). Further sections of the entry, such as "Grammatical forms," "Etymology," or "Meanings,"

<sup>6</sup> The oldest frameworks of that type to remain popular to this day, such as Django or Ruby on Rails, were not created until 2005.



appear as drop-down boxes (see Fig. 1). The user may choose to view any of them in more detail (see Fig. 2). By hovering the cursor over an abbreviation of a given source, the user may see its full citation. It is also possible to display an entry in a printable version that resembles a traditional dictionary entry.

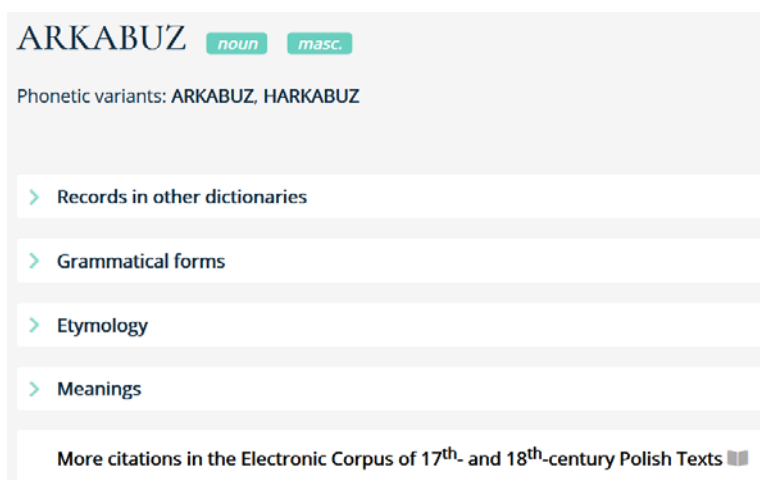


Figure 1: Truncated form of an entry.



Figure 2: An entry with unrolled drop-downs for “Grammatical forms” and “Etymology”.

## 6.4 Search Features

The dictionary search engine enables the viewer to find not only the entry, but also all of its phonetic variants, placed on the list of results alongside the entries themselves. Thus, it will return the entry for *arkabuz* even if the *harkabuz* “harquebus” form is typed into the search box. Subentries are searched alongside the main entries. The search engine also provides more advanced queries, and therefore various ways of filtering the entries. For example, a user can search for entries within a particular part of speech, with a given etymology, or a specific qualifier. It is also possible to search for entries with specific grammatical forms (e.g. dual number forms, no longer present in modern Polish).

## 6.5 Queries

The “Queries” function provides additional ways of searching the entries. It enables the user to access quotations of interest to them without having to open an entry. In this way, one may search for quotes from a specific source or ones that illustrate specific word combinations, set phrases, proverbs, or metaphors. A link underneath a quotation allows the user to quickly access the entry featuring it.



## 7 Integration with Other Resources

Several other electronic resources for 17<sup>th</sup>- and 18<sup>th</sup>-century texts began their development concurrently with e-SXVII. Originally, they functioned independently from the e-SXVII computer system; however, integrating those resources eventually became the natural next step in the development of the dictionary.

The most important resource created with the aim (although not exclusive) to improve the work of e-SXVII lexicographers is the *Electronic Corpus of 17<sup>th</sup>- and 18<sup>th</sup>-century Polish Texts* (hereafter KorBa).<sup>7</sup> The corpus, available online, currently includes 13.5 million tokens (and is expected to reach 25 million by 2023). The texts contained within it, presented as transliterations and transcriptions, have been annotated structurally (marked up for document structure), linguistically (marked up for foreign fragments), and morphosyntactically (lemmatised, marked up for parts of speech, and denoted for values of appropriate grammatical categories).<sup>8</sup> Each text was also provided with extensive metadata. The corpus is searched through the MTAS search engine (Brouwer, Brugman & Kemps-Snijders 2017), which utilises the CQL query language.

The creation of the corpus was a major aid for the dictionary editors. Firstly, it greatly increased the breadth of sources for dictionary quotations, as KorBa also includes material that was not used in the paper card index. The comparison of the frequency lists of lemmas present in the corpus to the KXVII index showed that KorBa contains many words absent from the dictionary card index. The corpus therefore allows for a large expansion of the dictionary's entry network. Secondly, the dictionary editors have gained easy access to sources – by using the corpus search engine, they can find forms of any lexeme or stable phrase that interests them and then copy an appropriate quotation to the dictionary form.

The ongoing developmental work on the dictionary and corpus is meant to integrate these resources, both from the point of view of the users and of the dictionary editors.

The first stage of integration has granted the users the ability to automatically issue queries about instances of a given token in the corpus. The links to the corpus search engine are available next to each inflectional form of the lexeme covered in a given entry, as well as below every entry article (in which case the search targets all inflectional forms of the lexeme).<sup>9</sup>

This solution, intended for the users of the dictionary, has also found use in the work of the dictionary editors, for whom it facilitates searching for appropriate quotations for dictionary entries. However, it was clear since the outset that functions that would connect KorBa to the dictionary editor's interface directly are needed. The first of these is the possibility of automatically downloading inflectional forms found in KorBa by clicking a button in the corpus editor platform. This selects all inflectional forms of the selected lexeme and displays them as hints on the platform. Grammatical markers used in the corpus are automatically translated into grammatical terms used in the dictionary. Each form proposed by the system must be approved by a dictionary editor before it is entered into the paradigm presented in the entry. Work is currently underway on automatic sampling of citations from KorBa marked with a source abbreviation and page number for placing in the appropriate section of the dictionary platform.

Further features are also planned to better utilise another resource gathering 17<sup>th</sup>- and 18<sup>th</sup>-century source texts: the digitalised card index of the dictionary. Experiments investigating the possibility of using djview4poliqarp (Bień 2016) for indexing the electronic version of KXVII are ongoing. Each card record will be assigned a label (in practice equivalent to the entry title), enabling editors to automatically find any and all cards to be used for the development of an entry (for further information, see Bilińska & Rodek 2020).

The future integration of e-SXVII and other linguistic resources for 17<sup>th</sup>- and 18<sup>th</sup>-century Polish will also cover the *Digital Library of Polish and Poland-related Ephemeral Prints from the 16<sup>th</sup>, 17<sup>th</sup> and 18<sup>th</sup> Centuries* (Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku, CBDU). In this case, the dictionary will be used as a source of information about a given historical lexicon – the meaning of the old lexemes taken from dictionary entries will be displayed to the readers of the old texts, helping them understand the content (for more on this subject, see Ogrodniczuk & Gruszczyński 2019).

## 8 Conclusion

The case of e-SXVII serves as an example of one path of further development for dictionaries originating in the second half of the 20<sup>th</sup> century (cf. e.g. Johansson & Battista 2014). Initially developed with traditional methods (based on paper card indices, intended for publication in several volumes over the course of years), they begin to be developed and shared through new digital solutions. The first and most natural change is the creation of a dictionary editor's platform and enabling access to entry articles online. The next stage of development stresses the creation of a convenient way of accessing sources (digital libraries of digitized source texts and dictionary archives as well as electronic text corpora). Later on, these electronic resources are subject to integration.

<sup>7</sup> The word *KorBa*, itself an abbreviation of the Polish phrase *korpus barokowy* 'Baroque corpus', is an alternative name for the corpus. In its original form, it included texts from a period mostly dominated by the Baroque style in Polish literature (17<sup>th</sup> century and 18<sup>th</sup> century up to 1772). Currently, the corpus is expanded to cover the entirety of the 18<sup>th</sup> century, thus also including Enlightenment texts. However, the *KorBa* name has become recognizable in the Polish lexicographic circles to the point where the decision was made to retain it.

<sup>8</sup> The structural and linguistic annotation was carried out by a group of transcribers, while morphosyntactic annotation was completed automatically, through a range of utilities adapted for analysis of old texts. For more on the subject, as well as on the creation of the corpus and its current form, see Bronikowska et al. 2016.

<sup>9</sup> These solutions have been presented in more detail during the 6<sup>th</sup> eLex Conference, held in 2019 in Sintra, Portugal, as part of a lecture titled *Integration of the Electronic Dictionary of the 17<sup>th</sup>-18<sup>th</sup>-c. Polish and the Electronic Corpus of the 17<sup>th</sup>- and 18<sup>th</sup>-c. Polish Texts* by R. Bronikowska, Z. Gawłowicz, M. Ogrodniczuk, A. Wiczorek, and M. Żółtak (relevant paper currently in development).



The new technical possibilities have a larger-than-anticipated influence on the bases of these lexicographic analyses. Far-reaching changes to both the working environment and the ultimate visions for the shape of the dictionary resulting from the specifics of its electronic form allow us to conclude that we are looking at a new lexicographic work, independent from the original print dictionary, although drawing readily on its materials and concept.

## 9 References

- Bień, J.S. (2016). Elektroniczne indeksy fiszek słownikowych. In *Kwartalnik Językoznawczy*, 2 (publ. 2018), pp. 16-27. Accessed at: [http://pmichal-kwartjez.home.amu.edu.pl/teksty/teksty2016\\_2\\_26/Bien.pdf](http://pmichal-kwartjez.home.amu.edu.pl/teksty/teksty2016_2_26/Bien.pdf) [16/03/2020].
- Bilińska-Brynk, J. & Rodek, E. (2020). Paper Quotation Slips to the Electronic Dictionary of the 17<sup>th</sup> and 18<sup>th</sup> Century Polish - Digital Index and its Integration with the Dictionary. In Gavriilidou, Z., Mitsiaki, M., Fliatouras, A. (eds.) *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion*, Vol. I, Democritus University of Thrace, pp. 465-470.
- Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M. & Woliński, M. (2016). The use of electronic historical dictionary data in corpus design. In *Studies in Polish Linguistics*, 11(2), pp. 47-56. Accessed at: <https://doi.org/10.4467/23005920SPL.16.003.4818> [23/01/2020].
- Brouwer, M., Brugman, H. & Kemps-Snijders M. (2017). MTAS: A Solr/Lucene based multi tier annotation search solution. In L. Borin (ed.) *Selected papers from the CLARIN Annual Conference 2016 (Aix-en-Provence, 26–28 October 2016)*. Linköping Electronic Conference Proceedings 136, pp. 19–37. Accessed at: <http://www.ep.liu.se/ecp/136/002/ecp17136002.pdf> [27/10/2019].
- Gruszczyński W. (2005). O przyszłości „Słownika języka polskiego XVII i 1. połowy XVIII wieku”. In *Poradnik Językowy*, 7, pp. 48–61.
- Johannsson, E. & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 169-179. Accessed at: [https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202014/euralex\\_2014\\_010\\_p\\_169.pdf](https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202014/euralex_2014_010_p_169.pdf) [02/03/2020].
- Majdak, M. (2012). Słownik języka polskiego XVII i 1. połowy XVIII wieku, Kraków 1996- IJP PAN. In *Poradnik Językowy*, 8, pp. 105-111. Also in M. Bańko, M. Majdak & M. Czeszewski (eds.) *Słowniki dawne i współczesne. Internetowy przewodnik edukacyjny*. Accessed at: <http://leksykografia.uw.edu.pl/slowniki/21/slownik-jezyka-polskiego-xvii-i-1-polowy-xviii-wieku-krakow-1996> [21/01/2020].
- Majdak, M. (2018). Elektroniczny słownik języka polskiego XVII i XVIII wieku IJP PAN. In M. Pastuch & M. Siuciak (eds.) *Historia języka w XXI wieku. Stan i perspektywy*. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 176-182.
- Ogrodniczuk M. & Gruszczyński W. (2019). Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus. In A. Jatowt, A. Maeda, S. Syn (eds.) *Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science*, vol. 11853, Springer, Cham, pp. 125-138. Accessed at: [https://doi.org/10.1007/978-3-030-34058-2\\_13](https://doi.org/10.1007/978-3-030-34058-2_13) [02/12/2019].
- Siekierska, K. (1998). Słownik języka polskiego XVII i 1. połowy XVIII wieku, historia przedsięwzięcia, założenia teoretyczne, plan prac, prognozy na przyszłość. In *Język Polski*, 1-2, pp. 82-90.

## 10 Language Resource References

- CBDU - Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku/Digital Library of Polish and Poland-related Ephemeral Prints from the 16<sup>th</sup>, 17<sup>th</sup> and 18<sup>th</sup> Centuries. Accessed at: <https://cbdu.ijp.pan.pl/> [15/10/2019].
- e-SXVII - Gruszczyński, W. (ed.). Elektroniczny słownik języka polskiego XVII i XVIII wieku/Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish. Accessed at: <https://sxvii.pl/> [19/03/2020].
- KorBa - Elektroniczny Korpus Tekstów Polskich z XVII i XVIII wieku (do 1772 roku)/Electronic Corpus of 17<sup>th</sup>- and 18<sup>th</sup>-century Polish Text (up to 1772). Accessed at: <http://www.korba.edu.pl> [19/03/2020].
- KXVII - Kartoteka Słownika języka polskiego XVII i 1. połowy XVIII wieku/Card-index of the Dictionary of the Polish Language of the 17<sup>th</sup> and First Half of the 18<sup>th</sup> Century. Accessed at: <https://www.rcin.org.pl/dlibra/publication/20029> [29/03/2020].
- SJPDor - Doroszewski, W. (ed.) (1950–1969). *Słownik języka polskiego*. Vol. 1-11. Accessed at: <https://sjp.pwn.pl/doroszewski> [1/04/2020].
- SStp - Urbańczyk, S. (ed.) (1953-2002). *Słownik staropolski*. Vol. 1-11, IJP PAN, Kraków. Accessed at: <https://pjs.ijp.pan.pl/ssstp.html> [1/04/2020].
- SXVI - Mayenowa, M. R. & Peplowski, F. (vol. 1–34), Mrowcewicz, K. & Potoniec, P. (vol. 35–37), Wilczewska, K., Woronczakowa L. et al. (vol. 27–37) (eds.). *Słownik polszczyzny XVI wieku*. Wrocław: Ossolineum, 1966-1994, Warszawa: IBL PAN, 1995-. Accessed at: <http://spxvi.edu.pl> [1/04/2020].
- SXVII - Siekierska, K. (ed.) (1999-2004). *Słownik języka polskiego XVII i 1. połowy XVIII wieku*, Vol. 1, Fasc. 1-5, Kraków.



**Acknowledgements**

Article prepared within the project “The extending of the Electronic Corpus of the 17<sup>th</sup>- and 18<sup>th</sup>- Century Polish Texts and its integration with the Electronic Dictionary of the 17<sup>th</sup>- and 18<sup>th</sup>-century Polish”, funded by the National Programme for the Development of the Humanities (NPRH) for years 2019-2023 (0413/NPRH7/H11/86/2018).





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Posters**

**Reports on Lexicographical and  
Lexicological Projects**







# The Development of the Open Dictionary of Contemporary Serbian Language Using Crowdsourcing Techniques

Lazić Konjik I., Milenković A.

*Institute for the Serbian Language of SASA, Serbia*

## Abstract

The paper introduces the idea of developing an Open Dictionary of Contemporary Serbian Language using the crowdsourcing method with a clear vision of the necessity of modernizing the work in the field of Serbian lexicography. The crowdsourcing procedure is expected to obtain the latest, present-day vocabulary, i.e. new words which have not yet been presented and described in existing dictionaries of the Serbian language. The crowdsourcing task requires participants to suggest new words, along with the description of their meaning and examples which they find most appropriate. There are two primary goals: (1) solving the problem of timely and comprehensively recording and presenting new vocabulary in real time and (2) creating a user-oriented dictionary of new words and expressions. The main idea is to create a dictionary which would offer simple explanations of the semantic, grammatical and pragmatic features of new words and expressions in the Serbian language. The practical, lexicographic objective is to show the need for the development of such a dictionary which would fill the gap that exists in Serbian descriptive lexicography and neography.

**Keywords:** e-lexicography, dictionary, neography, crowdsourcing, contemporary Serbian language

## 1 Introduction

In this paper we will briefly outline the idea of a new online dictionary – Open Dictionary of Contemporary Serbian Language (ODCS). The ODCS would represent a digital lexical database where new words and expressions, which have appeared in the Serbian language in the last ten years, but were not lexicographically processed, would be systematically collected, using a modern lexicographic methodology and explained in a user-accessible way.

We will present the goals which are set before the realization of the planned dictionary and the theoretical and methodological aspects of its development. Accordingly, the paper is structured as follows: in section 2 we present the objectives; section 3 gives a brief overview of the type and purpose of the ODCS; section 4 describes language material collecting procedures; sections 5 and 6 present the macro- and microstructure of the ODCS, and in section 7 we conclude and discuss the future work and perspectives of the ODCS.

## 2 Objectives

The appearance of new words in languages is an unceasing process. The role of the lexicographer is to record and present new words and expressions in a timely manner, both to the general (laic) and professional public. With the development of digital technologies, this long-standing need of both lexicographers and dictionary users has become attainable, moreover, it stands as a challenge before modern dictionaries (Apresjan 2000; Atkins & Rundell 2008; Hanks 2012). In addition to this demand, a priority is given to the requirement that the dictionary meets the needs and expectations of its future users (Atkins & Rundell 2008). This practically means that it is necessary to anticipate the reasons why the user is searching for a particular word and then try to provide a set of appropriate answers (Varantola 2002: 33). This, furthermore, indicates the necessity of defining and starting from a potential target group of users.

Since we are discussing a dictionary of new words, we emphasize that the target group of the ODCS is a wide range of potential users, non-professionals and professionals, who are usually Serbian native speakers (L1). This is why the main motive for planning and compiling dictionaries of new words is about meeting their communicative needs. By professional users we mean all the people who use dictionaries as reference sources in their work (Varantola 2002: 32) which, in the case of the ODCS, are primarily linguists and lexicographers.

Accordingly, the primary goal of the new ODCS would be to offer simple explanations of the semantic, grammatical and pragmatic features of new words and expressions in the Serbian language. Firstly, to help the broadest range of non-professional users understand the meaning of new words and to encourage them to use these words in naming objects, concepts and phenomena which are part of their everyday life. Secondly, the demands of professional users set a goal before the ODCS to become a modern and reliable lexicographic database of new words and expressions, as well as a reference source for various types of linguistic activities and research. And finally, to apply the user-produced data in the domain of the semantic interpretation of terms into further research of the interpretive user perspective of different types of terms and, in this regard, the ways of defining them (cognitivism) (cf. Bartminjski 2011: 93-168). This is part of our future work and we will give some remarks on it in the concluding section.

This way, the ODCS would fill the gap which has been noticeable in Serbian descriptive lexicography in the last ten years (the last dictionary of the Serbian language was published in 2011) as a basic lexical supplement to the existing descriptive dictionaries of the Serbian language (the one-volume DSL and the six-volume MSD). At the same time, it



would become a relevant source and reference tool for the development of future dictionaries. The final result would be a user-friendly, easy-to-use dictionary, where new words, which have recently appeared in the language and have, more or less incorporated into it, are explained in a simple, transparent and linguistically reliable way – so that potential users can easily understand and use them.

### 3 The Type and Purpose of the Dictionary

The proposed dictionary will introduce the present-day vocabulary of contemporary Serbian language covering the last ten years (approximately from the year 2010 until today) which has not been recorded in its descriptive dictionaries: the DSL, MSD and SASAD, nor in the existing dictionaries of foreign or new words: Ćirilov 1982, 1991; Klajn 1992; Otašević 1999, 2008a; Vasić, Prčić & Nejgebauer 2001; Bugarski 2019. These printed dictionaries cover the period up to 2008 (except for the dictionary Bugarski 2019, which, however, brings only lexical blends), so the comparison with them will determine the relevance for their inclusion into the Open Dictionary. Other criteria for the selection of the lemmas are described in section 5. The ODSL will have a descriptive character. There is no intention of it becoming a formal normative source, but certain information of this type will be included by presenting appropriate orthographic and semantic uses of words, in relation to Serbian descriptive lexicography (cf. Ristić 2006: 41-64, 79-92). This is very important because it is assumed that one of the reasons why users search for a certain word is its normative status. The focus will be on introducing words (vocabulary) which are, in conceptual or word-formation terms, a novelty in the Serbian language.

### 4 Collecting Language Material

The current practice in Serbian descriptive lexicography, especially neography which has, in recent years, stood out in many ways as an independent field (cf. Благоева 2016: 200), is grounded upon traditional methods of collecting words – manual excerption from newspapers and magazines, literary work, radio and television and colloquial language (see the above-mentioned Serbian dictionaries of new words). At the same time examples of recent Slavic and European efforts, as well as the papers, which present the state-of-the-art research into neology and ideas about the modern lexicographic treatment of neologisms, show that the procedures of collecting and excerpting new words for lexicographic presentation have been greatly modernized and improved by introducing the corpus method and digital performance (Sinclair 1991; Meyer 2004; Biber & Reppen 2015). Nevertheless, corpus-based lexicography is still, to a great extent, combined with the traditional manual method (e.g. Благоева 2011: 18; Karlsson 2000), but more frequently uses language tools which enable a semi-automatic or automatic selection of new words (Krek, Kosem & Gantar 2013; Klosa-Kückelhaus & Ilan Kernerman 2020). Concerning the data collection process, we can notice the emergence of a new form, the so-called user-generated dictionaries where word collecting is partially or fully performed by the crowdsourcing method (e.g. the Slovene *Sprotni slovar*,<sup>1</sup> the English *Macmillan Open Dictionary*,<sup>2</sup> the Swedish *Folkmun*<sup>3</sup>).

Recent approaches in lexicography show that the crowdsourcing method has been recognized as a suitable method for less complex lexicographic tasks (Čibej, Fiser & Kosem 2015: 71), so collecting new words is, based on previous experience, one of the tasks which can be performed successfully with this method (Rundell 2012; Sköldberg & Wenner 2020). Research has shown that successful crowdsourcing projects can attract a great number of volunteers if the public is well-informed and asked to help, especially when such projects are supported by non-profit organizations with the goal of “general welfare” (Holley 2010). European practice and some successful domestic projects show that there are many people outside formal institutions who have excellent skills and are willing to invest considerable time and effort into linguistic resources which are of general interest, such as free online dictionaries and encyclopedias, e.g. Dict.cc,<sup>4</sup> Wiktionary,<sup>5</sup> Lingobee,<sup>6</sup> Wikipedia,<sup>7</sup> or the language learning platform, e.g. Duolingo,<sup>8</sup> Memrise,<sup>9</sup> or domestic projects, e.g. Jezička laboratorija,<sup>10</sup> Prepis.org,<sup>11</sup> etc. (cf. Čibej, Fišer & Kosem 2015: 71; Fišer & Čibej 2017: 214-215).

Crowdsourcing is planned on being used in the development of the new ODCS as the basic way of obtaining material,<sup>12</sup> since there is no available current and up-to-date corpus of Serbian language (the existing corpora SrpKor<sup>13</sup> and SrWaC<sup>14</sup> cover the period up to 2013 and 2014) or appropriate language tools which would enable (semi-) automatic excerption. Such tools are being developed in the Serbian language community, but for now mostly in the field of terminology (e.g. Krstev et al. 2015; Stanković et al. 2016; Šandrih et al. 2020). The general purpose of crowdsourcing is to accelerate the

<sup>1</sup> Accessed at: <https://fran.si/> [07/07/2020]

<sup>2</sup> Accessed at: <https://www.macmillandictionary.com/open-dictionary/latestEntries.html> [07/07/2020]

<sup>3</sup> Accessed at: <https://folkmun.se/> [07/07/2020]

<sup>4</sup> Accessed at: <http://www.dict.cc/> [07/07/2020]

<sup>5</sup> Accessed at: <https://www.wiktionary.org/> [07/07/2020]

<sup>6</sup> Accessed at: <http://itrg.brighton.ac.uk/simola.org/lingobee/> [07/07/2020]

<sup>7</sup> Accessed at: <https://www.wikipedia.org/> [07/07/2020]

<sup>8</sup> Accessed at: <https://en.duolingo.com/> [07/07/2020]

<sup>9</sup> Accessed at: <https://www.memrise.com/> [07/07/2020]

<sup>10</sup> Accessed at: <http://lab.unilib.rs/> [07/07/2020]

<sup>11</sup> Accessed at: <http://www.prepis.org/> [07/07/2020]

<sup>12</sup> On the importance and advantages of electronic corpora as a source of information in lexicography in relation to traditional lexical files see Fillmore 1992; Atkins & Rundell 2008: 53.

<sup>13</sup> Accessed at: <http://www.korpus.matf.bg.ac.rs/> [07/07/2020]

<sup>14</sup> Accessed at: [https://www.clarin.si/noske/all.cgi/first\\_form?corpname=srwac:align](https://www.clarin.si/noske/all.cgi/first_form?corpname=srwac:align) [07/07/2020]



initial (and continuous) process of developing a database (corpus) which will serve as the starting point in forming a lexical vocabulary database, after the process of lexicographic evaluation. The crowdsourcing procedure is expected to obtain the latest, present-day vocabulary such as, for example, the lexicon which has emerged as a result of the outbreak of the coronavirus pandemic.

Crowdsourcing techniques will be combined with the traditional method of manual excerption of targeted sources such as scientific papers, MA theses and PhD dissertations, in which neology topics have been discussed in the last ten years, because of the great amount of new words and lexical material obtained in them.<sup>15</sup> We assume that, this way, a large amount of vocabulary which has remained in the gap between theory and practice, due to the discrepancies between lexicology and lexicography, will successfully be obtained.

Every Serbian native speaker will be able to participate in the development of the planned dictionary. The workflow basically follows the scheme presented in the paper Čibej, Fiser & Kosem (2015: 75-77). An appropriate user interface will be made with an integrated specific proposal form, thanks to which the language community will be involved in resource development. The crowdsourcing task requires participants to suggest new words, as a potential lemma candidate, along with the description of meaning. Also, if they wish, they can provide other information which they consider relevant to the description of the word they are proposing, such as e.g. the circumstances or the situation in which the given word is used, its synonyms, etc. Additionally, votes for other users' proposals whether a word is or isn't new will be enabled. Further verification (validation) of the entries made by the participants in the crowdsourcing task, the selection and lexicographic processing, in accordance with the theoretical concept of the dictionary, will be performed exclusively by lexicographers (see section 5). All the user provided data will be stored in a special database in the form of a list of tokens, available for online viewing and downloading. The dictionary base (i.e. the processed dictionary entries) will be kept separately.

## 5 The Macrostructure

New words and expressions in Serbian and other languages, include lexical neologisms (novel and recently borrowed words naming new realities and concepts from different fields), semantic neologisms, word-formation neologisms, neoarchaisms, recently introduced compounds of terminological character and new phraseology (Otašević 2008: 39-40; Ristić 2012: 9; Dragičević 2018: 237-247; Благоева 2011: 18). New words represent mainly the peripheral lexicon, neologisms, occasionalisms and potential lexis (individualisms), which have a variable status in the dynamics of lexical development. The characteristics of neologisms is that they undertake different phases of usualization, which begins with the phase of occasionalisms. The relation of neologisms to time is direct, while the chronology criterion for occasionalisms and individualisms is relative (Otašević & Sikimić 1991: 77-78; Dragičević 2018: 228-237). From the aspect of neography, the criterion of chronological status is superior, and the question of their status in the lexicon is mainly a question of the test of time. Accordingly, in digital neography, the question of selection, in terms of distinguishing a neologism in the narrower sense vs. a neologism in the broader sense of the term becomes secondary, when it includes individual vocabulary (individualisms, hapaxes, occasionalisms).

As opposed to the paper (printed) editions, the electronic dictionaries impose no restrictions on the extent and amount of language material that can enter the dictionary, which allows for the gathering of different layers of new lexical phenomena, including: jargon, speech lexis, newly borrowed words, acronyms, loan translations (calcs) and other types of new vocabulary (Котелова 1978: 25-26; Ristić 2012: 1, f.10; Bugarski 2006: 42 & 2019; Dragičević 2018: 222-247). This is a major advantage knowing that languages are constantly changing, evolving, new words and expressions appearing, old ones disappearing, the vocabulary is constantly regrouping, words moving from the peripheral registers into the general vocabulary funds and vice versa, which is a common and integral part of its development dynamics (Ristić 2012: 11). The recording of all new words, even one-day words which quickly disappear from the language, is significant not so much from the point of view of their acquisition and knowledge, but from the point of studying the principles, tendencies and perspectives in the development of the vocabulary of a language, from the synchronic and diachronic aspect. New technologies create a great opportunity for all lexicographers because they enable an exhaustive presentation of the most recent lexical changes and appearances which take place in one language's development and growth.

The main criteria for the selection of entries will be both chronological and lexicographic (cf. Otašević 2008: 19) – that is, all the words which have not yet been announced and recorded in corresponding Serbian dictionaries from the year 2010 (see section 3). Other, special ways of detecting neologisms, essentially, do not exist, but there are various ways of checking whether an unrecorded word is a lemma candidate in general (cf. Trap-Jensen 2020). The basic way of checking is the frequency factor. Namely, the high frequency of a word in the corpus indicates the institutionalization of the lexical unit, but the speakers' perception about the novelty of these words will also be taken into account (cf. Freixa & Torner 2020), thanks to the user participation in dictionary development. Therefore, for each user suggestion, it will be checked whether there is evidence that the word is used (based on reference sources, in the existing available corpora of the Serbian language, and given that it is a current vocabulary that dictionaries and corpora, as a rule, do not yet record, the

<sup>15</sup> Selected PhD and MA theses defended at the University of Novi Sad: The Faculty of Philosophy, Serbia: Rakić, K. (2016). Lexicological and lexicographical position of the current loanwords in the Serbian language; Savić, M. (2013). Semantico-Pragmatic Pseudoanglicisms in Serbian; Aćimović, S. (2013). Anglicisms in the Novels of Modern Serbian Women Writers of the Younger Generation; Rodić, S. (2013). Facebook communication on reality shows: the lexical influence of English on Serbian language; Jerković, V. (2012). Understanding Anglicisms in specialized teenage magazines in Serbian; Janjatović, V. (2010). The use of Anglicisms in Women's Fashion Magazines in Serbian, etc.



check will also be done online). If there is evidence showing that the word is in use, it will be selected and added to the dictionary after lexicographic processing. This way the corpus will exclude misspellings, tendentiously invented words, vulgar and obscene words, and other similar words which will still be stored in the primary database containing all user-produced suggestions. Schicchi & Pilato (2018) describe a semi-automatic method for creating neologisms and show that the number of tendentiously created words for commercial purposes in order to produce a creative name which will sell a product or service, can make a significant share in the vocabulary as a whole. The so-called vulgar and obscene lexicon, which is traditionally omitted from Serbian dictionaries, poses a special problem. While contemporary understandings indicate the need for an objective representation of all lexical units of a language, the real situation shows that the parameters for identifying such vocabulary have not yet been developed, neither in terms of content nor in terms of use (Ristić 2006: 91).

The advantages of digital technology provide the possibility of making rapid and permanent changes in the composition of dictionary entries (supplementation, rearrangement, etc.). Also, the lemmas will be presented in a concise or an extended form. In that way, the entry will be complemented with new contents and information which have not been customary for printed dictionaries of new words, and are, also, in line with the requirements of monolingual descriptive lexicography: the pronunciation of words, full forms of declension and conjugation, pragmatic information, thesaurus functions, established collocations, multimedia contents.

The dictionary will be incrementally updated and the entries will appear in the dictionary immediately after lexicographic processing, so we can say that the dictionary will be constantly complemented and created before the eyes of the user, reflecting the living nature of language and its continuous development. The database will be permanently available for download in at least one of the standard text formats (XML, CSV, JSON, YAML). A cross-section is planned yearly (once a year) in the form of individual dictionary versions.

The accompanying, not less important parts of the dictionary will introduce additional search criteria which is intended to meet the specific needs of users and different levels of their competence (Varantola 2002: 39). Different choices of displaying and searching will be offered according to: the standard alphabetical order, different functional-stylistic and semantic registers, etymology (foreign word / domestic word), the criterion new word / new meaning of an existing word. Labels for semantic fields in each entry can be used as a navigational tool to display a list of all entries from the given field, enabling thematic browsing through the collection (see Tasovac & Petrović 2015: 392). The thematic (semantic) classification of vocabulary in Serbian studies is mainly considered in connection with the dialectal and traditional vocabulary in dialectology, ethnolinguistics and dialectal lexicography (Miloradović 2012: 146; Tasovac & Petrović 2015; Lazić Konjik 2017), from which we will start, continue and develop in relation to the material. The labels for the semantic fields will also show which semantic domains are current in the contemporary linguistic picture of the world. In the case of a new meaning of an existing word the lemma will have in its exponent a symbol (mark) which will indicate that it's about a new meaning of an existing word, and not a new word.

In this way, the multi-layered content of the dictionary will be closer to potential users and will provide easy and quick access to the information they need.

## 6 The Microstructure

The basic characteristics of the microstructure provide that the design is entirely at the service of the function to be fulfilled by the dictionary in the presented context (Kiefer & Sterkenburg 2003: 351). The lexicographical processing is completely under the jurisdiction of the lexicographer.

- basic lemma (the most common, as well as all the acceptable spelling forms);
- the stressed form and alternatively, all acceptable pronunciation forms in use;
- the grammatical category of: word type, gender for nouns, verb type, gender forms for adjectives;
- etymology (original form of a foreign word);
- word-formation structure;
- inflectional grammatical forms (declension and conjugation);
- notes on pragmatic and stylistic values;
- definition (presenting the explanation of the new word's meaning, given in full sentences, with high frequency common words which are easy to understand);
- information on lexical and semantic compatibility, collocations in use;
- examples (whole sentences, as many examples as possible);<sup>16</sup>
- paradigmatic lexical relations (synonyms, antonyms, opposites);
- phraseology (all phraseologisms will be united in a separate dictionary part, connected to their primary lemmas);
- multimedia content.

## 7 Conclusion and Perspectives

In this paper we have introduced the idea of developing an Open Dictionary of Contemporary Serbian using the innovative crowdsourcing method, with a clear vision of the necessity of modernizing the work in the field of Serbian lexicography. In addition to the questions of the dictionary concept, the methodology of making and organizing the content on the micro and macro plan, which are mostly considered, the development of a new dictionary always opens numerous theoretical questions, especially in the domain of the presentation of semantic content and interpretation of the

<sup>16</sup> The usage degree of the entry-word will be indirectly indicated by the number of examples and collocations cited.



ways in which words function in language and texts. This leads to current cognitive theoretical approaches in which the problem of the boundary between the so-called linguistic and non-linguistic knowledge and ways of defining different types of concepts is discussed, which will also be the subject of our future work. Cognitivism, as pointed out by J. Bartminski, requires respect for the interpretive perspective adequate to the competence of the user of the language that is the subject of analysis (Bartminski 2011: 97). We expect that a significant contribution to research in this area will be given by the data obtained from our users.

Apart from the fact that such a dictionary would be an important source of data for the understanding of novel lexical meanings and word usage, it would also serve as an important language infrastructure and source for linguistic studies of contemporary Serbian, for monitoring its development, language standardization and linguistic planning. All of this together stands as a prerequisite for the modernization of scientific practice in the field of language science and other humanities in the modern IT society.

With the help and financial support of the competent institutions, the ODCS would be available, free of charge, to all interested users, researchers and all who wish to develop applications, programs and tools for the Serbian language based on this dictionary as a source of lexical information. A well-structured data set, as will be the outcome of this project, can encourage further scientific analysis in the domain of language technologies, both within the dictionary itself or linking up with other structured data collections, such as e.g. (morphological) e-dictionaries, electronic corpora or other future digitized dictionaries of the Serbian language.

## 8 References

- Apresjan, J. (2000). *Systematic Lexicography*. Oxford: Oxford University Press.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bartminski, J. (2011). *Jezik – slika – svet. Etnolingvističke studije*. D. Ajdačić (ed.). Beograd: SlovoSlavia.
- Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Bugarski, R. (2006). *Žargon – lingvistička studija*. Drugo, prošireno izdanje. Beograd: Biblioteka XX vek.
- Bugarski, R. (2019). *Srpske slivenice, monografija sa rečnikom*. Novi Sad: Akademska knjiga.
- Čibej, J., Fišer, D. & Kosem, I. (2015). *The role of crowdsourcing in lexicography*. Accessed at: [https://elex.link/elex2015/proceedings/eLex\\_2015\\_05\\_Cibej+Fiser+Kosem.pdf](https://elex.link/elex2015/proceedings/eLex_2015_05_Cibej+Fiser+Kosem.pdf) [05/15/2020].
- Ćirilov, J. (1982). *Rečnik novih reči*. Beograd: Narodna knjiga.
- Ćirilov, J. (1991). *Novi rečnik novih reči*. Beograd: Bata.
- Dragičević, R. (2018). *Srpska leksika u prošlosti i danas*. Novi Sad: Matica srpska.
- DSL (2007). *Rečnik srpskoga jezika*, Novi Sad: Matica srpska.
- Fillmore, Ch. (1992). Corpus Linguistics vs. Computer Aided Armchair Linguistics. In J. Svartvik (ed.). *Directions in Corpus Linguistics*. Berlin and New York, pp. 35-60.
- Fišer, D., Čibej, J. (2017). The potential of crowdsourcing in modern lexicography. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Dictionary of Modern Slovene: Problems and Solutions*. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 212-229.
- Freixa, J. & Torner, S. (2020). Beyond Frequency: On the Dictionarization of New Words in Spanish. In *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), pp. 131-153.
- Hanks, P. (2012). The corpus revolution in Lexicography. In *International Journal of Lexicography*, 25/4, pp. 398-436.
- Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? In *D-Lib Magazine*, 16 (3/4). Accessed at: <http://www.dlib.org/dlib/march10/holley/03holley.html> [05/15/2020].
- Jovanović, J. (2019). Nova upotreba prideva human u savremenom srpskom jeziku. In *Novorečje*, 1, Beograd: Alma, pp. 5-8. Accessed at: <http://alma.rs/Novorecje/Novorecje1.pdf> [05/15/2020].
- Karlsson, F. (2000). Lexikografie a korpusova lingvistika. In F. Čermák, J. Klimov, V. Petkevič (eds.) *Studie z korpusové lingvistiky*. Praha: Karolinum, pp. 427-454.
- Kiefer, F., Sterkenburg, P. (2003). Design and production of monolingual dictionaries. In P. Sterkenburg (ed.) *A Practical guide to lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing, pp. 350-365.
- Klajn, I. (1992). *Rečnik novih reči*. Novi sad: Matica srpska.
- Klosa-Kückelhaus, Ilan Kernerman (2020) (ed.). Global Viewpoints on Lexicography and Neologisms. In *Dictionaries: Journal of the Dictionary Society of North America*, 41(1). Accessed at: [https://muse.jhu.edu/issue/42292#info\\_wrap](https://muse.jhu.edu/issue/42292#info_wrap) [07/07/2020].
- Krek, S., Kosem, I. & Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika* (A Proposal for a new Dictionary of Contemporary Slovene). Version 1.1. Accessed at: [http://trojina.org/slovar-predlog/datoteke/Predlog\\_SSSJ\\_v1.1.pdf](http://trojina.org/slovar-predlog/datoteke/Predlog_SSSJ_v1.1.pdf) [05/15/2020].
- Krsteš, C., Stanković, R., Obradović, I. & Lazić, B. (2015). Terminology Acquisition and Description Using Lexical Resources and Local Grammars. In T. Poibeau & P. Faber (eds.) *Proceedings of the 11th Conference on Terminology and Artificial Intelligence, Granada, Spain, 2015*. LexiCon (Universidad de Granada), pp. 81-89.
- Lazić Konjik I. (2017). Leksika tradicionalne kulture prema tematskim poljima. In P. Piper & V. Jovanović (eds.) *Slovenska terminologija danas*, Beograd: SANU, pp. 613-623.
- Meyer, Ch. F. (2004). *English Corpus Linguistics. An Introduction*. Cambridge: Cambridge University Press.
- Miloradović, S. (2012). Lingvistički atlas – „centralni instrument“ savremene dijalektologije. In *Zbornik radova*



- etnografskog instituta SANU: *Terenska istraživanja – poetika susreta*, 27, pp. 141-151.
- MSD (1967–1976). *Rečnik srpskohrvatskoga književnog jezika*, I–VI, Novi Sad (– Zagreb): Matica srpska (– Matica hrvatska).
- Otašević, Đ. (1999). *Rečnik novih i nezabeleženih reči*. Beograd: Enigmatski savez Srbije.
- Otašević, Đ. (2008). Nove reči i značenja u savremenom standardnom srpskom jeziku. Beograd: Alma.
- Otašević, Đ. (2008a). *Rečnik novih reči*. Beograd: Alma.
- Otašević, Đ. & Sikimić, B. (1991). Odnos okazionalizama prema vremenu. In *Naš jezik*, XXIX, 1–2, pp. 77-81.
- Ristić, S. (2006). *Raslojenost leksike srpskog jezika i leksička norma*. Beograd: Institut za srpski jezik SANU.
- Ristić, S. (2012). *O rečima u srpskom jeziku*. Beograd: Institut za srpski jezik SANU.
- Rundell, M. (2012). The wisdom of crowds: can it work for dictionaries? Accessed at: <http://www.macmillandictionaryblog.com/the-wisdom-of-crowds-can-it-work-for-dictionaries> [05/15/2020].
- SASAD (1959–2019). Речник српскохрватског књижевног и народног језика, I–XXI, Београд: САНУ.
- Schicchi, D., Pilato, G. (2018). WORDY: A Semi-automatic Methodology Aimed at the Creation of Neologisms Based on a Semantic Network and Blending Devices. In L. Barolli, O. Terzo (eds.) *Complex, Intelligent, and Software Intensive Systems*. CISIS 2017. Advances in Intelligent Systems and Computing, vol. 611. Springer, Cham. Accessed at: [https://link.springer.com/chapter/10.1007/978-3-319-61566-0\\_23](https://link.springer.com/chapter/10.1007/978-3-319-61566-0_23) [05/15/2020].
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sköldbberg, E. & Wenner, L. (2020). Folkmun.se: A Study of a User-Generated Dictionary of Swedish. In *International Journal of Lexicography*, 33(1), pp. 1-16.
- Stanković, R., Krstev, C., Obradović, I. Lazić, B. & Trtovac, A. (2016). Rule-based Automatic Multi-word Term Extraction and Lemmatization. In N. Calzolari et al. (eds.) *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016*, pp. 507-514.
- Šandrih, B., Krstev, C. & Stanković, R. (2020). Two approaches to compilation of bilingual multi-word terminology lists from lexical resources. In *Natural Language Engineering*. Cambridge: Cambridge University Press. DOI: 10.1017/S1351324919000615.
- Tasovac, T., Petrović, S. (2015). Multiple Access Paths for Digital Collections of Lexicographic Paper Slips. In *eLex 2015 – Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age*. Accessed at: <https://ellex.link/ellex2015/conference-proceedings/paper-25/> [05/15/2020].
- Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in The Danish Dictionary. In *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), 11-25. doi:10.1353/dic.2020.0002.
- Varantola, K. (2002). Use and Usability of Dictionaries: Common Sense and Context Sensibility? In M. Corraeard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins, Euralex 2002*, pp. 30-44.
- Vasić, V., Prčić, T. & Nejgebauer, G. (2001). *Rečnik novijih anglicizama*. Novi Sad: Zmaj.
- Благоева, Д. (2011). Принципи за съставяне на словник за неологичен речник. In: *Многообразие в единството*, 1, pp. 17-21.
- Благоева, Д. (2016). Славянската неография – минало и настояще. In: *За словото – нови търсения и подходи: Юбилеен сборник в чест на чл.-кор. проф. д.ф.н. Емилия Пернишка*. Издателство на БАН „Проф. Марин Дринов“, pp. 200-210.
- Котелова З. Н. (1978). Первый опыт лексикографического описания русских неологизмов. In Н. З. Котелова (ed.) *Новые слова и словари новых слов*. Ленинград, pp. 5-26.
- Плотникова А.А. (2009). *Материалы для этнолингвистического изучения балканословянского ареала* (Plotnikova A.A. Materijali za etnolingvističko proučavanje balkansko-slovenskog areala / Prevod M. Ilić.) М: Институт славяноведения РАН. (1st ed. 1996).
- Поликарпов А.А. (2013). Модель жизненного цикла знака: к теоретическим основаниям исторической лексикологии и дериватологии. In: М. И. Чернышева (ed.) *Славянская лексикография. Международная коллективная монография*, М.: Издательский центр «Азбуковник», pp. 679-702.
- Поликарпов, А.А., Кукушина, О. В. & Токтонов, А. Г. (2008). Проверка теоретически предсказанных неогравитологических закономерностей данными русской корпусной неогравитологии. In: М. И. Чернышева (ed.) *Теория и история славянской лексикографии, Научные материалы к XIV съезду славистов*, Москва, pp. 392-426.

## Acknowledgements

This paper has been developed as a result of the financing Contract 451-03-68/2020-14/200174 with the Ministry of Education, Science and Technological Development of the Republic of Serbia. The research is also related to the activities of the Cost Action CA16105 enetCollect *European Network for Combining Language Learning with Crowdsourcing Techniques*, to whom we, by this occasion, express our sincere gratitude.



# THE NEW ONLINE ENGLISH-GEORGIAN MARITIME DICTIONARY PROJECT. CHALLENGES AND PERSPECTIVES.

Tenieshvili A.

*Batumi State Maritime Academy, Georgia*

## Abstract

The New Online English-Georgian Maritime Dictionary project (NEGMD) concerns two main issues: the compilation of dictionary itself and the creation and adoption of Georgian maritime terminology. Optimizing Georgian maritime terminology would greatly add to the development of the whole maritime field of Georgia and it will be especially important for the education of highly qualified seafarers in our country, who will be occupied at sea or on shore-based maritime jobs. The aim of our report on the NEGMD is to present the current status of the project, to show its specifics and the guiding principles during the compilation of the entries of this dictionary and also to present reasons illustrating the urgency of this project.

The creation of an optimal Georgian maritime terminology and the compilation of the NEGMD will increase the motivation of students of maritime education and training (MET) institutions in Georgia to study, will improve the quality of the curriculum at maritime higher education institutions of Georgia and will contribute to adoption of Georgian Maritime terminology in the field of Maritime Education and Training and also in other maritime fields in Georgia.

**Key words:** dictionary compilation, dictionary entry, corpora, term derivation, term definition

## 1. INTRODUCTION

In this paper we will present a progress report on the planned new online English-Georgian Maritime Dictionary project. We will discuss the challenges we will have to face when compiling the NEGMD and we will outline the perspectives the dictionary will offer when compiled. The present report has three purposes: 1) compilation of new maritime dictionary, coinage of Georgian maritime terms and 3) consequent creation of corpus of maritime texts in Georgian. The compilation of the NEGMD is of paramount importance to the development of the maritime semantic field. It will lead to the extension of Georgian maritime terminology, because lexical and terminological gaps in Georgian will be filled. The NEGMD will contribute to strengthening the country's position in the international maritime arena.

The issue of compiling the NEGMD was first raised at the Maritime Administration of Georgia, where we as alumni of the World Maritime University did our internship. We followed up on this by presenting the paper “Why It Is Necessary to Create and Adopt Georgian Maritime Terminology” at the international conference “International Terminology: Translation and Standardization”, held at Tbilisi State University in October 2018. The paper was given project status and in January 2019 we applied for an ELEXIS travel grant and we were invited for a research visit to the Dutch language Institute (INT) based in Leiden, the Netherlands. Under the professional guidance of the INT, specific research questions and the methodological concept of the dictionary project were addressed. Thanks to the help of INT, the project has been given sufficient basis for launching.

## 2. OUTLINE AND METHODOLOGY

### 2.1. THE OUTLINE OF THE PROJECT

The NEGMD project is very important for Georgia as a sovereign country, for Georgian linguistics and lexicography and for the Georgian maritime semantic field. Georgia is a sovereign country now, and terminology is one of the means of establishing the country as an independent state. Although maritime terminology is the most globalized semantic field in the world, each country, especially maritime nations like Georgia, must have their own maritime terminology. It is especially important for reaching the following aims:

- 1) Stimulation of developing Georgian maritime terminology;
- 2) Application of the developed maritime terminology in textbooks, maritime documentation and materials;
- 3) Economy of space and facilitation of understanding among specialists and students of the maritime field;



- 4) Improvement of the system of maritime education and consequently of the national maritime industry;
- 5) Contribution to the lexicography of the Georgian language;
- 6) Development of lexicography of the maritime semantic field;
- 7) Establishing a basis for a Georgian maritime corpus with further integration of the Georgian maritime corpus in the general corpus of the Georgian language;
- 8) General contribution to Georgian linguistics.

With the work on the NEGMD we have three objectives: 1) the dictionary; 2) coinage of Georgian maritime terms; 3) an organized corpus of maritime texts and materials in Georgian. The last goal is not foreseen in the near future, because it is only gradually that new terms will find their way into textbooks for maritime educational institutions, and from there to maritime students.

Prior to the project, we formulated the following research questions, most of which were answered during the research visit to the Dutch Language Institute (INT) in June 2019:

- Compliance of term definitions with the ISO 704 standard;
- Application of Sketch Engine Tools for extracting relevant sample sentences from the English and Georgian corpora;
- Application of QTerm, MultiTems databases to compile terminological databases;
- Familiarization with dictionary writing systems including LEXONOMY to start the dictionary compilation;
- Receiving instructions regarding terms' definitions;
- Receiving instructions regarding entries' compilation.

For the project it is important to invest time in establishing a balanced corpus for this specific domain, as a lot of documents are not available in digital form or are copyright protected. The materials available at the library of the Maritime Academy we are currently working for are also relevant to illustrate the usage of maritime terms in Georgian.

Since the dictionary is a kind of terminological project and as terms and concepts are two intertwined phenomena, we would like to focus our attention on the essence of terminology. In the words of Geeraerts: "Terminologies – the lexical components of specialized languages – emerge from theoretical and technological innovation: new scientific insights and novel tools enrich the conceptual and practical environment of the specialists, and in the process expand their vocabularies". ("Handbook of Terminology", vol. I, Foreword, 2015:14).

In order to compile a dictionary that will comprise various kinds of information, including definitions of terms, corpora examples, encyclopedic information, related words, etc., it was necessary to conduct some research and to seek advice from specialists experienced in similar work. As the dictionary is intended to have both terminological and lexicographic features, the opinions of terminologists and lexicographers are important and need to be taken into consideration. In addition to dictionary compilation, the purpose of our work is to conduct research in order to fill the lexical/terminological gaps in Georgian maritime terminology. Filling lexical/terminological gaps is very important as the NEGMD will not be complete before all necessary Georgian maritime terms have been developed. Although this work can be conducted simultaneously with the process of dictionary compilation and, consequently, new Georgian maritime terms can be developed and included in the dictionary, adoption and assimilation of terms in the language is a matter of time, and certain efforts are to be made by Maritime Education and Training (MET) institutions of Georgia to speed up the process.

## 2.2. THE METHODOLOGY USED

Before we discuss our project, we would like to make a distinction between terminological and lexicographic approaches of dictionary compilation. Kageura (2015) argues that there is a general understanding that terminographical work focuses on concepts, definitions and designations (although many terminographical practices are not necessarily limited to these aspects" (ISO 704, 2009). "Lexicography", however, deals with words or lexical items in general and with a full range of linguistic information related to words, including grammatical features such as POS, meanings, usages, discourse types, register, etc. depending on the type of dictionary ("Handbook of Terminology", vol. I, 2015:96).

The terminological approach involves moving from a concept to a term, whereas the lexicographic approach implies moving from a word to a meaning. The unique characteristic of a term is that a new concept leads to a term. Lexicographers describe how people use the words; terminologists try to establish how words should be used. We could add here that terminology has a more prescriptive character, whereas lexicography is characterized by a more descriptive approach. Ten Hacken (2000:22) says: "On the basis of study of terms, Temmermem (2000) argues that terms are not crucially different from words in the sense that both are based on prototypes. This implies that terminological definitions should be interpreted in the same way as lexicographic definitions".



For the compilation of the NEGMD we have chosen to use a combined approach that implies processing terminology while taking the lexicographic approach. During this practical work we checked the compliance of terms' definitions of the dictionary with corresponding requirements of the ISO standard 704 (2009), with terminologists at the INT and continued the work on the compilation of dictionary entries, having taken recommendations from essential sources, comments and remarks of the specialists into consideration.

Although the NEGMD has a terminological character, we also apply some lexicographic tools and a lexicographic approach when compiling dictionary entries. Thus, we define the term in both English and Georgian and we also provide samples from English and Georgian corpora, thus taking the descriptive lexicographic approach.

As for the selection of terms, we are guided by the following principles: firstly, we try to consider general maritime terms; then we gradually move to more specific terms and later we try to select terms from all maritime sub-branches trying to keep the balance between different maritime sub-branches, such as, General Maritime English (GME) followed by Specific Maritime English (SME): Navigation, Mechanical Engineering, Electrical Engineering, Logistics, Maritime Law, Maritime Economics, etc.

As the NEGMD is a dictionary of maritime terminology, its macrostructure will mainly contain nouns. Possible verb-forms and other related words belonging to different parts of speech will be included in the "Related Words" section of the dictionary/terminological entry. Kageura (2015: 83) says that "Terms" are by definition content-bearing elements, as they represent concepts inside a domain. From the viewpoint of parts of speech, terms are mostly nouns. Although, sometimes verbs, adjectives or adverbs can be considered as terms they constitute a much smaller part than nouns". For example, the adverbs "offshore" and "inshore" are frequently used in maritime language and therefore represent separate terminological entries in our dictionary.

It is worth noting that the initial selection of terms to be considered for our dictionary is based on the principle of induction. Thus, we started with terms of General Maritime English, then moved to basic nautical terms, then we collected general terms from the maritime sub-fields of marine engineering and marine electrical engineering, logistics, etc. In the future we are going to cover terminology from all maritime sub-branches including the most specific terms of maritime fields and its sub-fields considering different terminological lemmas. We would characterize this work as a "layer on layer" approach, involving different levels of maritime terminology. In order to illustrate usage of the term in the language we select sample sentences from corpora of English and Georgian languages. During the process of sample sentences selection, we are guided by such principles as: concordance, sorting, sampling and filtering. Thus, we consider not only one maritime sub-branch but will try to cover all sub-branches simultaneously starting from elementary maritime terms belonging to General Maritime English (GME) identifying concepts from all maritime sub-branches and then moving to more specialized terms of different fields of Specialized Maritime English (SME), in parallel considering and working on entries covering all maritime sub-branches.

### 3. DESCRIPTION OF THE PROJECT

#### 3.1. THE THEORETICAL BASIS OF THE PROJECT

During the work on the dictionary we are guided by different sources such as the ISO standard 704 "Terminology Work. Principles and Methods", also taking into consideration recommendations regarding the compilation of specialized dictionaries given by L'Homme in her article "The processing of terms in dictionaries":

Recommendations for the compilation of specialized dictionaries:

- Dictionaries should consider user needs and include highly specialized but also less specialized items;
- Terminologists or specialized lexicographers should make more use of evidence found in corpora as a basis for making decisions about terms;
- Dictionaries should include more data on terms (e.g., collocations, valence patterns, images);
- Bilingual and multilingual dictionaries should account for inter-linguistic differences;
- Dictionaries should describe relationships between terms;
- Specialized dictionaries should contain encyclopedic or pragmatic information;
- Definitions should be structured in order to display key conceptual components" (L'Homme 2006:182).

When developing Georgian maritime terminology, specialists are to be guided by Georgian language rules, taking into consideration international norms to reach "balance" on the basis of such a combined method. Coinage of new Georgian maritime terms is the task of Georgian terminologists, yet the definition of terms mentioned by Roldan-Vendrel in the article "Emergent neologisms and lexical gaps in specialized languages" would be relevant to reach this aim:



“The structure of the definitions of the terms should:

- i. Describe the notion in exclusive reference to a specialized domain;
- ii. Distinguish that notion from others that may be interrelated within the language system;
- iii. Establish the semantic features according to which the relationships of antonymy or complementarity will be set;
- iv. Build a link between the term and its superordinate and subordinate terms;
- v. Expose the relationships of hyperonymy, hyponymy, synonymy, antonymy, complementarity, etc. that may exist between them (Roldan-Vendrel 2012:21).

Fernandez (2009:15) refers to the following characteristics of terminology stated by Cabre: “Terminology is itself interdisciplinary, especially in its most recent manifestations. It integrates contributions from several of the language sciences, namely:

- I. Knowledge theory: the relations between concepts and their possible nomenclature;
- II. Communication theory: the types of communicative situations that can emerge and their characteristic discourse;
- III. Language theory: how terminological units are related to natural language (Cabre 2003).

Since our dictionary is a terminological dictionary it is expedient to consider the essence of terminology in general. In this respect we would like to add some definitions of terminology and terms. Depecker (2009:v) says that: “Terminology science is to make the link between “sign”, “concept”, and “object” clear. The aim of terminology work is to ensure that a “sign” designates a precise “concept” and that the “concept” fits the “object” it describes” (“Handbook of Terminology”, vol. I, 2015:67). For the NEGMD we consider a concept as a unit of knowledge, and then move from the concept to the term. Therefore, each dictionary entry has only one meaning, since terms are monosemantic. In the ISO Standard 704 it is stated: “For a standardized terminology it is desirable that a term be attributed to a single concept” [2009:38]. Although monosemy of terms is sometimes discussed, as in Steurs (2009): “Although, one-to-one relation between concept and term in a clearly delineated specialist domain is preferable, very often, polysemy and dynamic changes in the meaning-form relation are witnessed”. Regarding terminology work the ISO Standard 704 very precisely states: “The goal of terminology work as described in ISO 704 is clarification and standardization of concepts and terminology for communication between humans”.

### 3.2.THE PRACTICAL BASIS OF THE PROJECT

NEGMD is a compilation of information currently available in the relevant English and Georgian sources. Entries of the NEGMD include the following information: English term, transcription, audio file, translation of the term, definition of the term in English, definition of the term in Georgian, illustration of the term in a relevant context (samples from the English language corpus), illustration of the term a relevant context (samples from the Georgian language corpus), related words section, graphical depiction of the term in the form of pictures or drawings/figures. We think that if dictionary entries are compiled in this way, it will increase learner’s interest and motivation to study better. The illustration of the usage of terms in the context will facilitate correct understanding and internalization of the term by the student. We decided to include corpora examples as “The lexical and syntactic environment in which a word appears turns out to be the most reliable indicator of the meaning it conveys in any particular instance (when several readings are theoretically possible)” (The Oxford Guide to Practical Lexicography, 2008:294. In the case of the dictionary/terminological entry as “related words” we are guided by the following criteria: a) grammatical relation, e.g. mentioning relevant verb-forms and b) synonymy relations, i.e. lexical identity, etc.

During dictionary compilation we are mainly guided by the following principles:

- 1) To bring one definition of a term in order to protect the principle of terminology monosemantic nature;
- 2) To select the definition that meets demands of students and specialists of the field at the same time;
- 3) To select corpora examples that illustrate the usage of a term in a best possible way;
- 4) To show semantic and grammatical paradigm of a word in order to enrich the knowledge of the dictionary user as terminological dictionary is mainly composed of noun-terms.

One of the most important issues to be taken into consideration during dictionary compilation is establishing the profile of the future user of the dictionary i.e. “User’s Profile”, confer Atkins and Rundell (2008: 486), who urge “First, catch your user. The user profile critically affects both what goes into the entry and how it is presented”. Our dictionary is aimed at different audiences involved in the maritime field, from students of Maritime Education and Training (MET) institutions to qualified specialists in the field, employed either on board ships or on shore-based jobs. Our dictionary bears features of both monolingual and bilingual dictionaries, as it offers translation of the terms and its definitions both in English and Georgian. Therefore, it can be used both by Georgian users and English-speaking users as the dictionary comprises definitions and corpora examples both in Georgian and English. Due to variety of information it includes, the



dictionary can be classified as learner's dictionary and scholarly dictionary at the same time. Thus, our aim is to design a dictionary that will enable the users to decode and encode meanings of words.

Terms are monosemantic, but in some cases we also deal with homonymy, as is the case with the word "port". Thus, in the NEGMD we have three terminological entries related to the word "port":

### I. port n

[ pɔ:t ]

საზღვაო ნავსადგური, პორტი

place on a coast or shore which boats use to load and unload or shelter from storms

პორტის აკვატორიის ნავმისადგომის მიმდებარე ნაწილი (ჩვეულებრივ, აუზის სახისა, რომლის ორ ან სამხრეთ განთავსებულია ნავმისადგომი ნაგებობები), სადაც სატვირთო ოპერაციები წარმოებს

### II. port n

[ pɔ:t ]

გემის მარცხენა მხარე, მარცხენა ბორტი

left side of a ship, looking forward; larboard. to port: on or towards the left side of a ship, etc. *opposed to starboard*

ბაკბორტი, ბაკბორდი, გემის მარცხენა მხარე (ბორტი) თუ მას კიზოდან ცხვირისკენ ვხედავთ

### III. port n

/pɔ:t/

სარკმელი

opening in engine cylinder through which gases enter or leave a cylinder

შემშვები ან გამომშვები ნახვეტი, ხვრელი, არხი (მრავასი, ტუმბოსი და ა.შ.)

At present stage of compilation, we select terms given in the textbooks' glossaries of different maritime sub-branches as a term bank. Maritime dictionaries of other languages can also be used as term banks in the future.

## 3.3.CREATION/COINAGE OF GEORGIAN MARITIME TERMINOLOGY

Lexical and terminological gaps constitute a major problem of Georgian maritime terminology. Thus, one of the aims of our project is to stimulate and speed up work on coinage of missing maritime terms in Georgian. Some English Maritime terms have equivalents in Georgian, whereas some do not and this is the problem we intend to solve by addressing issues that have been discussed in the present report.

For example, if we take such terms as:

- Draft, draught – the distance between waterline and the lowest point of the vessel (keel) – (Georgian) გემის წყალშიგო
- Displacement – the amount of the water that is displaced by the ship when it moves forward – (Georgian) წყალწყვა
- Air draft – the distance between waterline and the highest point of the vessel (mast, etc.) – (Georgian) - ..... (lexical/terminological gap).

The first two terms have equivalents in Georgian but the third one "air draft" does not. The meaning of the term can be rendered to Georgian students only via definition. When we included the term "air draught" in our dictionary, we just omitted its Georgian meaning, giving definitions in English and Georgian and relying on corpora of these languages. The Georgian equivalent of the maritime term "air draught" will be given in the dictionary when competent specialists fill this "lexical/terminological gap" in Georgian. Before this task is fulfilled the entry looks in the way given below:

### air draft col

-

distance from the surface of the water to the highest point on a vessel

მანძილი წყალხაზიდან გემის უმაღლეს წერტილამდე

#### Corpus examples

The vessels will be modified with 5m air draft to cope with working under jetties and will also feature rope guards to protect the vessel and selected electronics.

-

In the case of new maritime terms, Georgian students should be offered meaning equivalents to terms already existing in English and other languages. Neologisms and neoterms<sup>1</sup> are means to solve this problem in Georgian. Usually, neologisms enter languages to designate new concepts and this is conditioned by technological progress and new achievements or innovations in the field. Specific to the development of Georgian maritime terminology is the fact that new terms have to be invented both for new and already existing concepts. This is very important, confer Cabre (2012:1): "Neologisms represent the constant changes of a society and are a clear indication of the vitality of a language".

<sup>1</sup> "A neoterm created to designate a concept is a type of neologism and is called a neoterm. Although most neoterms designate new concepts, some designate established concepts" [ISO 704 (2009:34)].



ISO 704 (2009:51) states that “the following term formation methods apply to the English language and may also apply to other languages:

- Creating neoterms;
- Using existing terms: (adding one or more morphological elements or affixes to a root or a word);
- Translingual borrowing;
- Derivation, compounding (complex terms, phrases, blends) abbreviations.

In ISO 704 (2009:38-39) we read: “The following principles should be followed in the formation of terms and appellations: transparency; consistency; appropriateness; linguistic economy; derivability and compoundability; linguistic correctness; preference for native language”.

The creation and coinage of Georgian maritime terminology require us to define methods of creation and development of Georgian maritime terminology, such as: transliteration, derivation, loans: direct loans, loan translations, calques, neologisms, adoption of international terms. There are different methods of terminology derivation. Sometimes even calques can be used. Considering that English is recognized by the International Maritime Organization (IMO) as the official language of maritime industry we would like to recommend using English calques. In this way internationalization of English maritime terms will take place. It is also important to ensure the co-existence of national terms and terms that enter the language in the above-mentioned ways. Thus, the issue of creation and derivation of new Georgian maritime terms is of importance as there are a lot of “lexical/terminological gaps” in maritime Georgian. For new maritime terms Georgian students should be offered meaning equivalents of maritime terms already existing in English and other languages. We will have to deal with neoterms. While creating Georgian maritime terminology, specialists have to be guided by Georgian linguistic rules. In ISO 704 (2009:41) it is stated in this respect that “when neoterms or appellations are coined, they should conform to the morphological, morphosyntactic and phonological norms of the language in question”. For example, in English we often see terms created on the basis of “noun + noun” model/formula but it does not mean that if some lexical calques are permitted, grammatical calques should be allowed too. Otherwise it may lead to deforming the language.

The development of new Georgian maritime terms is to be solved in a nationwide cooperation between specialists of different organizations, confer Fernandez (2009:18): “The first consequence of this quiet revolution in terminology is the need for much more flexible communication between descriptive terminologists, lexicographers, disciplinary specialists and the standardization organizations”. The State Chamber of Language of Georgia and relevant linguistic institutions will be responsible for the organization and coordination of this work.

During the work on the NEGMD project “lexical/terminological gaps” in Georgian maritime terminology will be revealed and thus this project will stimulate the creation of Georgian maritime terminology that implies several organizational issues, such as: revision of existing maritime dictionaries, choosing the right policy for the creation of Georgian maritime terminology, choosing the methods of creation of maritime terminology and establishing a project team.

#### 4. CONCLUSIONS

The NEGMD project will contribute to both the maritime semantic field and Georgian linguistics. It will bring the following benefits: stimulation of creation/coinage of Georgian maritime terminology, application of the newly formed maritime terminology in textbooks, maritime documentation and materials, economy of space and facilitation of understanding among specialists and students in the maritime field, improving the system of maritime education and training and, consequently, of the national maritime industry, contribution to the lexicography of Georgian; creation of the lexicography of the maritime semantic field, constructing the basis for establishing Georgian maritime corpus with its further integration in the corpus of the Georgian language; in this way making general contribution to Georgian linguistics.

The general objective of our project is to improve the system of Maritime Education and Training (MET) of Georgia, thus expanding the Georgian maritime field. The project will take several years, but as it can be launched on the Internet as soon as a certain number of terminological entries have been compiled, it will be available for those involved in the process of studies at MET institutions and for maritime professionals of Georgia way before its completion.

#### 5. REFERENCES

- Atkins B.T., Rundel M. (2008) The Oxford Guide to Practical Lexicography, Oxford University Press
- Cabre M. (2012). Neology in Specialized Communication. In *Terminology*, 18(1) pp. 1-8.
- Fernandez T. (2009). Terminology and Terminography for Architecture and Building Construction. In *Terminology*, 15(1), pp.10-36.



- Depecker Laic (2015). How to Build Terminology Science? In *Handbook of Terminology*, vol. I, pp. 34-44
- Hacken PuisTen (2015). Terms and Specialized Vocabulary: Taming the Prototypes. In *Handbook of Terminology*, vol. I, pp.3-13
- Humble J, Palacios G. (2012). Neology and Terminological Dependency. In *Terminology*, 18 (1), pp. 59-85
- ISO Standard 704 (2009), "Terminology work – Principles and methods"
- Kageura Kyo (2015), Terminology and Lexicography. In *Handbook of Terminology*, vol. I, pp.45-49
- Geeraerts D. (2015). Foreword to "Handbook of Terminology" vol. I, pp. xvii- xix
- Le Serrec A. (2010). Automating the Compilation of Specialized Dictionaries. In *Terminology*, 16(1), pp.77-106.
- L'Homme Marie Claude (2006). The Processing of Terms in Dictionaries. In *Terminology*, 12(2), pp.181-188.
- Roldan-Vendrel M. (2012). Emergent Neologisms and Lexical Gaps in Specialized Languages. In *Terminology*, 18(1), pp.9-26.
- Steurs F., Wachter K., Malsche E, (2015). Terminology Tools. In *Handbook of Terminology*, vol. I, pp.229-249







# Issues in linking a thesaurus of Macedonian and Thracian gastronomy with the Languag system

Toraki K., Markantonatou S., Vacalopoulou A., Minos P., Pavlidis G.

*Institute for Language and Speech Processing, Athena R.C., Athens, Greece*

## Abstract

As part of our project entitled “GRE-Taste: The Taste of Greece,” we have been developing a trilingual (Greek, English and Russian) thesaurus of food served in restaurants in Eastern Macedonia and Thrace. To this end, we have designed a web thesaurus development environment, and have defined facets and subfacets corresponding to the following major categories: Foods (both as ingredients and as dishes); drinks; food sources (and parts thereof); places of origin; preparation methods; functions; state; and, nutrition. For each concept, the preferred (most common) and non-preferred (synonym and hidden) terms are entered; nutritional, cultural and other information appear in separate fields, as do the relationships between concepts (e.g. between a dish and its ingredients, cooking methods or place of origin). In this paper, we examine how the Languag thesaurus can be used to code foods, and as well as the issues involved in the process, such as those related to confusing descriptions and the absence of corresponding Greek dishes. We offer a suggestion for the enrichment of the Languag thesaurus aimed at producing an outcome that might ensure harmonization and interoperability among different applications. We also put forth a proposal that might help resolve Greek terminology challenges encountered in the description and classification process of foods and also regarding issues that arise related to other gastronomic concepts.

**Keywords:** multilingual thesauri; culinary terminology; culinary lexicography; Languag thesaurus; food classification; food description

## 1 Introduction

As part of our project entitled “GRE-Taste: The Taste of Greece,” we have been developing a trilingual thesaurus of food served in restaurants in Eastern Macedonia and Thrace. We have designed and implemented a lexicographic web environment that accommodates a set of texts retrieved from restaurant menus in the study area. This has enabled the development of a thesaurus containing information on food-related dishes and concepts, which is complemented by dietary and cultural information concerning the dishes and their ingredients. The objective of this project is to support local travellers in their quest for gastronomical and cultural experiences by developing a multilingual tool for the search, retrieval and presentation of information on food according to the following specific criteria: Name; ingredients; source; preparation method; state; place of origin; function in the meal; health issues, and others.

In this paper, we will present the efforts to harmonize our work with the Languag thesaurus, an international tool for food description, with linked data in mind. We have recorded the lexicographic issues we have encountered to date in practice, and offer some suggestions regarding the naming, description and classification of dishes.

## 2 Background

A lot of food-related information is furnished by the media, focusing on various ingredients and preparation guidelines, as well as concerning quality, health, nutrition and other relevant aspects. On the one hand, all of this information may not only be overwhelming, but it may also prove confusing to consumers who must be informed on issues concerning their desires and needs, and also on the effects produced by what they are offered. The terminology used must be appropriate, familiar and properly understood in order to facilitate the culinary experience and also to allow consumers to make healthy choices with ease (Himmelsbach et al. 2014). Additionally, there is a lot of ongoing research internationally on food-related matters, such as food knowledge bases and food semantics, food description and classification, and food search and discovery, most of which aim at extracting food-related information from different data sources by way of specific criteria (Durazzo et al. 2019; Gateau et al. 2019; Haussmann et al. 2019; Zulaika et al. 2018). Drawing upon this research, corresponding systems are being developed that take into account the different needs and aspects of food naming, processing, uses and also other elements, such as local culture, health and nutritional issues.

Food classification and description systems cover specific user needs. Therefore, for the same food product – e.g. a pork product – the classification may differ depending upon whether the item appears in a nutrient database, a consumption database or a contaminant database. These systems have been implemented by regional or international organizations such as the FAO,<sup>1</sup> EuroFIR,<sup>2</sup> the European Union,<sup>3</sup> the Codex Alimentarius Commission<sup>4</sup> and others (Ireland et al. 2002; Ireland & Møller 2000; Ireland & Møller n.d.; FAO 2015). Examples of such systems include: The Languag and Agrovoc Thesaurus; EuroFIR; Eurocode/EFG and INFOODS, among others (Ireland & Møller 2016). Food systems at the national

<sup>1</sup> <http://www.fao.org/infoods/infoods/en/>

<sup>2</sup> <https://www.eurofir.org/>

<sup>3</sup> <http://www.efsa.europa.eu/>

<sup>4</sup> <http://www.fao.org/fao-who-codexalimentarius/en/>



level are also being developed, and aim to cover local needs by taking into account specific cultural, economic, social and other conditions. Another interesting advance in food description is the development of food ontologies, such as FoodOn.<sup>5</sup> In addition to the relationships found in a thesaurus, this “field-to-fork” food ontology includes additional terms, such as “has quality,” “has part,” “is immersed in,” “output of” and so on (Dooley et al. 2018).

## 2.1 The Languag system

Langua is a multifaceted thesaurus system containing terms for the description of foods from different vantage points (i.e. food groups, cooking methods, preservation methods, consumer group and geographical origin). The use of a multifaceted structure permits the description of a food product from several perspectives and allows a search for foods based on a variety of criteria, such as the ability to look for baby food containing cereals or to seek potato-based dishes that have been fried. In order to group foods, Languag uses a number of classification systems, each of which serves a different purpose or a different application area. Each system uses its own description and classification system, which means that information about food is not modelled in the same way or contains the same degree of detail. A unique code (i.e. a Languag code) is assigned to each concept regardless of the classification system, and this can be used to identify a particular food or any other concept described in Languag. Therefore, the Languag thesaurus can be used by specific food databases from different countries, ensuring harmonization and interoperability among different applications. Controlled terms are used for the representation of concepts describing specific foods, as well as other related issues (processes, methods, states, etc.). Concepts are structured hierarchically into the following facets, which correspond to the different angles mentioned above:

- A. Product type
- B. Food source
- C. Part of plant or animal
- E. Physical state, shape or form
- F. Extent of heat treatment
- G. Cooking method
- H. Treatment applied
- J. Preservation method
- K. Packing medium
- M. Container or wrapping
- N. Food contact surface
- P. Consumer group / dietary use / label claim
- R. Geographic places and regions
- Z. Adjunct characteristics of food

Each facet is used independently. More than one term from each facet can be used, depending on specific needs and uses. Facet A is the basic facet for the description of food products by type. In the case of multi-ingredient foods, Languag suggests indexing major ingredients by weight, without taking water into consideration; however, specific mixture terms can also be used if one constituent is the first ingredient and the other from the second to fourth ingredient (Ireland & Møller 2013).

The structure of the thesaurus follows the rules for the construction and display of thesauri in ISO international standards (ISO 25964-1; ISO 25964-2). Languag is published in a basic English version (Møller & Ireland 2018a) and in a multilingual version in English, Czech, Danish, French, German, Italian, Portuguese and Spanish (Møller & Ireland 2018b). Compared to the aims of our thesaurus, those of Languag are somewhat different, in that the latter represents food mainly from the perspective of its production and distribution on the market; in contrast, our thesaurus aims at representing the conceptual domain that emerges from the included menus, whose main aim is to facilitate and attract customers to an establishment that offers and/or serves prepared food.

## 3 Project description

The lexicographic web application that we have designed and implemented provides the interface for the presentation of the content of the restaurant menus we have collected and for the development and display of the thesaurus. The menus were digitized using OCR technology and the resulting data were entered into a database comprising the set of texts to be used as a source for the thesaurus entries. The content of this menu-derived text collection is entered into specifically-structured fields, such as those of dish name, category, description and place of origin.

The thesaurus is organized into facets and subfacets that correspond to major categories that have been identified through the analysis of the names of dishes on the menus (Markantonatou et al. in preparation); examples of these are the main ingredient of a dish, the way the dish is prepared and the way a piece of meat is obtained from the animal (the “cut”). Several of these categories also exist in Languag: Foods (two subfacets for foods – one as ingredients and the other as dishes), drinks, food sources (and parts thereof), places of origin, preparation methods, functions (courses in the meal), state of food, cuts of meat and nutrition.

When recording an entry, the term is entered in the three main languages of the project – Greek, English and Russian – as well as in Latin, if there is a scientific name (for animals and plants). The most common term is noted as the preferred one in each language.<sup>6</sup> Alternative and dialectic or idiomatic variants are recorded as synonymous or hidden terms. Any

<sup>5</sup> FoodOn: <https://foodon.org/>. More information on food ontologies is given by Haussmann et al. (2019).

<sup>6</sup> This is because, at present, there is no standard food terminology in Greek specifically dealing with restaurant menus.



relationship between one concept and any other of the same or of a different category is also recorded, such as one between a dish and its ingredients, cooking methods or place of origin. Detailed nutritional information is coded in specific facets, while cultural information and any other useful elements (e.g. recipes) are entered as free text in dedicated note fields. The thesaurus contains more than 1,550 concepts denoted by more than 3,500 terms; concepts interrelate with 35 relation types, producing more than 3,250 relation instances. Thus, this may be viewed as a highly complex project that attempts to harmonize very different issues related to lexicography, terminography and other areas, given the following: Firstly, there is a variety of names and ways of preparation of dishes; secondly, no index is available of controlled food terms in Greek; and, thirdly, no similar research focusing on dishes from restaurant menus has been conducted in Greece to date.

The unique feature of our thesaurus is that all dishes and most ingredients are drawn from the menus of approximately 120 restaurants in Macedonia and Thrace. Consequently, the thesaurus represents the actual language used on the market. In order to list foods by category, to define relationships between concepts, and to link and harmonize our thesaurus with international resources, we also draw on data from official sources, such as the National Code for Foodstuffs and Beverages, European regulations and Languag.

The screenshot displays the web interface for the GRE-TASTE thesaurus. On the left, a hierarchical tree structure shows the classification of food terms, with 'μπακαλιάρος τηγανητός' (fried cod) selected. The right side shows the detailed entry for this term, including its classification, synonyms, and relations.

**μπακαλιάρος τηγανητός [#1453]** Τροφές → Τροφές-πιτάτα → ψάρια και θαλασσινά

→ μπακαλιάρος (πιτάτα)

☐ Facet

**Terms** [New Term](#)

Term	Article	Type	Language
μπακαλιάρος τηγανητός	{ο}	Προσμηούμενος	Ελληνικά
τηγανητός μπακαλιάρος	{ο}	Συνώνυμος	Ελληνικά
fried cod	Empty	Προσμηούμενος	Αγγλικά
pan-fried cod	Empty	Συνώνυμος	Αγγλικά

**Relations** [New Relation](#)

Type	Concept
Constituency and Function of the food	
Βασικό συστατικό	πιστός μπακαλιάρος [<Τροφές> > <Τροφές-συστατικά> > ψάρια και θαλασσινά > ψάρι > μπακαλιάρος
Τρόπος παρασκευής	τηγάνισμα [<Τρόποι παρασκευής> > <θερμική ή χημική κατεργασία>]
Λειτουργία στο γεύμα	κύριο πιάτο [<Λειτουργίες>]
Nutritional and Dietary information	

Figure 1: The web environment for the GRE-TASTE thesaurus.

Figure 1 shows the entry for “fried cod.” On the left part of the screen, a tree-like representation of the hierarchical structure of the thesaurus is provided, which allows the lexicographer to put the concept in the appropriate place in the hierarchy with a graphic interface. On the right part of the screen, information related to the concept is offered and can be edited. The interface for handling the terms denoting the concept is located in the “Terms” area, while the interface for editing the relations between concepts is found in the “Relations” section.

### 3.1 Semantic and terminological issues

Taking into consideration the fact that, on the one hand, the names of foods on menus do not usually adhere to any particular rules and that the same dish may have different names in different restaurants and, on the other hand, that we wish to keep the names used on the menus in our system, we had to make decisions on the naming of food products, especially dishes, as well as on syntax and the selection of preferred and non-preferred terms.

The grouping of foods in the thesaurus is based on the main ingredient, the definition of which frequently does not depend on quantity alone, but also on the essential property of the product connected with particular local and cultural data. Such groups are meat dishes, fish and seafood dishes, vegetable dishes and pasta dishes, among others, while the facility for multi-hierarchical relationships in the platform allows us to classify foods into more than one hierarchy. So, for instance, specific legume-based dishes, such as “arakás” (peas), are classified under legumes as well as under “laderá”; the latter is a special and prominent type of Greek dish, one basically cooked in olive oil. Thus, we decided to create a separate category for “laderá” (although olive oil cannot be considered to be the main ingredient).

There are problems in Greek culinary terminology and lexicography when it comes to the description of food. A common issue is that there are several names for the same dish or for very similar dishes. For example, restaurants serve “Greek salad” or “horiátiki,” both with the same ingredients; consequently, these have to be listed as synonyms in the thesaurus, with “horiátiki” as the preferred term, based on frequency. Another similar example is lettuce salad. In Greek, it may be “maroulosaláta,” “maroúli” or “saláta maroúli,” but it may also have a particular name in a specific restaurant, e.g. “to maroúli tis Elénis” (Helen’s lettuce). These dishes have the same main ingredient but may differ in other respects. Thus, lettuce salad may or may not include onion, rocket, etc., while specially named dishes have a unique description, e.g. the salad “to maroúli tis Elénis” also contains yoghurt sauce and prosciutto. In our thesaurus, we decided to also list



alternative and optional ingredients, so as to cover different types of dishes, presenting them as specific concepts in addition to the more “generic” entry.<sup>7</sup>

### 3.2 Working with Languag, comparisons and issues thereof

As already mentioned, the various food coding systems have different levels of detail, depending on the kind of applied system, intended use, needs and so forth. As a result, the coding is different, for example, if the system focuses on nutrient intake (mostly covering processed foods and foods as consumed) or on hazard occurrence in food (interest in raw commodities) (Ireland & Møller 2016). Our system contains entries of dishes served in restaurants and, additionally, of ingredients used to prepare those dishes. To implement the Languag encoding, we had to decide which coding used in it was the most appropriate in our case, of course in combination with the general decisions we have taken concerning the project objectives, the intended users, the structure of our thesaurus, the depth of indexing, etc. Taking into account such parameters, we decided to use EFSA coding as a more complete system and as the one with the greatest detail in description and classification analysis, and also the EuroFIR, but mainly for reasons of formality (it is a requirement for member countries, but is usually too generic), and any other system if the aforementioned two did not cover a particular food product.

We chose to work mainly on facet A – that is, food products – as this is the most important one of all for the description of the foods covered in our project. At the same time, however, facet A presents difficulties as regards the identification and selection of the appropriate coding and the harmonization with our data, not only due to language issues but also to cultural, geographical, environmental and other particularities. The other facets are also used for coding, but concept correspondence is more straightforward, in their case.

Consequently, our concerns are the following: Firstly, to describe and classify foods using the Languag system; and, secondly, to decide on the names of foods, especially those of dishes. Here are some examples illustrating the kind of challenges we have faced with Languag thus far:

- Both our thesaurus and Languag are based on the description of the main ingredient (facet A), which, as noted in 3.1, may depend not only on quantity (as in Languag) but also on cultural and historical data connected with particular dishes. The use of polyhierarchical relationships such as those found in Languag is the solution for cases where more than one place is necessary for coding food (either a dish or any other food product).
- A basic problem is that Languag does not contain a number of culture-specific foods from Greece. This is something that we had expected, but we think that such dishes should be included, not only because they are of local interest but even more so because of the need for understanding, communication and compatibility between local cuisines and cultures.
- In Greek cuisine, we find several dishes sharing the same overall concept but with a varying main ingredient. Such is the case of the different croquettes and “keftédes.” The menus studied propose a large variety of dishes in which the main ingredient can be tomatoes, fava (split peas), eggplants, pumpkins, meat, fish, potato, cheese or another ingredient. Some of these – such as “meatballs” (“keftédes”), “fish balls” (“psarokrokétes”) and “potato croquettes” (“patatokrokétes”) – are found in Languag, but “cheese croquettes” (“tyrokrokétes”) and others are not. Furthermore, the use of a generic concept – such as “vegetable-based dishes” – does not cover the requirement that these dishes be coded as members of a particular food family. Thus, in the “Methods of Preparation” facet, apart from the subfacet for “Cooking Methods,” we have introduced another one for “Shape, Texture and Form.” Consequently, “keftés” is a term in this subfacet; “avgolémono” (egg and lemon sauce), “kebáp” and “gemistá” (stuffed vegetables) are other examples.
- The same issue can be observed in the case of legume-based dishes; in other words, not all types of legume are to be found in Languag. In the case of pea-based dishes, for example, we only find “mushy peas,” which is a rather uncommon type of dish in Greece (“arakás”).
- “Stifádo” is another typically Greek dish not included in Languag. It is a cooking method which goes as far back as to Byzantine times. On menus, “stifádo” may describe specific dishes such as “rabbit stew”<sup>8</sup> (“kounéli stifádo”), “veal stew” (“moschári stifádo») or “cuttlefish stew” (“soupiés stifádo”). For this reason, we have assigned a particular code for “stifádo” as a cooking method in our thesaurus, while retaining the term “stifádo” as part of the name of particular dishes, and classifying them within the group of the basic ingredient (rabbit dishes, veal dishes, cuttlefish dishes, and the like).
- The ingredients of some composite dishes may also be offered as distinct dishes. Examples of these are “codfish with garlic dip” (“bakaliáros skordaliá”), “pasta with minced meat” (“makarónia me kimá”) and “sautéed potatoes with bacon and mushrooms” (“patátes soté me béikon kai manitária”). The general rule for dish classification is applied here as well – namely, according to the main ingredient – while dishes can also be assigned additional codes corresponding to further classifications, if necessary. An example of the latter is “makarónia me kimá,” which is coded as a pasta dish but is also assigned a code in meat dishes. As for what applies in Languag, composite dishes are indeed coded in it, though this does not cover all special dish types. In such cases, we can classify the dish under the closest category, or we may decide that a new code is necessary if the dish has a connection to a specific cultural or local background. Thus, “patátes soté me béikon kai manitária” (sautéed potatoes with bacon and mushrooms) is assigned the Languag code for “potatoes, meat and vegetable meal,” followed by the description: “The group

<sup>7</sup> More detailed information is given in Markantonatou et al. (2019). Synecdoche issues are also presented there. For instance, in order to distinguish references to plants from references to dishes, we add the word “dish” in parentheses in the dishes subfacet, i.e. “maroúli” (lettuce) for the plant and “maroúli (dishes)” for dishes based on “maroúli.”

<sup>8</sup> We follow the general trend seen on Greek menus and translate “stifádo” as “stew,” but we also note that this translation only partially describes this special way of cooking, which uses shallots whose shape – most importantly – must be preserved. This demand dictates the overall cooking procedure employed.



includes any type of composite dish based on potatoes, meat, and vegetables. More detailed information on the characterising ingredients can be added with additional facet descriptors.” One such facet descriptor is the manner of cooking (sautéed). But in the case of “bakaliáros skordaliá” (codfish with garlic dip), which is both a ceremonial and popular Greek dish, a new code is assigned.

- Another issue is making decisions with regard to items offered on menus for which there is no corresponding entry in Languag. For example, we often find the description “handmade” (“handmade tzatziki,” “handmade marinated sardines,” and so forth). We have opted to ignore these descriptions in our thesaurus because these features do not distinguish the dish (in other words, they do not characterize it). It is open to discussion, however, whether this is a feature of interest for restaurant customers that our thesaurus may have to take into account in future.<sup>9</sup>
- Problems also arise from differences between the names of foods and their varieties, ranging from the description of specific concepts (that could be enriched through additional content not appearing at present in the international version of the thesaurus) to differences in the classification of concepts on the menus and in Languag. For example, Languag includes “Greek salad” as a specific dish under salads, describing it in the same way as “horiátiki,” but it does not mention that “horiátiki” is an alternative name for “Greek salad” (although it is a rather well-known term for it). In addition, Languag does not include “tomato salad” (“ntomatosaláta”) or “tomato and cucumber salad” (angourontomáta), two types of salad often served in Greek restaurants and included in our source menus. Green salad may also vary as lettuce salad in the example above.
- In some cases, similar situations may be coded within different facets, which may result in inconsistencies or may create difficulties in selecting the right classification concept. For example, fat content in food is found in facet P (for label claim), in facet Z (for fat content in Eurocode2), and in facet H (for fat removed), while for specific products with fat content in facet A we find mixture terms for milk, milk powder, mayonnaise, salad dressing and cheese.
- It is not always clear what is defined by the terms used, often because of the different classification systems that are included on the same platform that do not bear any relation to each other. This is why Languag managers often emphasize that the term alone is not enough to select a concept and that the description text is more useful. For example, searching for salted seafood such as fish, prawns and other seafood preserved by salting, in facet A we find only one specific product represented by the mixture term “salted cod” (with the scientific name “bacalao”), and the general category “salted seafood” in the same system (EFSA FOODEX2) with the description “The group includes Seafood (any non-mammal, non-fish marine animal) product essentially preserved by salting” (in other words, fish is not included in this category). A solution is offered, of course, in facet J with the descriptor “preserved by salting.” As far as cod is concerned, in facet A we also find “cod, dried” (scientifically named “gadus”), including, as noted in the description, “any type of dried cod” (that is, salted or unsalted). This could potentially cause confusion as to where the Greek food “pastós bakaliáros” (salted and dried cod) should be classified.
- Similar confusion is created in Languag in the case of salads with various basic ingredients (such as potatoes, pasta, rice and others). Thus, in EFSA, “pasta salad” is classified under salads as a special salad dish, while in EuroFIR “macaroni salad” and all other salads are coded as “prepared salads” together with “potato salad,” “rice salad,” “tuna salad” and others. In our thesaurus, each salad is classified according to its basic ingredient (e.g. pasta, potato, rice, tuna and so forth), while all these dishes are also connected through the concept of “salad” in the “Functions” facet.
- Another general issue is the lack of recipes for dishes in Languag, which sometimes may arise as a consequence of an inadequate listing of basic ingredients.
- Not all content found in Languag facets is mirrored exactly in our thesaurus. For instance, the ingredients needed to prepare a dish are included in our subfacet “Ingredients,” while in Languag they are found in facet H (TREATMENT APPLIED, category ADDED).

#### 4 Conclusions and further work

Languag is a useful tool for food coding systems, and the uniform Languag code for foods that it provides makes it equally useful for harmonization and interoperability between food databases. The problems, however, presented above show that it cannot be used as is for applications such as the one described here. The GRE-Taste Food Thesaurus is based on data collected from restaurant menus, and is intended to be used mainly by customers in restaurants and similar establishments. The food served varies widely in terms of the content of dishes; the terminology used is also very diverse, as it represents several language-specific, local, idiomatic characteristics, often not easily understood by users. Consequently, merely adjusting our data to Languag is not enough; working with Languag coding, we have to add new information, which in some cases will be able to supplement the description of existing concepts, though sometimes new concepts will need to be added for dishes that do not exist as separate entries within it.

In the future, we aim to continue enriching our thesaurus through the implementation of our classification scheme and the structure of the ontology we have designed. This task involves research on diverse issues such as the semantic and syntactic structure of the names of the dishes, dish-naming strategies and the formal problems concerning ontology, due to the rather disorderly nature of the content of restaurant menus; however, these issues lie beyond the scope of this discussion.

Food “constitutes a rich and complex cultural system [...] embracing history and geography,” together with language, social studies, race and ethnic identity and other disciplines, as noted by Faber & Vidal Claramonte (2017: 156). Languag and the other food systems show less interest in specific local and cultural parameters. Our project aims to play a vital role in this direction as well, by improving and supplementing existing systems with this local wealth of information on food connected with specific historical, cultural and social data. In this way, it is hoped that our project will also contribute to the tourism economy, to communication between different peoples and cultures, and to sustainable health and nutrition.

<sup>9</sup> Facet Z in Languag provides the code Z0109 for HOME PREPARED food.



## 5 References

- Dooley, D.M., Griffiths, E.J., Gosal, G.S., Buttigieg, P.L., Hoehndorf, R., Lange, M.C., Schriml, L.M., Brinkman, F.S.L., & Hsiao, W.W.L. (2018). FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. In *Npj Science of Food*, 2(1), 23. Accessed at: <https://doi.org/10.1038/s41538-018-0032-6> [22-04-2020]
- Durazzo, A., Camilli, E., D'Addezio, L., Sette, S., Marconi, S., Piccinelli, R., Le Donne, C., Turrini, A., & Marletta, L. (2019). Italian composite dishes: Description and classification by LanguaL™ and FoodEx2. In *European Food Research and Technology*, 246(2), pp. 287-295.
- EFSA. (2015). *The food classification and description system FoodEx 2 (revision 2)*. EFSA Supporting Publications, 12(5) Accessed at: <https://doi.org/10.2903/sp.efsa.2015.EN-804> [23-04-2020].
- Faber, P., & Vidal Claramonte, M.C.Á. (2017). Food terminology as a system of cultural communication. In *Terminology*, 23(1), pp. 155-179.
- FAO (Food and Agriculture Organization of the United Nations). (2015). *Guidelines on the collection of information on food processing through food consumption surveys*. Accessed at: <http://www.fao.org/3/a-i4690e.pdf> [22-04-2020].
- Gateau, B., Stahl, C., & Pedretti, O. (2019). Creating a semantic food knowledge base with cooking recipes for a meal recommender system. In *Semantics 2019*, Karlsruhe, 10 September 2019. Accessed at: <https://2019.semantics.cc/sites/2019.semantics.cc/files/Stahl-SEMANTICS2019-publishing.pdf> [25-04-2020]
- GCSL (General Chemical State Laboratory of Greece). (2018). *National Code for Foodstuffs and Beverages*. Accessed at: [http://www.gcsl.gr/index.asp?a\\_id=365&txt=y&show\\_sub=1](http://www.gcsl.gr/index.asp?a_id=365&txt=y&show_sub=1) [23-4-2020].
- Hausmann, S., Seneviratne, O., Chen, Y., Ne'eman, Y., Codella, J., Chen, C.H., McGuinness, D.L., & Zaki, M.J. (2019). FoodKG: A semantics-driven knowledge graph for food recommendation. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (eds.), *The Semantic Web – ISWC 2019*. Cham: Springer, pp. 146-162.
- Himmelsbach, E., Allen, A., & Francas, M. (2014). *Study on the impact of food information on consumers' decision making: final report*. TNS European Behaviour Studies Consortium. Accessed at: [https://ec.europa.eu/food/sites/food/files/safety/docs/labelling\\_legislation\\_study\\_food-info-vs-cons-decision\\_2014.pdf](https://ec.europa.eu/food/sites/food/files/safety/docs/labelling_legislation_study_food-info-vs-cons-decision_2014.pdf). [23-04-2020].
- Ireland, J.D., van Erp-Baart, A., Charrondière, U.R., Møller, A., Smithers, G., & Trichopoulou, A. (2002). Selection of a food classification system and a food composition database for future food consumption surveys. In *European Journal of Clinical Nutrition*, 56(2), S33–S45. Accessed at: <https://doi.org/10.1038/sj.ejcn.1601427> [23-04-2020].
- Ireland, J.D. & Møller, A. (n.d.) *Guidelines for food classification and description in food databases*. Accessed at: <http://www.languaL.org/download/Posters/Ireland%20and%20M%20C%20B%20%20%20Food%20classification%20and%20description.pdf> [23-04-2020].
- Ireland, J.D., & Møller, A. (2016). Food classification and description. In B. Caballero, P. M. Finglas, & F. Toldrá (eds.) *Encyclopedia of Food and Health*. Waltham, MA: Academic Press, pp. 1-6.
- Ireland, J.D., & Møller, A. (2013). Describing a food using LanguaL™ facets A-Z. In *Food Comp 2013, Wageningen, The Netherlands, 13-25 October 2013*. Accessed at: [http://www.languaL.org/download/Presentation/LanguaL\\_facets\\_A-Z\\_2013-10-17.pdf](http://www.languaL.org/download/Presentation/LanguaL_facets_A-Z_2013-10-17.pdf) [23-04-2020].
- Ireland, J.D., & Møller, A. (2000). Review of international food classification and description. In *Journal of Food Composition and Analysis*, 13(4), pp. 529-538. Accessed at: <https://doi.org/10.1006/jfca.2000.0921> [23-04-2020].
- ISO 25964-1:2011 Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval.
- ISO 25964-2:2013 Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies.
- Markantonatou S., Vacalopoulou, A., Minos, P., Toraki, K. & Pavlides, G. (2019). Thesaurus of gastronomy in Macedonia and in Thrace. In: *12th ELETO Conference «Hellenic Language and Terminology» Proceedings, Athens, Greece, 7-9 November 2019*. Accessed at: <http://www.eleto.gr/en/papers.htm#12thPapers> [28/04/2020].
- Møller A., Ireland J. (2018a). *LanguaL™ 2017 – The LanguaL™ Thesaurus*. Technical Report. Danish Food Informatics. Accessed at: DOI: 10.13140/RG.2.2.23131.26404/ [23-04-2020].
- Møller, A., & Ireland, J. (2018b). *LanguaL™ 2017—Multilingual Thesaurus (English – Czech – Danish – French – German – Italian – Portuguese – Spanish)*. Technical Report. Danish Food Informatics. Accessed at: DOI: 10.13140/RG.2.2.13274.64964 [23-4-2020].
- Zulaika, U., Gutiérrez, A., & López-de-Ipiña, D. (2018). Enhancing profile and context aware relevant food search through knowledge graphs. In *12th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2018) Proceedings*, Punta Cana, Dominican Republic, 4-7 December 2018. Accessed at: <https://doi.org/10.3390/proceedings2191228> [28/04/2020].

## Acknowledgements

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T1EDK-02015).





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Software Demonstrations**









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Software Demonstrations**

**Lexicography and Language Technologies**







# XD-AT: A Cross-Dictionary Annotation Tool

González M., Buxton C., Saurí R.

*Oxford University Press, United Kingdom*

## Abstract

Linking lexical datasets to each other is a key strategy for expanding and enriching their content with additional data from other resources. However, different resources show significant differences in the degree of granularity of the lexicographic information. Thus, while extending more coarse-grained datasets with content from fine-grained ones seems a feasible task, the other way around cannot be tackled directly. For this reason, linking datasets at the level of meaning rather than word level is essential. But also, for the same reason, word alignment at the level of meaning is a challenging task not yet solved. Within this context, we created XD-AT, a web-based annotation tool aimed to assist humans to annotate linked sense pairs across dictionaries. In this work, we focus in XD-AT's main functionalities, capabilities and potential extensions, such as reusability and adaptability. For example, although XD-AT has been implemented to classify the type of relationship between linked senses from an English monolingual dictionary and the English side of bilingual ones, XD-AT can also be extended into a more general annotation tool for marking up any type of cross-dictionary mappings at the sense level.

**Keywords:** annotation tool; sense linking; dictionary mark-up; meaning overlap

## 1 Introduction

This work presents XD-AT, a web-based annotation tool for marking up relations across dictionaries taking place at the sense level. XD-AT has been developed within the framework of Prêt-à-Llod, an EU funded project devoted to developing multilingual linked data and language technology (i.e., Linguistic Linked Open Data). The context of XD-AT within the project is particularly concerned with the exploration of methodologies for linking dictionaries at the level of meaning. This is been an area of significant activity in the past decade with the successful linking of key lexical databases, such as WordNet or Wikipedia, for language technology purposes. Examples of such success are BabelNet (Navigli & Ponzetto, 2012) and the work summarized in Gurevych, Eckle-Kohler & Matuschek (2016). Efforts in this area have more recently turned into the alignment of dictionary content due to the benefits that dictionary sense linking can contribute. Very significantly, it opens up the possibility of expanding existing lexicographic content with additional data from other sources, for example for building specialized multilingual lexicons (Schmidt 2009), or for creating new bilingual dictionaries (Gracia et al. 2019; McCrae, et al. 2017; Saurí et al. 2019).

Currently there are a number of dictionary writing systems (DWSs) that are used to generate and edit dictionary content (cf. Abel 2012). Some are quite well-known off-the-shelf proprietary solutions, such as T-LEX,<sup>1</sup> and IDM DPS<sup>2</sup>, while others have been developed within the open source paradigm, like DEBWrite (Rambousek & Horák 2015) and Lexonomy (Měchura 2017). Similarly, there already exist text annotation tools (TATs), some of which are specifically for sense tagging text, such as WebAnno (Eckart de Castilho et al. 2016), STSAnno (Batanović et al. 2018), and Ubyline (Miller et al. 2016). However, to the best of our knowledge, there is no tool among either DWSs or TATs that provides the functionality for cross-dictionary sense alignment annotations. For example, for mapping senses of different dictionaries that refer to the same meaning, or for qualifying whether aligned senses differ in any way.

XD-AT aims to supply such functionality, which involves addressing very specific challenges. Similar to DWSs, our tool had to be able to display in a clear, differentiated way, the several parts in the structure of a dictionary entry, such as definitions, example sentences, and labels. Moreover, similar to TATs, the system had to facilitate the classification of the targeted annotation elements given a closed set of classes (e.g., sense-link vs. non-sense-link). In addition, the system had to extend beyond the functionality of both DWSs and TATs in order to be able to allow for annotations on pairs of (as opposed to single) dictionary units, which entails a degree of structural (and therefore layout) complexity due to the hierarchical nature of dictionary information. Therefore, XD-AT addresses this gap among annotation tools for language resources of different kinds.

To tackle XD-AT's development, we focus on a very specific use case: to mark-up differences in sense granularity between two linked senses from different dictionaries in order to, with the resulting annotations, train a machine learning-based classifier able to determine these distinctions automatically. The annotation strategy defined by lexicographers for that task guided XD-AT functional requirements. For example, we wanted it to allow for multiple annotators on the same data in order to avoid any annotator bias and also to be able to compute inter-annotation agreement (IAA). Other requirements that were essential for us were: easy access, clear information layout, and user-friendliness. All these were taken into account in the design and deployment of the tool, together with the goal of enabling its reusability in other cross-dictionary sense annotation tasks in the future.

<sup>1</sup> <https://tshwanedje.com/tshwanelex/>

<sup>2</sup> <https://www.idmgroup.com/content-management/dps-info.html>



## 2 Sense alignment and granularity differences

This section presents the specificities of our particular annotation use case with the aim of facilitating the understanding later on, of the requirements that drove the design and deployment of the tool. As said, the goal was to obtain manual annotations on granularity differences between the two dictionary senses in a sense link. We define *sense link* as a pair of senses, each from a different dictionary, which represent the same meaning for a given *lexeme*. For lexeme we understand a combination of a lemma and a lexical category. For example, *water* (noun) and *water* (verb) are two different lexemes. By contrast, what in a dictionary may be considered as independent lexical items (e.g., homographs like *lie* (verb) ‘to not tell the truth’ and *lie* (verb) ‘to adopt or be in a horizontal position’) will be taken here as belonging to the same lexeme.

When aligning senses for a lexeme in one dictionary with the equivalent senses for the same lexeme in another dictionary, we can see that in some cases they fully align (i.e., they refer exactly to the same meaning), whereas in others one of the senses (or both) extends beyond the meaning conveyed by the other. This is illustrated in Figure 1 below, which presents the senses for lexeme *fog* (noun) from an English monolingual dictionary (left), and its translation into Spanish from a bilingual dictionary (right). As can be observed, sense 1.2 in the monolingual perfectly aligns with sense 2 in the bilingual, while sense 1 in the bilingual covers the meaning of both senses 1 and 1.1 in the monolingual.

EN Monolingual	EN-ES
<b>fog</b> <b>NOUN</b> <b>1.</b> [mass noun] A thick cloud of tiny water droplets suspended in the atmosphere (...) <b>1.1</b> [in singular] An opaque mass of particles in the air. <b>1.2</b> [Photography] Cloudiness which obscures the image on a developed negative (...) <b>2.</b> [in singular] A state or cause of perplexity or confusion.	<b>fog</b> <b>noun</b> <b>UNCOUNTABLE AND COUNTABLE</b> <b>1.</b> (Meteorology) niebla (feminine) <b>2.</b> (Photography) velo (masculine)

Figure 1: Senses for lexeme *fog* (noun) in a monolingual (left) and bilingual (right) dictionary.

Taking into account these differences in the semantic extent and overlap of two senses, we defined the four different types of meaning relationship between two linked senses illustrated in Table 1: *perfect*, *narrower-than*, *wider-than*, and *partial*.<sup>3</sup> *Perfect* match indicates that each sense aligns completely throughout the full extension of the other one. In other words, the entire meaning expressed by one sense is also expressed and covered by the other one. In contrast, *narrower-than* and *wider-than* account for sense pairs where a sense in one dictionary has a broader meaning than the sense in the other dictionary. This occurs when the meaning of one sense fully overlaps with the other one but does not fully enclose it (*narrower-than*), or the other way around (*wider-than*). Finally, *partial* is used when each of the two senses’ meaning extends beyond the reference of the other and thus, there is a part of the meaning covered by each sense that is not included in the other one.

	Perfect match	Different sense granularity		Different sense boundaries
Meaning alignment	$S_A$ $S_B$	$S_A$ $S_B$	$S_A$ $S_B$	$S_A$ $S_B$
Grounding relationships	$S_A$ fully overlaps with $S_B$ and encloses it.	$S_A$ fully overlaps with $S_B$ but does not enclose it.	$S_A$ partially overlaps with $S_B$ and encloses it.	$S_A$ partially overlaps with $S_B$ and does not enclose it.
Sense link classes	Perfect	Narrower-than	Wider-than	Partial
Symbol	=	<	>	~

Table 1: Types of sense link based on differences in sense granularity.

In order to conduct the human classification task with confidence, annotators therefore needed access to the information related to both dictionary senses as well as the possibility of selecting among the four values just presented. In addition, we wanted to meet the specifications for first-class annotation tools (e.g., user friendliness, support for managing the annotation tasks, ability to deal with multiple annotations, etc.). In the following sections, we explain the requirements we identified and how they were deployed into XD-AT.

## 3 General Framework Specifications

**Web-based access.** XD-AT is a web-based application with database support. One of the main advantages of web

<sup>3</sup> This classification has also been adopted in McCrae, ELEXIS Monolingual Word Sense Alignment Task (2020)



applications is that they allow working from any device connected to the Internet using a browser, without the need to install additional software. Although this was not a hard requirement for our experiment, we consider it would bring a valuable benefit for future re-usability and scalability of the tool, especially in the particular set-up where external contributors may participate in the annotation task.

**User roles.** XD-AT is also a multi-user application that requires authentication. The current version defines three types of user roles: *annotators*, *judges* and *managers*, and a synthetic role named *automatic*. Only managers can create new users (either other managers or annotators), handle the assignment of annotation tasks across users, inspect the annotation progress, and enable the judge review of certain assignments. In turn, annotators can only see their own assignments and progress. The judge is the user role created to resolve classification discrepancies when there are multiple annotations for the same sense link; and the automatic role is used to store classifications produced by automatic means (e.g. by an algorithm).

**Data storage.** The entire information handled by XD-AT is stored in a relational database. This covers the user credentials mentioned above, the dictionary information to be displayed, the organization of sense links collections into batches, batch assignments to annotators, all annotations and re-annotations produced by annotators, judges and automatic roles, as well as application-specific definitions, such as the list of lexical categories, polysemy degrees, and sense link classes. The latter is indeed a relevant feature in XD-AT. On the one side, it simplifies the adoption of the tool by others who can reuse the pre-defined closed set of possible annotation labels, and annotation tasks characteristics. On the other side, it still allows decoupling the application-specific characteristics from the tool implementation, which endows the application with higher adaptability to other annotation tasks that may use a different set of classes or prefer to organize the annotation tasks using different criteria.

Next sections give more details of all these functionalities, and how they are displayed in the interface.

## 4 Annotation Panel

The information to be displayed on the interface was carefully selected. It was important to be able to provide annotators with all the information needed to ground their decisions, but also that the display was as light as possible to avoid visual stress during the annotation work.

**Dictionary information.** For correctly classifying a sense link, annotators required access to all the other existing links for either sense in the targeted link. Figure 2 shows a screenshot of the annotation panel. The left side displays information for the monolingual dictionary, while the right one shows the piece for the bilingual counterpart. The area in between the two frames displays symbols for the different types of links the annotator must choose from: = (perfect match), < (narrower-than), > (wider-than), ~ (partial match), ? (donotknow), unlink (for cases that in fact are not links). Additional sense links for the senses in the targeted link are shown in two further frames at the lower part of the frame, along with their current type class (in green in between the panels).

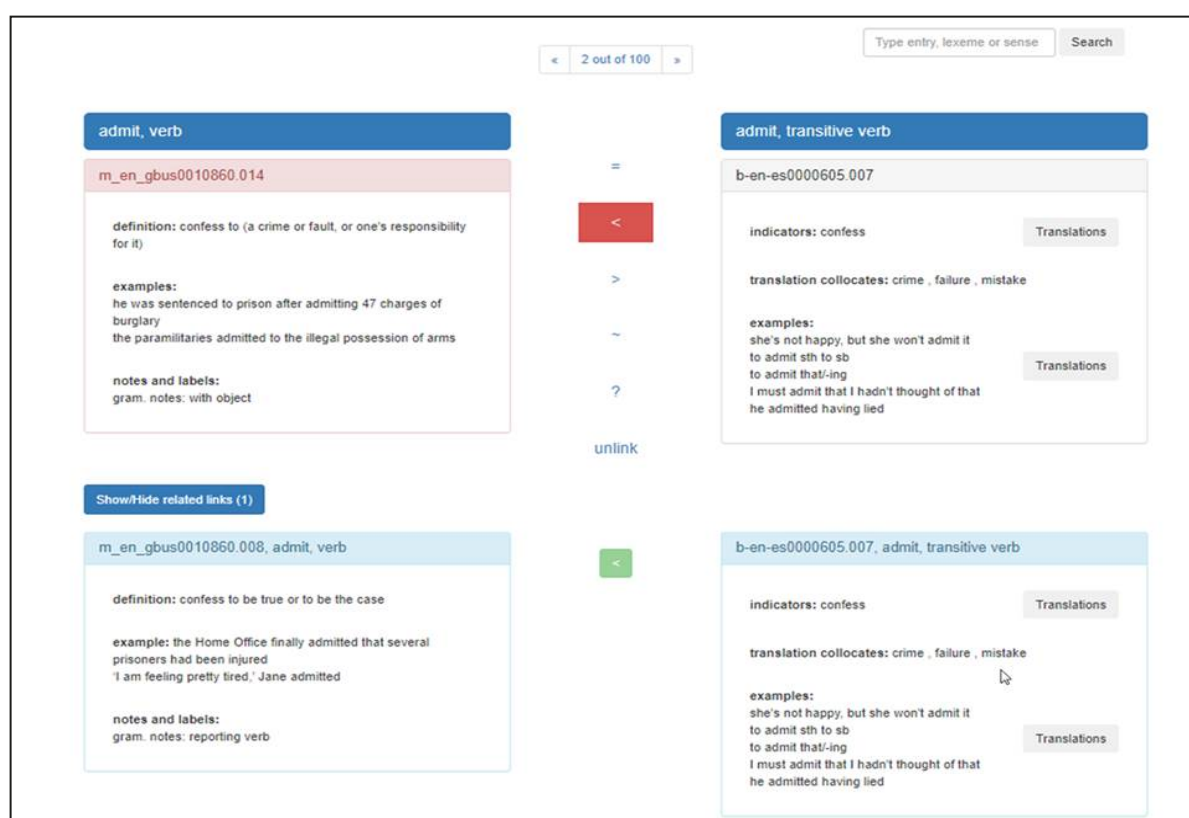


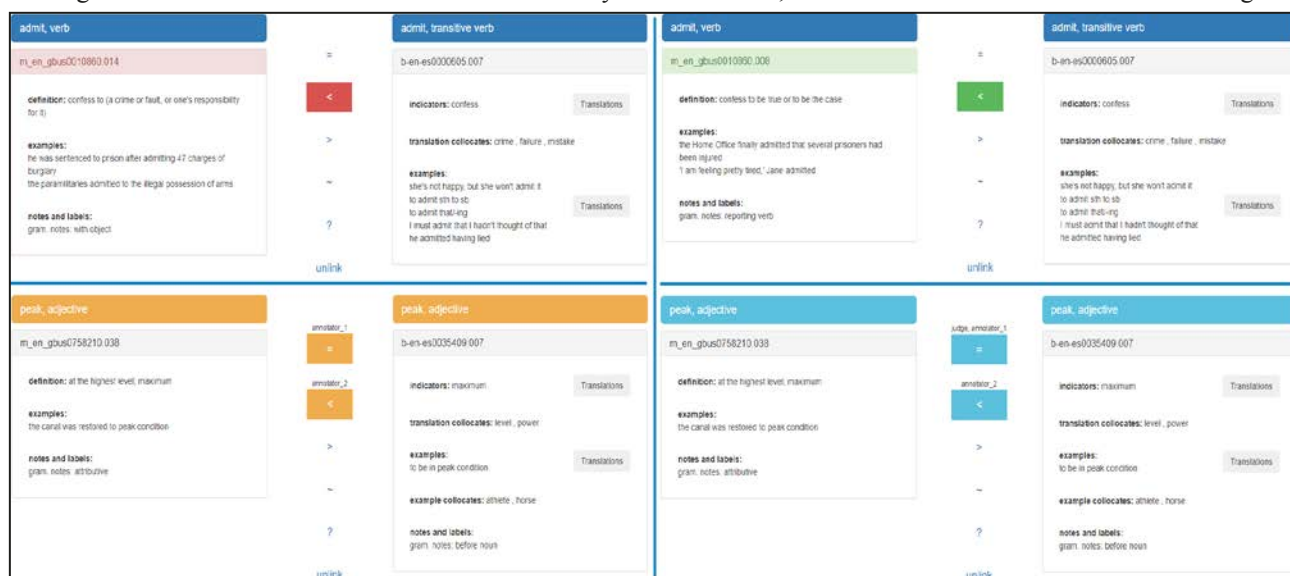


Figure 2: Screenshot of the annotation panel.

For the monolingual dictionary, the interface shows definitions, examples, grammatical<sup>4</sup> and technical notes,<sup>5</sup> word forms,<sup>6</sup> domain, region and register labels,<sup>7</sup> and domain and semantic classes from a taxonomy.<sup>8</sup> Similarly, for the bilingual dictionary, it shows indicators,<sup>9</sup> collocates,<sup>10</sup> examples, and the additional labels and notes described above. Translations were also included but hidden because annotators pointed out that they were not needed in most of the cases. Thus, the interface included a button that shows lemma's and examples' translations on demand. As a matter of fact, results from the experiments in Kouvara et al. (2020) supported the hypothesis that translations are not essential for this particular task (although they may help to discern some difficult cases). In our experiments involving bilingual dictionaries for Spanish, Chinese and Russian, none of the annotators knew either Russian nor Chinese, but some knew Spanish. However, the inter-annotators agreement shows little differences among the three datasets.

**Search functionality.** Annotators did not consider it essential to have access to the whole sense inventory for the linked lexeme in either dictionary, and therefore that information was not included as part of the default display in order to avoid visual clutter. Instead, the tool includes a search engine that allows retrieving this information for any given substring or identifier, in case annotators are interested in looking it up. The goal is to assist them by providing them access to any piece of information they may need without having to reach to the dictionary sources externally from XD-AT. Such a functionality is only possible because of having stored the dictionary data in a relational database.

**Pre-annotated labels.** To facilitate annotators work, the system was designed so that it was able to offer annotators with a pre-annotated choice (the one estimated as the most likely) for them to validate. More specifically in the context of our project, sense links were already pre-classified based on the set of heuristics (Kouvara et al. 2020), which in turn were based on the number of links held by each sense in either side of the link (i.e. in either dictionary). Hence, the annotator task consisted in correcting, or confirming, the class pre-assigned to each link. To help annotators to distinguish automatic labels from those that had already been corrected, we created the colour scheme shown in Figures



3a and 3b: red was used for automatically computed labels (via the role *automatic*), and green for manually annotated ones (given by the annotator). That way, it was visually easy to identify on which links the annotator had already taken a decision.

Figure 3: XD-AT colour scheme: a) automatic class (red), b) human label (green), c) disagreement between annotators (orange), d) reviewed by the judge (blue).

**Judge disagreements review.** XD-AT also includes a judge review mechanism to resolve disagreements between annotators and assign a final label in those cases. Given a particular batch, it finds all sense links for which there is at least one classification difference among the annotators, and shows them all in an annotation panel designed for that purpose (Figures 3c and 3d). The color scheme in this view consists of orange for displaying all annotators' choices,

<sup>4</sup> In Figure 2, grammatical notes for both senses of *admit* (verb) in the monolingual: with object and reporting verb.

<sup>5</sup> Technical notes for *sodium hydroxide* (noun): 'Chemical formula: NaOH'.

<sup>6</sup> Word forms for *world* (noun): [usually] 'the world'.

<sup>7</sup> Register labels for *think big* (idiomatic): [informal]; region label for *barbie* (noun): [chiefly Australian, New Zealand]; domain label for *arteriosclerosis* (noun): [Medicine].

<sup>8</sup> Semantic class for *arteriosclerosis* (noun): [physiological state]; and domain class: [Pathology].

<sup>9</sup> In Figure 2, *admit* (verb) has a single indicator: 'confess'.

<sup>10</sup> In Figure 2, *admit* (verb) has three collocates: 'crime, failure, mistake'.



and light blue for indicating the judge's final choice. Note also the small text on top of the annotation label buttons. These are the annotator names along with their choice. This view cannot be seen by annotators so as to avoid them influencing each other's decisions.

## 5 Annotation Task Management

XD-AT organizes annotation data into batches, which are sets of annotation units (e.g., here, sense links) of a specific length. The notion of batch has been a great enabler for better organizing and managing the annotation task. In our case, batches consisted of 100 sense links each.

**Batch creation.** An important feature in XD-AT is the automatic creation of batches of data to annotate based on carefully selected criteria. Batches are sets of annotation units (here, sense links). For our case in particular, the full collection of links was split into lexeme subsets according to their lexical category (noun, verb, adjective, adverb/preposition, or other) and polysemy degree (single-sense, small, medium, or large size). The rationale behind that is that senses for certain lexical categories and, most likely, polysemy degrees may be harder to annotate than others. Thus, the aim with the batch creation functionality is to help annotators focus on similar annotation cases at a time.

**Batch assignment.** Once batches are created, a manager user needs to assign them to annotators. Ideally, batches could be evenly distributed and automatically assigned among annotators, so that each annotator is assigned a similar number of batches of each type. However, we kept this as a manual operation so that the manager can adapt assignments to the annotators skills to increase the quality of the results.

Finally, XD-AT features two further functionalities to manage and monitor the annotation task progress: assignment of multiple annotations and annotation history track. They are presented next.

**Multiple annotation assignment.** Batches can be assigned to several annotators at once in order to obtain multiple annotations on the same data. This is a key feature because it supports several functions: firstly, it makes it possible to calculate IAA as an estimate on the difficulty of the task; secondly, it allows us to identify areas of major disagreement among annotators (and therefore presumed data complexity) that can then inform the development of well-grounded annotation guidelines; and last but not least, it helps pinpoint annotators that tend to disagree with others more often, a piece of information that can then be used to adjust batch assignment in order to ensure highest data accuracy.

**Annotation history track.** Finally, XD-AT stores all re-annotations over each sense link; that is, all the labels that the same annotator may have assigned to a sense link at different moments in time due to second thoughts or hesitation about that case. The re-annotation history is useful to analyze data complexity and task difficulty. Among other things, it can help identify common lexicographic features among annotation units that create more difficulties, or pinpoint particular batches that are more challenging than others.

## 6 Exporting Annotated Data

XD-AT is able to export the annotations in a machine-readable format (CSV and JSON) according to three use cases:

- **Baseline:** These are the labels created by automatic roles. In our downstream experiments, we use these to compare the accuracy of different machine learning models trained on the manual annotations, hence the name.
- **Shared annotations:** This is the collection of labels assigned by all annotators to sense links in the shared batches (i.e., the batches adjudicated to multiple annotators). This subset is used for, e.g., computing IAA or identifying areas of major disagreement. It does not include judge labels.
- **Gold standard:** This includes the final classification labels for all the batches in the dataset. In the case of shared batches, it takes the judge's decisions in case of disagreement among annotators. This subset discards all links classified as unlink or uncertain.

In our specific annotation task, the export output contains the following information fields: 1) sense link type, which is the final label only; 2) polysemy degree of the lexeme that the linked senses belong to; 3) lexical category of that same lexeme; 4) dictionary to which each of the linked senses belongs; 5) annotation batch ID; 6) annotation timestamp; and 7) annotator ID. The goal is to provide as much information as possible along with annotations, so that it can be used in downstream tasks.

Finally, XD-AT also exports the list of links that were re-annotated, grouped by annotator and batch ID. This information, together with the list of links labelled as donotknow, can be used to further analyse the complexity of the task, discern patterns in difficult cases, and also guide enhancements of XD-AT in future revisions.

## 7 Final Remarks

XD-AT was developed for annotating distinctions of sense granularity between dictionary senses that refer to the same meaning (that is, that are already aligned). However, it can be upscaled into a more general tool for also marking up any type of cross-dictionary alignments and relations at the sense level.

We are not aware of any other tool developed so far for that purpose. Possible extensions include:

- Improving user management functionalities, e.g. use of other authentication methods, ability to enable/disable



users, assignment of multiple roles to the same user, etc.

- Facilitating the analysis of IAA scores online. For example, by adding interface areas for inspecting the inter-annotator agreement results in an interactive way, by selecting the batches (or users) among which to compare annotations.
- Improving export functionalities by, e.g., including interface areas for selecting batches or dictionaries subsets to export, or for sampling based on users or other fields, in addition to the classification labels.
- Publishing the tool publicly. To date, XD-AT is for internal use only since the above extensions are work in progress. Nonetheless, we are open to receive requests for using the tool and suggestions for making XD-AT a more flexible tool able to embrace other use cases.

## 8 References

- Abel, A. (2012). Dictionary Writing Systems and Beyond. A S. Granger, & M. Paquot (Ed.), *Electronic Lexicography* (p. 83–106.). Oxford, United Kingdom: Oxford University Press.
- Batanović, V., Cvetanović, M., & Nikolić, B. (2018). Fine-grained Semantic Textual Similarity for Serbian. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)* (p. 1370-1378). Miyazaki, Japan: ELRA.
- Eckart de Castilho, R., Mújdricza-Maydt, É., Muhie Yimam, S., Hartmann, S., Gurevych, I., Frank, A., & Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (p. 76-84). Osaka, Japan: The COLING 2016 Organizing Committee.
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382.
- Fleiss, J., Levin, B., & Cho Paik, M. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Hoboken, New Jersey: Wiley.
- Gracia, J., Kabashi, B., Kernerman, I., Lanau-Coronas, M., & Lonke, D. (2019). Results of the Translation Inference Across Dictionaries 2019 Shared Task. *Proceedings of TIAD-2019 Shared Task*. 2493, p. 1-12. Leipzig, Germany: CEUR Workshop Proceedings.
- Gurevych, I., Eckle-Köhler, J., & Matuschek, M. (2016). *Linked Lexical Knowledge Bases: Foundations and Applications*. Morgan & Claypool.
- Kouvara, E., González, M., Grosse, J., Sauri, R. (2020). Determining Differences of Granularity between Cross-Dictionary Linked Senses. *Congress of the European Association for Lexicography (Euralex 2020)*. Alexandroupolis, Greece.
- McCrae, J. (May / 2020). *ELEXIS Monolingual Word Sense Alignment Task*. CodaLab Competition: <https://competitions.codalab.org/competitions/22163>
- McCrae, J., Bond, F., Buitelaar, P., Cimiano, P., Declerck, T., Gracia, J., Piasecki, M. (2017). 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets. co-located with 1st Conference on Language, Data and Knowledge (LDK 2017). *Proceedings of the LDK 2017 Workshops*. 1899. Galway, Ireland: CEUR Workshop Proceedings.
- Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. A I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Ed.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, (p. 662-679). Leiden: Lexical Computing.
- Miller, T., Khemakhem, M., Eckart de Castilho, R., & Gurevych, I. (2016). Sense-annotating a Lexical Substitution Data Set with Ubyline. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 828–835). Portorož, Slovenia: European Language Resources Association (ELRA).
- Navigli, R., & Ponzetto, S. (December / 2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217-250.
- Rambousek, A., & Horák, A. (2015). DEBWrite: Free Customizable Web-based Dictionary Writing System. *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 Conference*. (p. 443-451). Herstmonceux Castle: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd.
- Sauri, R., Mahon, L., Russo, I., & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level. *42nd Conference on Very Important Topics (CVIT 2016)*, 15.
- Schmidt, T. (2009). The Kicktionary – A Multilingual Lexical Resource of Football Language. A H. Boas (Ed.), *Multilingual FrameNets in computational lexicography : methods and applications* (p. 101-132). Berlin, Germany: de Gruyter.

## Acknowledgements

This work has been funded by the H2020 project “Prêt-à-LLOD: Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors” under grant agreement No 825182. Also, we are very grateful to Ashleigh Alderslade, Matthew Bladen, Emma Davies, Janet Gough, Denny Hilton, Eleanor Maier, Iona Ogilvie, Nick Rolfe, and Catherine Sangster, the expert lexicographers who contributed all the editorial knowledge we were lacking, and also helped to test the first version of the tool and annotate the data for our experiments. In addition, we would like to thank Eva Theodoridou and Anna Emberton for their support in managing and planning the work required for the project. The authors are responsible for any errors and problems.



# Augmented Writing and Lexicography: A Symbiotic Relationship?

Køhler Simonsen H.

Copenhagen Business School, Denmark

## Abstract

We live in an age of disruption and technological innovations, and lexicography as a scientific discipline and practice is witnessing a fundamental paradigm shift, cf. also (Fuertes-Olivera 2016), who talks about a “Cambrian Explosion”, (Simonsen 2016), who discusses the need for a new “Lexicographic Business Model” and (Tarp 2019), who refers to the paradigm shift in lexicography as “Tradition and Disruption in Lexicography”. Like many other disciplines, lexicography is operating within the framework of the “Fourth Industrial Revolution”, cf. (Schwab 2015), and it seems to be facing many fundamental challenges.

One of these challenges is Augmented Writing (AW), cf. (Banks 2019; G2.com 2019; Marconi 2017 and Simonsen 2020a, 2020b), who discuss AW and how it affects journalism, communication and lexicography respectively.

The objective of this article is to discuss AW from a lexicographical perspective and to what extent the two disciplines may form a value-adding symbiotic relationship. Based on empirical data from a test of 32 AW technologies, the article discusses this question and presents a number of theoretical considerations on how AW and lexicography might develop a symbiotic relationship drawing on Colson (2019), Fadel et al. (2017), Liew (2013), Tarp (2019), and Simonsen (2020a, 2020b).

**Keywords:** Augmented Writing; Writing Assistants; Lexicographically Augmented Writing

## 1 Introduction

It is always dangerous to make predictions, especially when it comes to the impact of technology. Even the quite famous corporate turnaround expert, Jim Keyes, the then CEO of Blockbuster, got it very wrong when he predicted, “Neither RedBox nor Netflix are even on the radar screen in terms of competition” (Rapier 2020). He was very wrong. As we all know, Blockbuster went bankrupt only two years later.

Some would no doubt argue that this has nothing to do with lexicography. Others would argue that similar disruptive developments are already taking place in lexicography. One thing is certain. We can all learn from history.

One example of direct relevance for this article is *Write Assistant*, which has almost outcompeted virtually all established and renowned dictionary publishers in Denmark. Admittedly, this is just one example and one small country, but the adoption curve of disruptive technology is almost exponential and very much international. Consequently, there is an imminent need for discussing AW and the role it may have in lexicography.

Fortunately, lexicography is a strong science and discipline, and it has helped people understand, communicate and learn for thousands of years and it has much to offer. This article discusses how AW and lexicography can form a symbiotic relationship.

## 2 Research Question, Method, Data and Delimitations

The underlying research question of this paper is to answer the overall question: How can AW and lexicography form a symbiotic relationship?

The article draws on empirical insights from a structured test of 32 different AW technologies, (see also Simonsen 2020a; 2020b for a detailed discussion of the 32 AW technologies). The structured test and analysis of the AW technologies focused on parameters such as task types, degree of autonomy, workspace integration and lexicographic augmentation potential.

The analysis and discussion in this paper are delimited to AW technologies supporting text production and text analysis (sentiment analysis).

## 3 Literature Review

Computer-Assisted Language Learning (CALL) has no doubt played an important role in the development of AW. A particularly relevant contribution of CALL applications and dictionaries of relevance is Abel (2009:5), who states that it



is crucial to categorize CALL applications based on their “central element and/or the starting point”. A similar categorization is used here. The central element and/or starting point of most AWs is AI and most AWs aim at providing automatic lexical error correction and text production. CALL applications typically have a dictionary as its central element or language learning as its primary purpose. AW technologies thus seem to differ from CALL applications.

For the past 50 years publishing houses, computer linguists and lexicographers have developed a large number of language technological solutions, which have largely led to increased efficiency for translators and communicators. One landmark development started already in the 1970s when researchers discussed the possibility of translators using segments of already translated texts (Kay 1997). This led to the development of translation memory systems and *Sdltrados* was one of the first TM systems in use. Today, translators and professional text producers primarily use web-based systems like *Sdltrados* or *Wordfast*. This development to some extent also plays a role in modern AW.

Another landmark development was the many language technology solutions developed by computer linguists, IT experts and lexicographers. L2 writing research has been central to lexicography and language technology for the past 50 years, and recent research seems to focus on computer-supported collaborative writing, which in many ways is something much more advanced than the single-user AW technologies analysed here (Strobl 2014). Other contributions on L2 writing research and computer-supported collaborative writing are discussed by Arnold et al. (2009), De la Colina and García Mayo (2007), Elola and Oskoz (2010), Kessler et al. (2012), Kost (2011), and Storch (2005), to mention just a few.

Other landmark developments, which may have served as inspiration for many AW technologies, are based on computer-based writing instructions for text producers and learners (Allen et al. 2016) and the tool Writing Aid Dutch, which offers students process-oriented writing support (De Wachter et al. 2014). Furthermore, Frankenberg-Garcia et al. (2019) discuss a writing assistant, which is designed to help EAP writers with collocations, and Wanner et al. (2013) published a seminal discussion of writing assistants and automatic lexical error detection.

However, the above writing aids or writing assistants are not based on AI, and AW is widely different from many existing CALL applications and other types of language technological solutions because AW to a very high degree is based on AI and very often do not even use lexicographical data as the “central element and/or the starting point” (Abel 2009).

Recent landmark developments include Granger and Paquot (2015), who outline theoretical blueprints of a needs-driven online academic writing aid, Strobl et al. (2019), who offer a very useful review of different technologies for digital support for academic writing and, of course, the very relevant contributions by Tarp et al. (2017) and Tarp (2019), who discuss new challenges in lexicography based on the L2 writing assistant *Write Assistant* referred to at the beginning of this article.

Consequently, we need to develop theoretical considerations on lexicography and AW – because AW seems to need lexicography. However, before I do that, it is time to reflect on the insights from the empirical data.

## 4 Analysis and discussion

The analysis of the 32 AW technologies combined with the literature review of relevant theoretical contributions led to three overall findings. The first important finding based on the test of the 32 AW technologies made it possible to create an overall typology. It was found that the surveyed AW services can be divided into five overall groups.

**Group 1:** Spelling and grammar checkers such as Grammarly or WhiteSmoke. This category of tools is most often fully workspace-integrated and helps the user with automatic spelling and grammar recommendations. As shown below in Figure 1, Grammarly also includes an automatic tone of voice detector in addition to its grammar checker.

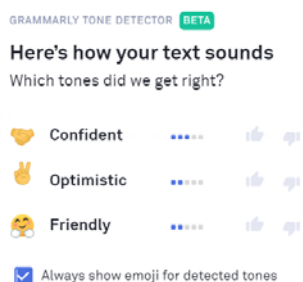


Figure 1: Grammarly.



**Group 2:** Text production robots such as TalktoTransformer or Articoolo. This category of tools is most often only browser-based and helps the user by autonomously producing texts based on just a few keywords. As shown below in Figure 2, TalktoTransformer automatically creates a text with just a few words using the GPT-2 Natural Language Understanding model. As will appear TalktoTransformer starts satisfactorily, but then the AI goes seriously astray.

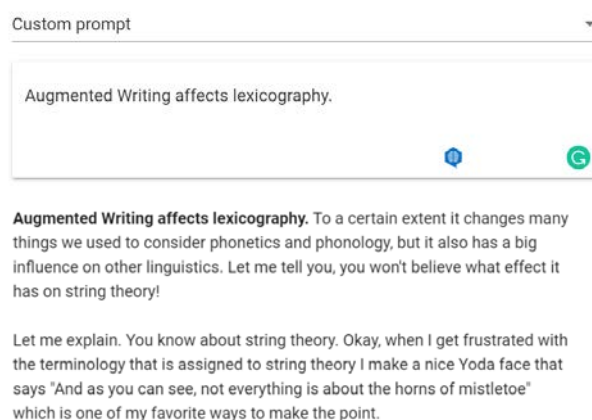


Figure 2: TalktoTransformer.

**Group 3:** L2 writing assistants such as Text Assistant. This type of tool is most often fully workspace-integrated and helps the user with context-aware recommendations in connection with L2 translation and L2 text production. The example in Figure 3 shows how Write Assistant predicts the next English word. Write Assistant is not AI-based and merely predicts the next word based on a language model and a 1:1 terminological relationship.

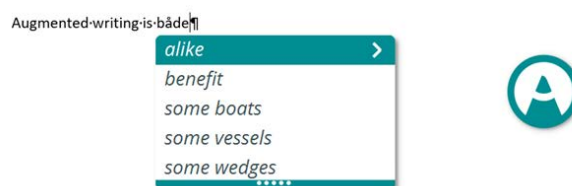


Figure 3: Write Assistant.

**Group 4:** Stylistic and tone of voice checkers such as Persado or MessagePath. This type of tool is most often workspace-integrated, particularly browser-based, and helps the user with stylistic and/or tone analysis of specific texts, for example, sales or marketing texts. Figure 4 below shows how the content and tone of voice analysis works in MessagePath.

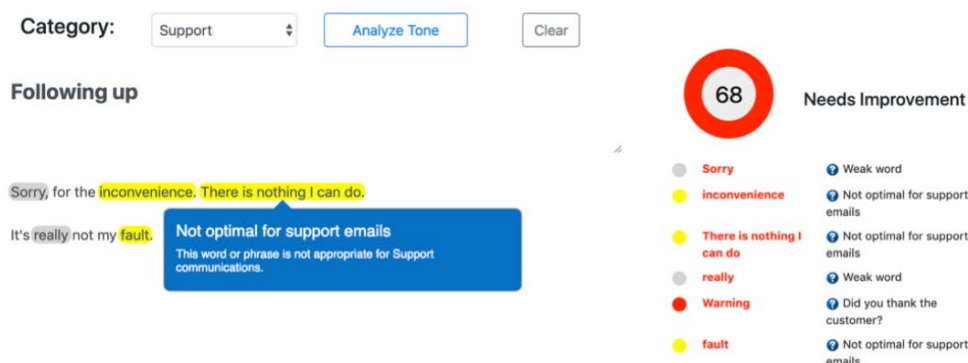


Figure 4: MessagePath.

**Group 5:** Special-purpose language pattern assistants such as Textio. This type of tool is most often browser-based and helps, for example, HR departments screening texts from candidates and producing job ads with the right sound. It also helps companies around the world produce insightful and inclusive texts based on data on age and gender bias.



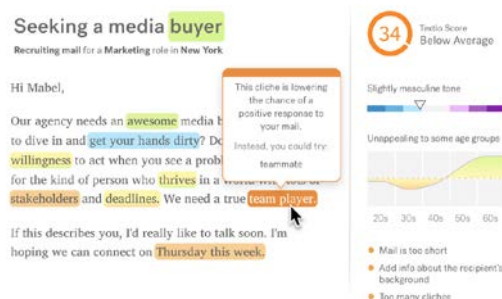


Figure 5: Textio.

Finally, the structured test and analysis of the AW solutions also revealed that many news agencies have already implemented special-purpose robot journalists designed to extract and produce specific news articles, for example, financial news, soccer news or football match reports. The robot journalist tools are designed to extract data from existing news media and produce specific news articles based on text templates. In other words, AW also plays an increasing role in the news industry.

The second important finding from the structured test is that the technological maturity of many AW technologies is very high and they already seem to be a major challenge to many conventional lexicographic services such as spellchecking and grammar dictionaries. Tarp et al. (2017; 2019) reached a similar conclusion that L2 writing assistants and context-aware dictionaries seem to have much to offer to producers of L1 and L2 texts. AW really seems to challenge the type of lexicographical products, which focus exclusively on the delivery of data and information. This argument is already seen in Simonsen (2020a; 2020b), who argues that we may have to “turn lexicography upside down” dividing specific tasks between man and machine. Figure 5 below shows Liew’s DIKIW model (Liew 2013) with my additions (dotted lines and vertical text).

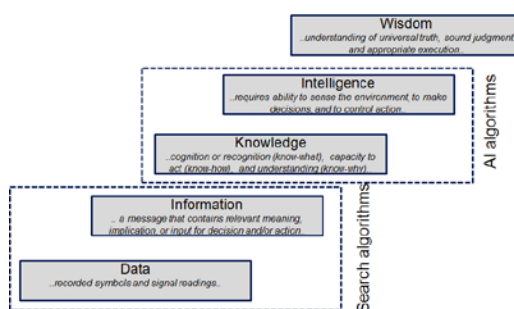


Figure 6: The DIKIW Model (my additions).

It is argued that lexicographical products, which solely focus on the first two levels in Liew’s model (delivering data and information), are already being replaced by powerful search algorithms. Many people do not look up words in a dictionary anymore. They merely double-click and then right-click on a word in MS Word and perform a smart search on Bing and/or Google, cf. also de Schryver (2012:130), who observed this experimentally. Furthermore, the structured test of the 32 different AW technologies also showed that the next two levels in Liew’s model (providing knowledge and intelligence) are increasingly being challenged by AI and even though existing autonomous AW solutions still leave much to be desired when it comes to quality and relevance, they are improving exponentially.

The third overall finding from the structured test was that AW platforms are moving into the lexicographical arena. This may have dramatic consequences for dictionaries providing users with knowledge and understanding as they may be in danger of being disrupted or replaced by AI (Simonsen 2020a; 2020b). AW solutions based on strong AI may very well become the next big disruptor in lexicography because the development of these AI technologies has priority in many countries. Their ease of use, ubiquity and degree of integration make them interesting for many users.

However, the test also revealed that the quality of the autonomous AW solutions such as TalktoTransformer leaves much to be desired. Most AW solutions will first try to understand the context of the input you feed into them using AI algorithms. Then they will locate the best text resources available and reconstruct it all to one coherent text through language models or NLP engines. So AW solutions are not necessarily based on curated data, but language models. To sum up, the test showed that AW needs curated lexicographical data, world knowledge and relational knowledge and thus needs to form a relationship with lexicography.

The output quality of many AW technologies can be improved significantly using curated lexicographical data. These lexicographical data should be available in special corpora and used when the AW attempts to locate the best text resources available. In other words, lexicography can help AW with curated lexicographic data and thus significantly



improve the output quality.

The output quality of many AW technologies can also be improved by adding world knowledge and relational knowledge to the actual output of the AW. Most AW technologies do not sufficiently understand context and lexicography might provide both world knowledge and relational knowledge (Simonsen 2020a; 2020b). In other words, lexicography can also help AW as condensation and description of world knowledge is central to lexicography.

Providing world knowledge and relational knowledge is not an easy task, but AW technologies could be equipped with an auxiliary post-editing window providing as much help as possible to the user when post-editing the output text. Similar arguments are found in Leroyer and Simonsen (2019), who have developed a framework for providing help to users when post-editing professional texts.

The suggested framework for the lexicographical augmentation of AW technologies takes its starting point in the division of labour between man and machine (Colson 2019), the layered understanding of data, information, knowledge and intelligence (Liew 2013) and last but not least, the idea of providing access to specially selected lexicographical data in the post-editing phases (Leroyer & Simonsen 2019). The suggested framework is based on OpenAI's Natural Language Understanding (NLU) model, which was trained to perform a single task of predicting the next word with a given set of words and a very large dataset (Rodriquez 2019). The GPT-2 model is used in TalktoTransformer and it does yield amazing output based on just a few words as it was demonstrated in Figure 2 above.

I argue that the symbiotic relationship between AW and lexicography could be consummated by building an AW where curated lexicographical data are simply part of the first priority training datasets. This would significantly improve the output quality of the AW.

When it comes to improving the output of AW technologies with world knowledge and relational knowledge it is much more complex. It is not about just inserting yet another fine-tuning layer in OpenAI's GPT-2 language model (OpenAI 2019). Human augmentation and intervention are needed. I argue that an external post-editing window is needed because human augmentation is required when adding world knowledge or relational knowledge in line with (Colson 2019), who makes the case for the division of labour between man and machine. This external post-editing window could be the final step in the output process of a lexicographically augmented AW technology. In other words, lexicographically augmented AW technologies might be what we need.

## 5 Conclusion

Building on Colson (2019), Banks (2019), Liew (2013), Marconi (2017), Simonsen (2020a, 2020b), Tarp et al. (2017) and Tarp (2019) and the empirical analysis, this article offered a discussion of selected AW services.

Based on the structured test it was first possible to develop an overall typology of AWs, which were categorized in five overall groups. The second finding based on the structured test was that the technological maturity of most AW technologies is very high. The third finding was that the output quality of most AW technologies leaves much to be desired and that what is needed is curated lexicographical data and world knowledge and relational knowledge.

The analysis and discussion of the 32 AW technologies also revealed that AW is or may develop into a major challenge to many conventional lexicographic services offering only data and information (Liew 2013). The discussion also revealed the weaknesses and lacking quality of some AW technologies and the discussion uncovered many considerations on how lexicography can augment AW or even form a symbiotic relationship with AW.

Lexicography has an important role to play in the development of new advanced text production technologies and the lexicographical augmentation of AW could be an important step in the right direction. In conclusion, lexicography has much to offer to AW especially when it comes to human augmentation of the automatic output from an AW service.

## 6 References

- Abel, A. (2009). Towards a systematic classification framework for dictionaries and CALL. In S. Granger and M. Paquot (eds), *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009*, Louvain-la-Neuve, 22-24 October 2009, 3-11.
- Allen, L., Jacovina, M. & McNamara, D. (2016). Computer-based writing instruction. In C.A. MacArthur, S. Graham & J. Fitzgerald (eds.) *Handbook of writing research*. New York, NY: Guilford, pp. 316-329.
- Arnold, N., Ducate, L., & Kost, C. (2009). Collaborative writing in wikis: Insights from culture projects in German classes. In L. Lomicka & G. Lord (Eds.), *The next generation: Social networking and online collaboration in foreign language learning* (pp. 115-144). San Marcos, TX: CALICO.
- Banks, C. (2019). What is an Augmented Writing Platform? Accessed at: <https://medium.com/swlh/what-is-an-augmented-writing-platform-b28fa588a1c5> [19/04/2020].



- Colson, E. (2019). What AI-Driven Decision Making Looks Like. Accessed at: <https://hbr.org/2019/07/what-ai-driven-decision-making-looks-like>. [19/04/2020].
- de Schryver, Gilles-Maurice (2012). Lexicography in the crystal ball: Facts, trends and outlook. In: Fjeld, Ruth V. & Julie M. Torjusen (eds). *Proceedings of the 15th EURALEX International Congress*, 7-11 August, 2012, Oslo: 93–163. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
- De la Colina, A. A., & García Mayo, M. d. P. (2007). Attention to form across collaborative tasks by low-proficiency learners in an EFL setting. In M. d. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 91-116). Clevedon: Multilingual Matters Ltd.
- De Wachter, L., Verlinde, S., D'Hertefelt, M., Peeters, G., Tounsi, L. (2014): How to deal with students' writing problems? Process-oriented writing support with the digital Writing Aid Dutch. In Rak, Rafal (Editor) The 25th International Conference on Computational Linguistics, Date: 2014/08/23 - 2014/08/29, Location: Dublin. *Proceedings of the Conference. System Demonstrations*; 2014; pp. 20 – 25.
- Elola, I., & Oskoz, A. (2010). Collaborative writing: Fostering foreign language and writing conventions development. *Language Learning and Technology*, 14(3), 51-71.
- Fadel, C., Bialik, M. & Trilling, B. (2017). Fire-dimensional uddannelse: Kompetencer til at lykkes i det 21. århundrede. 1. udgave, 1. oplag. Dafolo.
- Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), pp. 23-39.
- Fuertes-Olivera, P.A. (2016). A Cambrian Explosion in Lexicography: Some Reflections for Designing and Constructing Specialised Online Dictionaries. In *International Journal of Lexicography* 29(2): 226-247.
- Granger, S. & Paquot, M. (2015). Electronic lexicography goes local: Designs and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), pp. 118–141.
- G2.com (2019). Best AI Writing Assistant Software. Accessed at: <https://www.g2.com/categories/ai-writing-assistant> [19/04/2020].
- Kay, M. (1997). The Proper Place of Men and Machines in Language Translation. In *Machine Translation*. 12 (1–2): 3–23.
- Kessler, G., Bikowski, D., & Boggs, J. (2012). Collaborative writing among second language learners in academic web-based projects. In *Language Learning & Technology*, 16(1), 91-109.
- Kost, C. (2011). Investigating writing strategies and revision behaviour in collaborative wiki projects. In *CALICO Journal*, 28(3), 606-620.
- Leroy, P. & Simonsen, H. K. (2019). Google Translate som trussel eller redning for oversættelsesordbøger. In *LexicoNordica* 26, 2019.
- Liew, A. (2013). DIKIW: Data, Information, Knowledge, Intelligence, Wisdom and their Interrelationships. In *Business Management Dynamics*. Vol. 2, Issue 10, April 2013, 49-62.
- Marconi, F. (2017). NiemanLab. Predictions for Journalism (2017): The Year of Augmented Writing. Accessed at: <https://www.niemanlab.org/2016/12/the-year-of-augmented-writing> [19/04/2020].
- OpenAi. Accessed at: <https://openai.com/> [19/04/2020]
- Rapier, G (2020). 13 Quotes From Bosses Who Mocked Technology and Got It (Very) Wrong. Accessed at: <https://www.inc.com/business-insider/boss-doesnt-understand-technology-mocks-trend-wrong.html> [19/04/2020].
- Rodriguez, J. (2019). One Language Model to Rule Them All. Accessed at: <https://towardsdatascience.com/one-language-model-to-rule-them-all-26f802c90660> [19/04/2019]
- Schwab, K. (2015). The Fourth Industrial Revolution: What It Means and How to Respond. Foreign Affairs. Accessed at: [www.foreignaffairs.com/articles/2015-12-12/fourth-industrialrevolution](http://www.foreignaffairs.com/articles/2015-12-12/fourth-industrialrevolution) [19/04/2020].
- Sdltrados. Accessed at: <https://www.sdl.com/software-and-services/translation-software/sdl-trados-studio/> [19/04/2020].
- Simonsen, H. K. (2016). Hvor er forretningsmodellen? *En analyse af de forretningsmæssige udfordringer i forlags- og informationsindustrien med særlig fokus på opslagsværker. MBA-afhandling*. Institut for Økonomi og Ledelse. Aalborg Universitet.
- Simonsen, H. K. (2020a). Augmented Writing: nye muligheder og nye teorier. In *Nordiske Studier i Leksikografi* 15, 2019, Rapport fra 15. Konference om Leksikografi i Norden – Finland 4. juni–7. juni 2019 (In press).
- Simonsen, H. K. (2020b). Når Augmented Writing og leksikografi går hånd i hånd. In: *LEDA-nyt nr. 69* - april 2020, 3-13.
- Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, 14(3), 153-173.
- Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A. & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. In *Computers & Education*, 131, pp. 33-48.
- Tarp, S. (2019). Connecting the Dots: Tradition and Disruption in Lexicography. In *Lexikos* 29, 224-249.
- Tarp, S., Fisker, K., & Sepstrup, P. (2017). L2 writing assistants and context-aware dictionaries: New challenges to lexicography. In *Lexikos* 27(1), 494-521.
- Wanner, L., Verlinde, S. & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. In: Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M. & Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, 17-19 October 2013, Tallinn, Estonia, pp. 472-487.
- Wordfast. Accessed at: <https://www.wordfast.com/> [19/04/2020].
- Write Assistant. Accessed at: <https://www.writeassistant.com/da/> [19/04/2020].



# IdeoMania and Gamification add-ons for App Dictionaries

Caruso V., Monti J., Andrisani A., Beatrice B., Contento F., De Tommaso Z., Ferrara F., Menniti A.

University of Naples 'L'Orientale', Italy

## Abstract

The paper outlines the main features of a lexicographic mobile game developed by a group of students during a coding course. The app is a learning resource for Italian idioms based on pictorial strategies and theoretical assumptions from phraseological research. Special attention is also given to the pedagogical methodology (*Challenge Based Learning* or *CBL*) used during the course and its specific improvements for supporting Humanities students in learning coding topics. Learners actually work in groups to develop an app project on a topic they feel passionate about, so they are urged to acquire coding and specific content knowledge on the subject of their interest.

*Challenge Based Learning* is intended to be a flexible learning framework which can be used to improve students' knowledge in the electronic lexicography field due to an ongoing process of reflecting, researching and testing innovative solutions before releasing the final prototype. As an example, the type of gamification elements provided by *IdeoMania* can be added to any kind of lexicographic tool.

**Keywords:** Lexicographic Apps; Lexicography Learning Programs; Lexicography and Language Technologies; Phraseology and Collocation; Reports on Lexicographical and Lexicological Projects

## 1 Introduction

This paper reports on the main features and functionalities of *IdeoMania*, a mobile application for learning Italian idioms based on pictorial strategies and theoretical assumptions from phraseological research. Another topic of interest lies in the methodological framework used for its development, which can be used more extensively in lexicography since it offers learning strategies for improving lexicographic knowledge and skills among students.

*IdeoMania* has been developed during the four-weeks of *L'Orientale Apple Foundation* (or *LOR Foundation*<sup>1</sup>), a course in programming and app development for Humanities students held at the University of Naples 'L'Orientale' (Monti & Caruso 2019). A special pedagogical method, called *Challenge Based Learning* (or *CBL*, Nichols et al. 2016), supports beginners in coding with learning the necessary IT skills for releasing an app prototype at the end of the course hours. The learning strategy relies on students' involvement in investigating an issue by which they feel challenged and thus willing to offer solutions with an app tool.

## 2 Gamification and Lexicography

*IdeoMania* is based on gamification principles with the aim to develop an effective and funny tool to engage people in learning a foreign language, with special reference to its idioms. Gamification is generally understood as the use of game elements in situations which are not considered as a game or don't have game-like features (Deterding et al. 2011: 10). The effect of educational games on language learning, together with their impact on affective learning outcomes and knowledge has been widely analyzed (Hung et al. 2018). Only very recently digital game-based learning has been used in the field of e-lexicography (Mihaljević 2019): an analysis of existing educational games and their gamification elements on lexicographic sites shows that only a small percentage of online dictionaries employs gamification to engage its users. Besides, the use of gamification in dictionary apps as an effective support in vocabulary learning is still an under-researched field, although apps are used by an increasing number of language learners as is demonstrated by the success of Memrise or Duolingo apps. Dehghanzadeh and colleagues (2019) provide a systematic review of 22 contributions dating from 2008 through 2019 where positive effects of gamification were reported on learners' learning experiences and their learning outcomes.

## 3 *IdeoMania*: a Game for Learning Idioms

*IdeoMania* is the beta version of a stand-alone application which will also be available as an add-on to a mobile idiom dictionary for foreign speakers (Caruso et al. 2019). The game uses lexicographic descriptions contained in the dictionary entries which, in turn, will provide direct access to the game activities of *IdeoMania*. The software has been localized in English therefore other language versions can be released promptly and make idiom learning more accessible to less proficient speakers of Italian.

<sup>1</sup> For further information on the course and images on the hig-tech classroom: <https://lorientalefoundationprogram.wordpress.com>.



### 3.1 The *Idiomoji* Game Mode

Two different quiz modes are provided by the tool: one for learning new idioms and one for doing exercises by means of a recall activity. The first game is called *Idiomoji* and uses emoji to depict the literal meaning of idioms, the second is *Sfidioms* (from ‘sfida’ the Italian word for ‘challenge’) and is an interactive game to challenge friends on idioms.

The *Idiomoji* home page view is made up of face-down idiom cards. By tapping on one, players discover an idiom (see fig. 1) and try to guess its literal meaning using the emojis better suited to this aim among a selection provided on the screen. For example, fig. 1 shows the translation into emojis of “mettere la mano sul fuoco” (lit. “to put the hand on a fire”, figurative meaning “to be sure about a statement or piece of information because it is definitely true”), consisting in a pictorial representation of the four content words of the idiom, as is also required by the system (“component words to be guessed: 4”).

This game has been inspired by *Pinocchio in emoji italiano* (Chiusaroli et al. 2017), an experiment of translation of the famous Italian novel into emoji carried out on Twitter thanks to the *emoji italiano* bot on Telegram (Monti et al. 2016), which has proved the interest of app users in translation activities based on emojis.

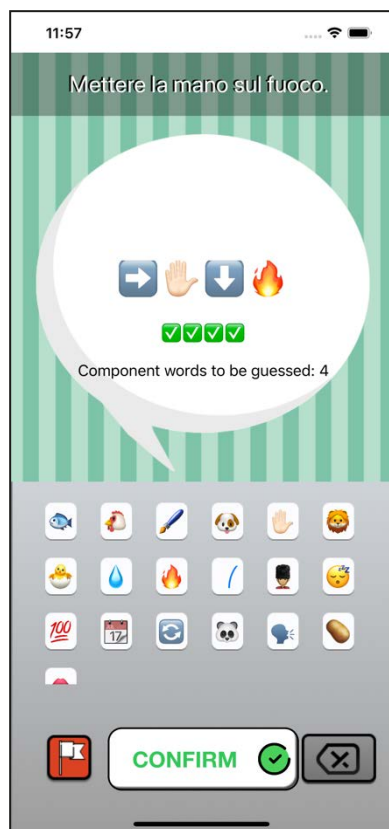


Fig. 1: Translation into emojis of the idiom “mettere la mano sul fuoco”

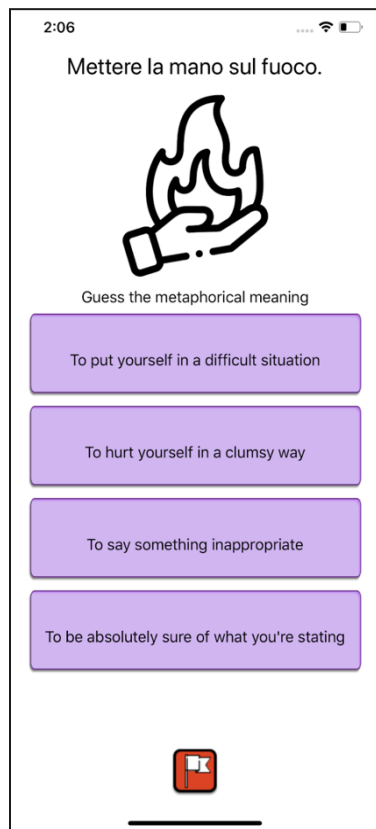


Fig. 2: Guessing the idiom metaphorical meaning



Fig. 3: Idiom card of “mettere la mano sul fuoco”

Once the correct translation is guessed, players get access to a second quiz on the current meaning of the expression, where they are asked to select one metaphorical meaning out of four semantic explanations, as fig. 2 shows.

In the end, when the correct answer is given, a lexicographic card is unlocked (fig. 3). This card, stored in a collection accessible anytime by the user, displays the metaphorical meaning of the idiom, its origins and possible equivalents in the player’s native language.

### 3.2 *Sfidioms*: a Role-play Activity to Make Practice with Idioms

The second game mode of *IdeoMania* is a practice activity to support idiom learning. In *Sfidioms* (fig. 4) players are asked to write down the idioms they come up with by looking at an emoji displayed on the screen. The quiz can be used as a single or multiplayer game to make the learning experience more entertaining.

Although the graphical user interface and the layout design still need to be improved, the app functionalities are fully implemented and the tool is going to be tested to assess both usability and the aforementioned learning function of the emojis to depict the idiom lexical structure.

## 4 Learning Idioms, Phraseological Theory and Possible Future Developments

The twofold quiz structure of *Idiomoji* is not only aimed at teaching a number of idioms but also at boosting metalinguistic awareness about the key features of their meaning, i.e. the literal and the figurative components, because these expressions “do not point to the target concept directly but via a source concept” (Dobrovolskij & Piirainen 2005:



40) using metaphorical, metonymic or synesthetic strategies. The literal meaning is responsible for different semantic and formal restrictions, as Dobrovolskij & Piirainen (2005) remark, therefore the two semantic components are both relevant to be proficient with this type of phrasemes. Besides, the guessing activity necessary for matching single words and pictograms not specifically created for this aim, such as emojis, is expected to be engaging but needs specific tests to assess its effectiveness as a learning strategy.

Nonetheless, drawings and illustrations have proved to be beneficial for idiom learning, as reported by Szczepaniak & Lew (2011) who claim that the positive effect can be explained in terms of the dual coding theory (Paivio 1986: 53-83): acquiring visual and verbal information at the same time facilitates comprehension and long-term retention because the depth of processing increases memorization. However, Boers (2009) reports on a distraction effect produced by pictures when learners must cope with unfamiliar and difficult words in the idiom lexical structure. In such instances, it is to be expected that the word-by-word translation methodology used in *Idiomoji* is beneficial to the learners.



Fig. 4: Sfidioms quiz mode of *IdeoMania*

## 5 The Pedagogical Framework

The study of the existing literature on idiom learning has also been carried out by the *LOR Foundation* students for defining the core concept of *IdeoMania*. Their findings were discussed during review sessions with the course “mentors”, experts who guide the learning process with the three-step phases of the *CBL* framework, called “Engage”, “Investigate”, and “Act” respectively. Each phase provides a set of instructions for transforming broad topics, called “Big Ideas”, into concrete and actionable app challenges. Following the constructivist assumption of learning by doing (Papert & Harel 1997; Sandholtz et al. 1997), the *Challenge Based Learning* urges students to build up autonomously the necessary knowledge on a topic they are passionate about to come up with an app solution for it. Teachers act as “mentors” in the learning process and are responsible for its success.

### 5.1 From Engagement to Action

The learning process starts with the Engage phase, during which students group together to work on a topic of common interest. Interactive brainstorming activities allow them to explore a closed set of broad ideas addressing humanities issues and turn them into a real-world problem with which students feel connected and, thus, urged to find an informed solution.

As an example, the *IdeoMania* project started from the broad theme of “Set phrases for making communication more effective” which was investigated by the *IdeoMania* team with the following “essential questions” (Nichols et al. 2016): How can idiom learning be improved? Which are the cognitive processes involved in memorizing idioms in other languages? What kind of difficulties have we faced, as learners, in learning idioms? What could be of help for students like us to learn idioms?

The Investigate phase starts from a “Challenge”, or a question which urges participants to study the chosen topic thoroughly. For example, the *IdeoMania* developers started with the following: “Can idiom learning become fun using images?”

Team members set out on a research phase (or “Investigate”) guided by a list of key topics (or “Guiding questions”) for finding valuable solutions to the challenge. They chose resources (questionnaires, web sites, scientific papers, interviews



with experts and so on) and planned the necessary actions to come up with an effective “Solution Concept”, like the following, written for *IdeoMania*:

Our idea is an app in which people can play with idioms using their pictorial representation. Our app will consist of two game modes: a single player mode and a team mode in real life. The team mode will make the game more challenging and fun also for native speakers who might be interested in playing games with their own language.

As soon as the app idea is outlined, teams start developing their tools in the so-called “Act” phase by using the same strategies commonly adopted in the design field.

## 5.2 Designing, Coding and Prototyping

App development during the Act phase is mostly concerned with UI (User Interface) and graphic design solutions to make for an effective app. Students attend design lessons throughout the course and learn the basic components of the screen layout (alignment, grids, proximity etc.), its textual contents (titles, labels, buttons) and the hierarchy between typographic elements by doing practical activities like dissecting existing apps or drawing sketches of their own projects. The first prototype is released by assembling screens originally drawn on paper and transforming them into an interactive mockup with fast prototyping tools like *Marvel*.

Coding lessons are also focused on implementation of UI components. After the first introductory hours on coding basics (i.e. constants, variables and operators) more complex topics (i.e. delegates or protocols) are explained to add basic UI components to the app, like dynamic cells of table views (see *Human Interface Guidelines*).

Creative design sessions and coding “nano-challenges” alternate with reviews and presentations of mockups. Mentors are generally the first users to test the app prototypes, but students are also encouraged to experiment their results with colleagues, friends, field-experts or whatever type of user outside the classroom. Given the immersive experience of the *LOR* classes, which last seven hours a day for four weeks, students receive daily feedback from mentors on their ideas and design solutions, prototyping is therefore a continuous workflow throughout the course.

As a closing remark, it should be outlined how *IdeoMania* testifies to new possible approaches to disseminating lexicographic knowledge by means of technology learning programs. The adopted framework is a standard methodology which can be adjusted both for teaching special programs in electronic lexicography or even for single lexicographic activities during class hours at any school level.

## 6 References

- Boers, F., Piquer Píriz, A. M., Stengers, H. & Eyckmans, J. (2009). Does pictorial elucidation foster recollection of idioms? In *Language Teaching Research*, 13(4), pp. 367-382.
- Caruso, V., Balbi, B., Monti, J. & Presta, R. (2019). How Can App Design Improve Lexicographic Outcomes? Examples from an Italian Idiom Dictionary. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubiček, S. Krek, C. Tiberius (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. Sintra, 1-3 October 2019. Brno: Lexical Computing CZ, pp. 374-396.
- Chiusaroli, F., Monti, J. & Sangati, F. (2017). *Pinocchio in emojiitaliano*. Sesto Fiorentino: Apice libri.
- Deterding, S., Dixon, D., Khaled, R. & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In A. Lugmayr (ed.) *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. New York: ACM, pp. 9-15.
- Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2019). Using gamification to support learning English as a second language: a systematic review. In *Computer Assisted Language Learning*, pp. 1-24.
- Dobrovolskij, D., Piirainen, E. (2005). *Figurative Language: Cross-Cultural and Cross-Linguistic Perspectives*. Oxford: Elsevier.
- Human Interface Guidelines, Accessed at: <https://developer.apple.com/design/human-interface-guidelines/ios/overview/themes/> [30/05/2020].
- Hung, H. T., Yang, J. C., Hwang, G. J., Chu, H. C., & Wang, C. C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, 126, pp. 89-104.
- Marvel, Rapid prototyping, testing and handoff for modern design teams, Accessed at: <https://marvelapp.com> [30/05/2020].
- Mihaljević, J. (2019). Gamification in E-Lexicography. In P. Bago, I. Hebrang Grgić, T. Ivanjko, V. Juričić, Ž. Miklošević, H. Stublić (eds.) *INFUTURE 2019: Knowledge in the Digital Age*. Zagreb, Croatia: Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, pp. 155-164.
- Monti, J., Caruso, V. (2019). L'Orientale Foundation: un programma per l'introduzione delle nuove tecnologie nei curricula umanistici. In S. Allegrezza (a cura di) *Didattica e ricerca al tempo delle Digital Humanities/ Teaching and research in Digital Humanities' era*, 8th Annual Conference, AIUCD 2019, Book of Abstracts, pp. 82-85. Accessed at: <http://aiucd2019.uniud.it/book-of-abstracts/> [30/05/2020].
- Monti, J., Sangati, F., Chiusaroli, F., Benjamin, M. & Mansour, S. (2016). Emojiitalianobot and emojiworldbot - new online tools and digital environments for translation into emoji. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. Final Workshop (EVALITA 2016), volume 1749 of CEUR Workshop Proceedings, Napoli, Italy. CEUR-WS.org, Accessed at: <https://dblp.org/db/conf/clic-it/clic-it2016> [30/05/2020].
- Nichols, M., Cator, K., Torres, M. & Henderson, D. (2016). *Challenge Based Learner User Guide*. Redwood City, CA: Digital Promise.



- Paivio, A. (1986). *Mental Representations*. Oxford University Press.
- Papert, S. Harel, I. (eds) (1991). *Constructionism: research reports and essays 1985-1990 by the Epistemology and Learning Research Group*. The Media Lab, Massachusetts Institute of Technology, Norwood, NJ: Ablex Pub. Corp.
- Sandholtz, J.H., Ringstaff, C. & Dwyer, D.C. (1997). *Teaching with Technology: Creating Student-Centered Classrooms*. New York: Teachers College.
- Szczepaniak, R., Lew, R. (2011). The Role of Imagery in Dictionaries of Idioms. In *Applied Linguistics*, 32(3), pp. 323-347.

### Acknowledgements

The authorship contribution is as follows: Johanna Monti is author of paragraph 2 and added notes on *Pinocchio in emojitaliano* in 3.1; Valeria Caruso is the author of the rest of the paper. Alessia Andrisani, Barbara Beatrice, Federica Contento, Zelinda De Tommaso, Fabiana Ferrara, Antonello Menniti developed the app described in the paper during *L'Orientale Apple Foundation* program.









**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Software Demonstrations**

**Lexicography and Corpus Linguistics**







# Skema: A New Tool for Corpus-driven Lexicography

Baisa V.<sup>1</sup>, Tiberius C.<sup>3</sup>, Ježek E.<sup>2</sup>, Colman L.<sup>3</sup>, Marini C.<sup>2</sup>, Romani E.<sup>2</sup>

<sup>1</sup> Lexical Computing, Brno, Czech Republic

<sup>2</sup> Università degli Studi di Pavia, Dept. of Humanities, Italy

<sup>3</sup> Instituut voor de Nederlandse Taal, Leiden, The Netherlands

## Abstract

In this paper, we describe the development of Skema and its features. Skema ['ski:mə] is a new corpus pattern editor system which supports the manual annotation of concordance lines with user-defined labels (each concordance has its own set of labels) and the editing of the corresponding patterns in terms of slots, attributes, examples and other features following the lexicographic technique of Corpus Pattern Analysis. Skema is integrated into the web-based Sketch Engine and can be used by any user for annotating both preloaded and user corpora. Each annotation label is linked to the pattern structure (stored in JSON format) which can be easily customized to individual projects, a generic pattern structure (i.e. a list of user-defined attributes) being available by default. The paper illustrates the use of Skema in three specific projects, i.e. *Woordcombinaties* for Dutch verbs, *Typed Predicate-Argument Structures* for Italian Verbs (T-PAS) and its sister project for Croatian Verbs (CROATPAS).

**Keywords:** corpus-driven lexicography; editor, pattern dictionary; sketch engine, corpus annotation; annotation schema

## 1 Introduction

Skema ['ski:mə] is a new corpus pattern editor system. It was implemented to facilitate the management of manual annotations in Sketch Engine (hence the name) (Kilgariff et al. 2014) allowing to associate word meaning with word use as is advocated by Corpus Pattern Analysis (Hanks 2013).

Corpus Pattern Analysis (CPA) is a lexicographic technique for mapping meaning onto words in text. It is based on the Theory of Norms and Exploitations (Hanks 2004; 2013). This theory distinguishes between normal or prototypical uses of words and exploitations of these, like patterns with anomalous collocates or unconventional metaphors. The focus of the analysis is on the prototypical syntagmatic patterns with which words in use are associated. Associating a “meaning” with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns.

In this paper, we describe the development of Skema and its features, and we illustrate the use of Skema in a number of projects, i.e. *Woordcombinaties* for Dutch verbs (Colman & Tiberius 2018), *Typed Predicate-Argument Structures* for Italian verbs (T-PAS; Ježek et al. 2014) and its sister project for Croatian CROATPAS (Marini & Ježek 2019).

## 2 The Skema editor

Sketch Engine has supported the manual annotation of corpora (concordance lines) for quite a long time (Baisa et al., 2015), but the management of annotations and the integration in Sketch Engine was clumsy and unstable, since different systems on different servers were involved. Therefore, a new interface with a modern look was implemented, which - more importantly - is much more stable and easier to maintain than the previous system. Below, we describe the two main components of Skema: the manual annotation of concordance lines (section 2.1) and the editing of the patterns (section 2.2).



Figure 1: From the left: annotating concordance lines, word sketch collocates and the annotation menu in the Sketch Engine interface



## 2.1 Manual annotation

Using the CPA technique, the lexicographer starts with analyzing and annotating a sample (usually 250 lines or more) of concordance lines with labels using the annotation feature (🔍) in the Sketch Engine interface. Patterns are identified manually by carefully examining similar concordance lines. The labels are chosen from a small pop-up menu (see Figure 1). It is a common practice to use numerical labels for the patterns, but any string can be used; labels with a dot (e.g. 1.a) are treated as sublabels and are grouped together under the main label.

It is important to note that each annotated concordance has its own set of labels. In the projects described below, the concordances are based on verb headwords, but any concordance (a query result) can be stored for later annotation.

The set of labels is maintained in the Annotation menu (Figure 1, on the right) which provides an overview of the labels together with the number of concordance lines annotated with that label and allows the user to add new labels, to sort the whole concordance by the labels or to show only lines annotated with a specific label. The user can also go to the Annotation manager, which is on a separate page (Figure 2).

The annotation is not only available in Concordance, but also in Word Sketch (Figure 1 in the center). Annotating per collocate can significantly speed up the process, since the label is assigned to all concordance lines containing the headword-collocate pair.

## 2.2 Editing patterns

Once the annotation is done, the lexicographer starts editing the patterns in the Annotation manager. Here, lexicographers have an overview of the list of all headwords (Figure 2). They can search the list and per headword, they can maintain the list of labels identifying the patterns (Figure 2 at the bottom). Labels can be renamed, added, removed and reordered in the Annotation manager, and these operations are synchronized with Sketch Engine.

Query	Label count	Frequency	Status	Edited	Editor
analyseren-v	5	6217	FINISHED	2020-02-27 15:40:59	inl02
annuleren-v	5	1382	FINISHED	2020-03-19 12:07:29	inl02
argumenteren-v	8	907	FINISHED	2020-03-20 16:05:58	inl02

Query	Label count	Frequency	Relative frequency	Status	Edited	Editor
annuire	1	3373	3.6048	NYS	2007-05-24 14:07:57.854677	deb
annullare	6	23377	24.9835	WIP	2014-05-21 16:24:55.512706	feltracco
annunciare	6	59948	64.0677	WIP	2014-03-05 21:31:58.316249	feltracco

Figure 2: The list of concordances and a list of labels with different pattern visualization (Dutch labels and patterns for *analyseren* ‘to analyze’ and Italian labels and patterns for *annunciare* ‘to announce’)

When a pattern has been edited, a preview of the pattern is shown next to the label. Each pattern corresponds to one of the labels used in the manual annotation. The visualization of patterns is also project-specific and can be customized. By clicking on one of the labels, the pattern (structured information) opens up and can be edited (see Figure 3).

Figure 3: Pattern editors for Dutch and Italian; the generated patterns at the bottom can be customized with colors, typography etc. Editing is done by selecting the right number of slots for a specific pattern and completing the information for the relevant



features for each slot (e.g. syntactic function, semantic type, lexical set). Each project can define its own features for the slots. Each project can thus have a different information structure linked to the slots and the different projects currently using Skema do indeed take advantage of the possibility to fine-tune the information in the slots to the particular needs of the project.

### 2.3 Skema: the technical solution

In the old version (called CPA editor; Baisa et al. 2015), the structured information was stored in a PostgreSQL database in several tables and every structural change in the pattern structure had to be reflected in the DB schema. Due to frequent changes, the schema became clumsy and hard to maintain. In Skema, the whole pattern structure is saved in JSON format in a SQLite database (one DB per corpus and project), so that changes are easily done at the level of the Skema pattern editor without a need to change the DB schema.

To use a customized pattern editor, Sketch Engine users need to be assigned to a specific project by Sketch Engine administrators. Otherwise, users will see a generic pattern editor with an option to save an arbitrary list of attribute-value pairs.

The system is currently not well-suited for parallel annotation by several annotators. Even though multiple annotators can work on different queries (stored concordances), the situation when more annotators (within one project and in one specific corpus) are editing the same concordance and changing the labels is not treated well at the moment and might lead to inconsistencies and conflicts. In the future, these situations can be treated by locking the annotation temporarily. In each project, the selected queries and their labels can be published as a read-only single-page website. At the moment, only two of the projects using Skema have such access: the English Pattern Dictionary of English Verbs<sup>1</sup> and the Italian T-PAS<sup>2</sup>. Currently, it is not possible for users to publish their own data, but Sketch Engine administrators can set up a new single-page website with a unified user interface (with simple customization options) similar to the two examples above.

## 3 Projects using Skema

In this section, we illustrate the use of Skema in three ongoing projects.

### 3.1 Dutch *Woordcombinaties* project ('Word combinations')

*Woordcombinaties* is a new online lexicographic resource from the Dutch Language Institute<sup>3</sup>, which merges a pattern dictionary of Dutch verbs, following the example of the Pattern Dictionary of English Verbs<sup>4</sup>, with a collocation application, following the example of Sketch Engine for Language Learning (SkELL)<sup>5</sup>. A demo<sup>6</sup> of the resource has recently been released describing the combinatorics of a selection of 150 verbs taken from a list of high frequency verbs for advanced learners of Dutch as a second language. For all verbs, example sentences and a kind of word sketch are provided, and for a subset, a pattern description is also available. The project is based on a corpus of approx. 230 million tokens consisting of newspaper material and domain specific texts from the Netherlands and Belgium, in order to reflect language variety in Dutch.

In the editorial process, Skema is used for editing patterns, whereas an in-house system (Tiberius et al. 2014) is used for editing the example sentences and word sketches. The data from both editors is integrated in the online *Woordcombinaties* application. The basic setup of Skema for *Woordcombinaties* is fairly similar to the other projects. A pattern consists of a number of slots and each slot has a number of features and attributes attached to it. Specific to the Dutch project are the features 'fixed element' and 'dummy'. Dummies (such as *iemand* 'someone', *iets* 'something') are used in addition to semantic types for the sake of readability. This practice is inspired by E-VALBU<sup>6</sup>, where complements in the patterns are also embedded in dummies so that semantic roles are more or less implicitly recognizable. In the patterns in the online application, only dummies are shown, not the semantic types. The fixed element was introduced to have a placeholder for prepositions and conjunctions separate from the dummy in prepositional complements and predicative modifiers as is illustrated in the pattern below, where there are two optional prepositional complements, one introduced by *in* ('in') and the other by *op* ('on').

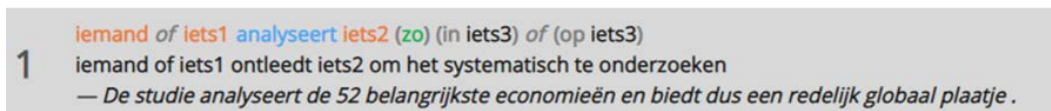


Figure 4: Pattern 1 of Dutch *analyseren* ('to analyse')

The visual rendering of the patterns in Skema has been customized completely to the *Woordcombinaties* project and is very similar to the layout and color-coding used in the online application of the project, providing a WYSIWYG preview to the lexicographer. For instance, the OR feature is displayed as the Dutch word *of* ('or') in the pattern (see Figure 4), and

<sup>1</sup> [pdev.org.uk](http://pdev.org.uk) [04/05/2020]

<sup>2</sup> [tpas.sketchengine.eu](http://tpas.sketchengine.eu) [04/05/2020]; The T-PAS project is currently using this feature only internally and enriching it with good examples (GDEX) from the corpus for each label. The results will be made public by the end of 2020.

<sup>3</sup> [ivdnt.org](http://ivdnt.org) [04/05/2020]

<sup>4</sup> [www.sketchengine.eu/skell/](http://www.sketchengine.eu/skell/) [04/05/2020]

<sup>5</sup> [woordcombinaties.ivdnt.org/](http://woordcombinaties.ivdnt.org/) [04/05/2020]

<sup>6</sup> [grammis.ids-mannheim.de/verbvalenz](http://grammis.ids-mannheim.de/verbvalenz) [04/05/2020]



different syntactic functions are marked by different colors. Note also that the pattern uses the inflected form of the verb. In addition to the pattern, a definition and a general example are given.

In *Woordcombinaties*, sublabels are used to distinguish subpatterns from main patterns. Subordinate clauses, idioms, proverbs and formulas are normally considered as subpatterns. For instance, the pattern in Figure 4 has two subpatterns: one where the object slot of the main pattern is realized by a subordinate clause introduced by *of* ('or') or a *wh*-word, and one where the object is realized by a quote.

In the *Woordcombinaties* version of Skema, patterns are complemented by two types of example sentences, a general example and selection of examples of lexical items instantiating a particular slot. The slots in a pattern are numbered and the active slot is highlighted so that the (GDEX sorted) selected examples are automatically linked to this slot (see slot 6 in Figure 5). This numbering of the slots is especially important if a slot with a particular syntactic function occurs more than once in a pattern, as in the example below.

	1	2	3	4	5	6
Function	Subject	Head	Object	Adverbial	Prepositional	Prepositional
Semantic type	Human					
Fixed element					in	op

Example 43	Lexical item 43	Slot
Voor die locaties wordt het water standaard geanalyseerd	op	pc (6)

Figure 5: Pattern for Dutch verb *analyseren* 'to analyze' illustrating automatic linking of slots to example sentences

Both types of example sentences are stored in the JSON file which can be downloaded from Skema. In addition to this, the full set of annotated concordances in the corpus is extracted through the Sketch Engine API and shown on demand to the user in the online version of *Woordcombinaties*.

### 3.2 The Italian T-PAS project

The T-PAS resource is a corpus-derived inventory of semantic structures for Italian verbs to be used for linguistic analysis, language teaching, and computational applications. It is developed at the Department of Humanities of the University of Pavia, in collaboration with the Fondazione Bruno Kessler (FBK, Trento) and the technical support of Lexical Computing Ltd. It is based on the Generative Lexicon theory of compositionality (Pustejovsky 1995) and on the corpus pattern lexical analysis proposed in Hanks (2004, 2013). It currently consists of 1160 analyzed verbs for about 8000 patterns, and ca. 190,000 annotated concordances. The verb sample of T-PAS was selected according to two criteria: a random sample of average polysemy verbs from the Sabatini Coletti 2008 dictionary (10% of 2 sense verbs, 60% of 3-5 sense verbs, 30% of 6-11 sense verbs), and coverage of the fundamental verb lemmas ("lemmi fondamentali") from De Mauro 2000<sup>7</sup>. The corpus used to extract the patterns is the itWaC reduced (935,698,409 tokens), a wide corpus gathered by crawling texts from the Italian domain in the web using medium frequency vocabulary as seeds (Baroni et al. 2009). The resource includes a repository of patterns, a hierarchically organized system of semantic types to classify the semantic properties of the verbal arguments, and a corpus of annotated concordances with pattern numbers, that represent instantiations of the corresponding patterns.

The T-PAS System of Semantic Types (Jezek 2018) is a hierarchy of general semantic categories obtained from manual clustering of lexical items found in the argument positions of verbal structures in the corpus. The System currently contains 180 semantic types that are organized hierarchically based on the "is a" (subsumption) relation (e.g., [Human] is an [Animate]). The System of Semantic Types, together with definitions and examples for each type, is made accessible to lexicographers through a customized function of Skema (Figure 6), so that they can easily and instantly consult it while editing the patterns.

#### System of Semantic Types

##### ANYTHING –

ST to use as a last resort when [Eventuality], [Entity] and [Property] are equally likely

##### PROPERTY +

A quality or characteristic of [Anything] (peso, altezza, bellezza, forma, eleganza, reputazione)

##### EVENTUALITY –

It can either be an [Event] involving movement, change or development or a fixed [State] (evento, relazione, cambiamento, situazione)

##### STATE +

A static [Eventuality] that does not involve activity, movement or development (pace, stabilità, situazione, equilibrio)

##### EVENT +

An [Eventuality] that involves movement, change, or development, unlike a [State]. An [Event] can either be a volitional [Activity] or a non volitional [Process]. (incontro, morte, visita, matrimonio, trattamento, tempesta, guerra, richiesta)

##### ENTITY –

[Anything] that exists independently of other things and has a distinct identity. [Anything] which is not an [Eventuality] nor a [Property] (forza, materiale, ambiente, economia, creatura, edificio)

##### ABSTRACT ENTITY +

An intangible [Entity], such as a [Concept] (idea, problema, concetto)

##### PHYSICAL ENTITY +

A tangible [Entity] (ponte, faccia, tavolo, auto, fiore, uccello, birra, merci, pietra, bambino, vulcano)

Figure 6: A selection of the top level of the hierarchy for the T-PAS System of Semantic Types in Skema.

<sup>7</sup> In De Mauro's classification, fundamental lemmas are the words that in all languages tend to cover on average about 90 percent of the occurrences of words in texts and discourse.



Importantly, pattern strings in the T-PAS customized version of Skema only show semantic and lexical information, that is, they include the verb, the semantic type of the arguments, a selection of the best examples of lexical items for the types, the role played by the arguments (i.e. Athlete, Doctor), the features (i.e. Female, Visible) associated to the types, and a preposition or a complementizer (*a, per, di*), should they be present in the pattern. The syntactic information is encoded in Skema, but it is deliberately not made visible in the pattern string, neither in Skema nor in the online version. This is because the resource is intended primarily as a semantic resource. The syntactic features available for pattern encoding by the lexicographer are: subject, object, prepositional complement (this includes indirect objects), adverbial, clausal, predicative complement, and QDM (quantifier, determiner, modifier) - the latter for argument slots with rigid syntax regarding these features, for example arguments that must be introduced by a determiner. One important feature of T-PAS is that it allows for syntactic alternation within the same pattern, as in Figure 7:

1 [Human] finire [Activity] | di [Activity]  
[Human] conclude, porta a termine [Activity] | di [Activity]

Figure 7: Syntactic alternation in T-PAS for the verb *finire* ‘to finish’, which allows for both a direct object and a clausal argument introduced by *di* to express the semantic selection [Activity] for the second argument.

Another main feature of T-PAS is that it encodes metonymic shifts on the arguments (Pustejovsky & Jezek 2008). The idea of registering metonymic instances in the patterns emerged from the need of addressing the divergence between the frequency of metonymic instances in the corpus and the lack of a proper way to record this kind of information in the resource. Therefore, we implemented Skema with the addition of a specific sublabel, .m (where “.m” stands for metonymic); metonymic sublabels are linked to their main label and reflect their syntactic structure, as well as the sense of the verb, which does not change. The metonymic sublabel encodes the new semantic type(s); the shift between the type in the label and the metonymic type is also registered, see the second line of sublabel 1.m in Figure 8. The metonymic sublabel has been applied to a preliminary sample of 30 verbs in Romani (2020); we are interested in extending the number of verbs annotated for metonymies in their arguments.

1 [Animate] bere [Beverage]  
[Animate] ingerisce, assume [Beverage]  
1.m [Animate] bere [Container {bicchiere | bottiglia}]  
[Animate] ingerisce, assume [Container] (che contiene [Beverage])

Figure 8: Metonymic sublabel in T-PAS for the verb *bere* ‘to drink’, where the semantic shift between the type [Beverage] and the metonymic type [Container] is registered.

Finally, in developing and customizing Skema for T-PAS, we devoted attention to the graphic layout and visualization of the patterns, in order to make them easily-readable and user-friendly. We used as few symbols as possible and conceived a system to properly combine lexical and semantic information (as in the metonymic sublabel in Figure 8).

### 3.3 CROATPAS

The CROATian Typed Predicate Argument Structure resource (CROATPAS, Marini & Ježek 2019) is the Croatian sister project of the T-PAS resource (see section 3.2). The two projects share the same corpus-based lexicographic methodology and a number of common features, such as the focus on metonymic shifts taking place within argument structures. The reference corpus linked to the resource is the Croatian Web as Corpus (Ljubešić & Erjavec 2011), which contains over 1.2 billion tokens. CROATPAS’s first release is scheduled for the end of 2020. At present, its inventory consists of 101 verb entries, 457 patterns, 106 metonymic subpatterns and 22,052 annotated corpus lines (Marini & Ježek 2020). Being a Slavic language, Croatian posed a certain number of issues which had to be tackled when Skema was implemented for CROATPAS, such as the graphical rendering of case inflection in pattern strings. The Croatian case system consists of seven cases, namely nominative, genitive, dative, accusative, vocative, locative and instrumental (Barić et al. 1997: 101). Since it is mainly noun endings that express the grammatical relations between sentence components, we soon realized that – if we planned to translate Croatian valency structures into CROATPAS patterns using non-inflected Semantic Types – we had to find an effective way to convey the morpho-syntactic information usually provided by case, since we could not even rely on fixed word order nor on an extensive inventory of prepositions. In addition to color-coding the different argument slots, the solution was adding case markings as bottom-right indexes to the Semantic Types in the pattern strings, as in the example portrayed in the figure below.

1 [Human | Institution | Software]<sub>ACCUSATIVE</sub> preporučuje [Activity]<sub>ACCUSATIVE</sub> (korištenje krema) | da [Activity] (nikako ne gubite notu) | [Activity] [Human]<sub>GENITIVE</sub>  
[Human], [Institution] or [Software] recommends [Activity] to [Human]

Figure 9 CROATPAS pattern 1 of the verb *preporučivati* ‘to recommend’

Another Croatian-specific feature to be taken into account when setting up Skema was verbal aspect. Croatian verbs usually come in pairs featuring both a perfective and an imperfective lexical variant, thus allowing language users to



choose between two different options according to the temporal constituency of the given event. Therefore, in CROATPAS each aspectual variant is treated as an independent verb entry.

## 4 Conclusion

This paper introduced Skema, a new corpus pattern editor system. Skema is a web-based editor integrated into Sketch Engine which combines two new features: annotating concordance lines with labels for patterns and management of these labels with the possibility of storing arbitrarily structured information for each pattern label.

In this paper, we described three projects which employ the technique of Corpus Pattern Analysis, all of which are using Skema. Another project which has been recently moved to Skema is the Pattern Dictionary of English Verbs (Hanks and Pustejovsky, 2005). Since all four projects use a very similar pattern structure, inter-language linking of patterns (verb meanings) should be relatively easy and the resulting dictionary (as envisaged in Baisa et al. 2016) of verb valencies would form a valuable resource for both researchers and language learners.

Skema is being actively developed and new features are expected to be added, such as user-customizable pattern structures, reliable collaborative annotation, support for online self-publishing of the data and an export function.

## 5 References

- Baisa, V., El Maarouf, I., Rychlý, P., Rambousek, A. (2015). Software and Data for Corpus Pattern Analysis. In *RASLAN*, pp. 75-86.
- Baisa, V., Može, S., & Renau, I. (2016). Multilingual CPA: Linking Verb Patterns across Languages. In *Proceedings of the XVII Euralex International Congress*, pp. 410-417.
- Colman, L., Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In *Proceedings of the XVIII EURALEX International Congress*, pp. 233-246.
- Barić, E., Lončarić, M., Malić, D. Pavešić, S., Peti M., Zenčević V., Znika M. (1997). *Hrvatska gramatika*. Zagreb: Školska knjiga.
- Baroni, M., Bernardini S., Ferraresi A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. In *Language resources and evaluation*, 43(3), pp. 209–226.
- De Mauro, T. (2000). *Grande dizionario dell'uso* (GRADIT). Torino: UTET.
- Ježek, E. (2018). Classi di nomi tra semantica e ontologia. In Masini, F. and F. Tamburini (eds.) *CLUB Working Papers in Linguistics*, 2, Bologna: CLUB (Circolo Linguistico dell'Università di Bologna), pp. 117-131.
- Ježek E., Magnini B., Feltracco A., Bianchini A., Popescu O. (2014). T-PAS: A resource of Typed Predicate Argument Structures for Linguistic Analysis and Semantic Processing. In *Proceedings of LREC 2014*, Reykjavik, Iceland, pp. 890-895.
- Kilgariff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1(1), pp. 7-36.
- Marini, C., Ježek E. (2019). CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (Clic-it)*. Bari, Italy.
- Marini, C., Ježek, E. (2020). Annotating Croatian Semantic Type Coercions in CROATPAS. In *Proceedings of the Sixteenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-16)*. France, Marseille, pp. 50-55.
- Hanks, P. (2004). Corpus pattern analysis. In *Proceedings of the 11th EURALEX International Congress*, pp. 87-98.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: The MIT Press.
- Hanks, P., Pustejovsky, J. (2005). A pattern dictionary for natural language processing. In *Revue Française de linguistique appliquée*, 10(2), pp. 63-82.
- Ljubešić N., Erjavec, T. (2011). hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Habernal I., Matoušek V. (eds.) *Text, Speech and Dialogue, Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer-Verlag, pp. 395-402.
- Pustejovsky J. (1995). *The Generative Lexicon*. Cambridge MA: The MIT Press.
- Pustejovsky J., Ježek E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In *Rivista di Linguistica (Italian Journal of Linguistics)*, vol. 20, pp. 181-214.
- Romani, E. (2020). Searching for Metonymies in Natural Language Texts. A Corpus-based Study on a Resource for Italian Verbs. BA Thesis, University of Pavia, Pavia, Italy.
- Sabatini, F., Coletti, V. (2007). *Il Sabatini Coletti 2008. Dizionario della Lingua Italiana*. Milano: RCS Libri S.p.A. [https://dizionari.corriere.it/dizionario\\_italiano/](https://dizionari.corriere.it/dizionario_italiano/). [04/05/2020]
- Tiberius, C., Niestadt, J. & Schoonheim, T. (2014): 'The INL Dictionary Writing System'. In Kosem, I. and Rundell, M. (eds.) *Slovenščina 2.0: Lexicography*, 2 (2), pp. 72–93.

## Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.



# CROATPAS: A Lexicographic Resource for Croatian Verbs and its Potential for Croatian Language Teaching

Marini C., Ježek E.

University of Pavia, Italy

## Abstract

This paper revolves around CROATPAS (Marini & Ježek 2019), a digital lexicographic resource for Croatian verbs able to frame verbal polysemy and metonymic shifts, which is currently being developed at the University of Pavia. Just like its Italian sister resource T-PAS (Ježek et al. 2014), CROATPAS is a corpus-derived collection of verb argument structures whose argument slots have been manually annotated using a specific set of semantic labels called Semantic Types. At the moment, the resource contains 101 verb entries linked to 457 different verb senses (called *patterns*) and over 22,000 annotated corpus lines (Marini & Ježek 2020). The possible applications of CROATPAS are endless. However, given the status of Croatian as an *under-resourced* and Less Commonly Taught Language, this paper focuses on its potential as a language teaching tool, putting forward some hypothetical vocabulary and grammar teaching suggestions. Even though CROATPAS is still in the early stages, its user-friendly interface, bilingual nature and focus on verb semantics bode well for its future as a tool for the teaching of Croatian as a Foreign Language.

**Keywords:** Croatian; semantic resource; verb; language teaching; Less Commonly Taught Language

## 1 Introduction

In this paper, we present a new corpus-derived lexicographic resource for Croatian called CROATPAS (Marini & Ježek 2019). The resource focuses on verb semantics and polysemy; it currently contains 101 verb entries, 457 verb senses and is linked to over 22,000 annotated corpus lines (Marini & Ježek 2020). CROATPAS relies on a sound theoretical background and a well-established lexicographic methodology, which are thoroughly dealt with in section 2. Even though its possible applications are endless, here we explore its potential as a language teaching tool, putting forward some of the vocabulary and grammar teaching activities already offered by its pattern inventory (see § 3). In section 4, an overview of the currently available resources for Croatian verbs is provided, stressing their differences with respect to CROATPAS. In light of its user-friendly interface, bilingual nature and focus on verb semantics, we believe our resource has the potential to become a truly useful tool for the teaching of Croatian as a Foreign Language.

## 2 CROATPAS

The CROATian Typed Predicate Argument Structures resource (CROATPAS, Marini & Ježek 2019) is a digital lexicographic resource containing a corpus-based collection of manually annotated Croatian verb valency structures with the addition of Semantic Type labels on their argument slots (henceforth, SemTypes). Since each semantically typed verb argument structure – informally called *pattern* – is linked to a different verb sense, the resource is primarily tailored for investigating verbal polysemy. For instance, *[Human]<sub>nominative</sub> pije [Beverage]<sub>accusative</sub>* and *[Human]<sub>nominative</sub> pije [Drug]<sub>accusative</sub>* are two of the identified patterns of the Croatian verb *piti* (English, *to drink*), which correspond to the following senses: that of *drinking a beverage* and that of *swallowing a medicine*, respectively. The resource is also able to frame recurrent metonymic shifts in verb arguments, which are encoded in *sub patterns* nested within the patterns they stem from (see § 2.2).

At the moment, CROATPAS contains 101 verb entries, which are linked to 457 patterns, 106 metonymic sub patterns and over 22,000 annotated corpus lines (Marini & Ježek 2020). Like its Italian sister project T-PAS (Ježek et al. 2014, containing 1160 analysed verbs, 8000 senses, and approximately 190,000 annotated concordances), the Croatian resource is being developed at the University of Pavia with the technical support of *Lexical Computing Ltd.* and will be available online in late 2020. The four components the resource relies on are:

- a representative corpus of contemporary Croatian
- a set of semantic labels to tag argument slots with
- a sound corpus-based lexicographic methodology to identify different verb senses
- the adequate corpus tools

As far as the corpus choice is concerned, the Croatian Web as Corpus (hrWac 2.2, Ljubešić & Erjavec 2011) was chosen as the reference corpus for CROATPAS in order to maximise compatibility with the Italian T-PAS resource (Ježek et al. 2014), since the latter is linked to a reduced version of the Italian Web as Corpus (ItWaC, Baroni & Kilgariff 2006). As for the semantic labels, CROATPAS takes advantage of a shallow ontology of approximately 180 SemTypes developed for the T-PAS project and called System of Semantic Types (Ježek 2019). In terms of corpus tools, the main software CROATPAS makes use of is a pattern editing environment linked to the *Sketch Engine* (Kilgariff et al. 2014) called *Skema* (Baisa et al. 2020). Skema was originally developed by *Lexical Computing Ltd.* for the Italian T-PAS resource, but



it was later customised not only to the needs of CROATPAS, but also of several other projects focused on different languages, such as the *Woordcombinaties* projects for Dutch verbs (Colman & Tiberius 2018). Finally, as far as methodology is concerned, the resource relies on a customised version of Corpus Pattern Analysis (CPA, Hanks 2013), a lexicographic methodology resting on the idea that meaning should be mapped onto its prototypical contexts of use, which was first put into practice in the *Pattern Dictionary of English Verbs* (Hanks & Pustejovsky 2005). CPA usually follows four steps: 1) 250 corpus lines are randomly sampled for each verb; 2) the different verb senses are identified through extensive lexicographic analysis; 3) pattern strings are created in Skema labelling argument slots with the right SemTypes and, finally, 4) numbers are assigned to the corpus lines exemplifying each identified pattern, in order for each semantically tagged valency structure to be justified by corpus evidence.

## 2.1 Understanding Patterns

From a theoretical point of view, both T-PAS and CROATPAS rely on the Generative Lexicon framework and its principles for strong compositionality, namely the principle of co-composition and the principle of Semantic Type Coercion (Pustejovsky 1995 & 1998; Pustejovsky & Ježek 2008; Ježek 2016).

According to the principle of co-composition, lexical items expressing verb arguments are to be considered as semantically active in the contextual generation of verb meaning as the verb itself. In light of this, verbal polysemy can be traced back to the compositional operations taking place between the verb and the SemTypes associated to its surrounding arguments. For instance, as we can see from Figure 1, the Croatian verb *voziti* (English, *to drive*) takes on different meanings depending on what is said to be *driven*.

1	[Human] <sub>NOMINATIVE</sub> vozi [Vehicle] <sub>ACCUSATIVE</sub> {bicikl   auto   tramvaj   avion} [Human] drives, rides or flies a [Vehicle]
2	[Human] <sub>NOMINATIVE</sub> vozi [Human] drives, travels by car
3	[Human] <sub>NOMINATIVE</sub> vozi [Human] <sub>ACCUSATIVE</sub> u [Location] <sub>ACCUSATIVE</sub> u {školu   bolnicu} [Human] accompanies [Human] to [Location] by car

Figure 1: Some of the patterns connected to the verb *voziti* (English, *to drive*) in CROATPAS

If a *[Human]* drives a *[Vehicle]* as in pattern (1), then he or she is “operating that vehicle in order for it to move”. However, if a *[Human]* drives another *[Human]* to a *[Location]* as in pattern (3), then he or she is “accompanying that person to a certain destination, usually by car”. The meaning of each pattern string is explained in English in the line underneath the pattern, which takes the name of *sense description*. The choice of the right SemType for each argument slot in a pattern is made by the lexicographer on the basis of the lexical items found in that slot in the corpus lines linked to that pattern: for instance, the SemType *[Vehicle]* in pattern (1) is justified by corpus examples featuring direct objects such as {*bicikl* = bicycle | *auto* = car | *tramvaj* = tram | *avion* = plane}, which make up a the so-called *lexical set* for the SemType *[Vehicle]* in this context.

## 2.2 Understanding Subpatterns

As for Semantic Type Coercions, that is how metonymic shifts are called in a Generative Lexicon perspective (Pustejovsky & Ježek 2008). Coercions take place when a verb’s selectional requirements in terms of semantic typing are not satisfied by one of its arguments, but no change in verb meaning is observed. For instance, if we look at the corpus line highlighted in Figure 2 – *Ako ne voziš BMW, ti si nitko i ništa* (English, *If you do not drive a BMW, you are nobody and nothing*) – we can see that, even though the example conveys the same meaning encoded in pattern 1 from Figure 1, the direct object of *voziti* (English, *to drive*) is not a *[Vehicle]* *per se*, but a *[Business Enterprise]* producing *[Vehicle]*s.

rivo u njega. Možda zato što dosada nikada nije vozila golf ili je	vozila	1.1.m	BMW ili neki drugi auto koji ima drugačiju tehniku. Na kraju sebi
li kao i najobičniji mobitel Al ' košta Ista stvar sa autima. Ako ne	voziš	1.1.m	BMW ti si nitko i ništa. Samo bi trebalo vidjeti i ove koji voze BM
ne voziš BMW ti si nitko i ništa. Samo bi trebalo vidjeti i ove koji	voze	1.1.m	BMW dal uopće imaju za kavu ali se pred susjedima mora poka:
0 eura. JA JA (anoniman posjetitelj) 24.6.13. 08:40 Svaka šuša	vozi	1.1.m	BMW i Mercedes, i još sluša kurzu u njima, fuj 24.6.13. 00:44 ee
aštvom kaže Dinko Vodanović. Crna kronika ' Živim od socijale,	vozim	1.1.m	BMW, imam uvijek 70 tisuća kuna u džepu, brani me Čedo Prod
o taj novac i ne seljačim se s time da jedem žgance za ručak al	vozim	1.1.m	BMW na kredit. Moj je, platio ga u keš i nisam nikome dužan. OI
a 52 puta bio strelac Svetski šampion Valentino Rosi testirao je	vozilo	1.1.m	Ferarija na stazi Mudjelo, na kojoj je trijumfovao u posljednjih ser

Figure 2: Some of the metonymic corpus lines linked to the sub pattern 1.1.m for the verb *voziti* (English, *to drive*) in CROATPAS.

When someone refers to a *[Vehicle]* in terms of the *[Business Enterprise]* – and specifically the *Automobile Company* – that produces it, then we are witnessing a Semantic Type Coercion from *[Business Enterprise]* → *[Vehicle]*, which can be listed as an instance of the pervasive metonymy *Producer/Product* (Pustejovsky 1995: 25). Both T-PAS and CROATPAS encode Coercions as sub patterns ending in “.m” (which stands for *metonymic*), as you can see in Figure 3.



1	[Human] <sub>NOMINATIVE</sub> vozi [Vehicle] <sub>ACCUSATIVE</sub> {bicikl   auto   tramvaj   avion} [Human] drives, rides or flies a [Vehicle]
1.1.m	[Human] <sub>NOMINATIVE</sub> vozi [Business Enterprise : Automobile Company] <sub>ACCUSATIVE</sub> {Ferrari   BMW} [Human] drives a [Vehicle] produced by certain [Business Enterprise]

Figure 3: Pattern 1 and its metonymic sub pattern 1.1.m for the verb *voziti* (English, *to drive*) in CROATPAS

### 3 CROATPAS as a Language Teaching Tool

After presenting the main features of CROATPAS, we now turn to its applications. Its potential uses are countless and they entail, *inter alia*, corpus-based linguistic research on verbal polysemy and metonymies (Marini & Ježek 2020), but also computational applications (e.g. machine translation enhancement) based on multilingual pattern linking with other monolingual CPA-inspired resources, such as T-PAS for Italian (Ježek et al. 2014), PDEV for English (Hanks & Pustejovsky 2005) or *Woordcombinaties* for Dutch (Colman & Tiberius 2018).

Moreover, given the status of Croatian as both an *under-resourced* (Tadić et al. 2014) and a Less Commonly Taught Language (LCTL) in need of attention (Mikelić Preradović et al. 2019), CROATPAS could also become a useful tool for teachers of Croatian as a Foreign Language. In the following, we focus on the latter application.

According to the widely accepted principles of Communicative Language Teaching (Brown & Lee 2015: 31), the best practice in language teaching promotes exposure of language learners to real-life communication and meaningful input embedded in naturally occurring language, which is exactly what corpus-based language teaching has been offering over the past few decades, especially thanks to tools such as corpus concordances and collocation lists (*ivi*, 62). More recently, due to the increasing integration of technology in our daily lives and the development of the Web, we have also witnessed the rise of Computer Assisted Language Learning (CALL), which is characterised, among other things, by an increased “interactive communication and collaboration [among language learners] via the Internet” (*ivi*, 238).

Although existing research has proved the effectiveness of corpus usage on both grammar learning, vocabulary learning and language awareness (Chan & Liou 2005; Liu & Jiang 2009; Lee et al. 2018), research on LCTLs is scarce (Ward 2016). According to the results of a recent study on teaching Croatian as a Foreign Language involving learners with different native languages and proficiency levels (Mikelić Preradović et al. 2019), student response to corpus-based material in experimental classes was mostly positive: beginners showed higher levels of appreciation for the introduction of corpora than advanced learners, while most intermediate learners enjoyed discovering corpus patterns, but were easily overwhelmed.

Given the well-known difficulties connected to learning verb/noun collocations even at advanced level (Nesselhauf 2003) and the attested positive influence of bilingual CALL tools on verb/noun collocation learning (Chan & Liou 2005), we believe that using CROATPAS in the classroom could help learners improve their understanding and usage of Croatian verb patterns.

In the rest of this section, we suggest some vocabulary and grammar teaching activities which could already be carried out with CROATPAS (see § 3.1 and 3.2), and we discuss the possibility of integrating some SkELL-inspired features in its interface (see § 3.3).

#### 3.1 Teaching Vocabulary with *Lexical Sets* and *SemTypes*

In addition to being linked to corpus examples providing evidence for their existence, CROATPAS's patterns feature a selection of manually identified verb collocates in each argument slot (see Figure 4). These collocate lists take the name of *lexical sets* (Hanks & Ježek 2008), they are portrayed in braces after their respective SemType and can be exploited in vocabulary teaching activities focusing on the semantic areas of their respective SemTypes.

1	[Human = Doctor   Drug   Activity : Medical Treatment] <sub>NOMINATIVE</sub> {Isus   terapija} izliječi [Animate = Patient   Body   Part of Body] <sub>ACCUSATIVE</sub> {djecu   slomljeno srce   umorne oči} (od [Illness]   [Injury]) <sub>GENITIVE</sub> {od {rana}}
2	[Human = Doctor   Drug   Food   Activity : Medical Treatment] <sub>NOMINATIVE</sub> {lijekovi   tablete   zdrava prehrana} izliječi [Illness   Injury] <sub>ACCUSATIVE</sub> {bolesti   depresiju   rane u duši}
3	[Human = Patient] <sub>NOMINATIVE</sub> {pacijent} se izliječi (protiv   od [Illness]) <sub>GENITIVE</sub> protiv   od {raka   virusa HIV-a}

Figure 4: The first three patterns of the verb *izliječiti* (English, *to heal*) in CROATPAS

For instance, since the verb *izliječiti* (English, *to heal*) involves SemTypes such as [Drug], [Activity: Medical Treatment], [Illness] and [Injury] in its pattern strings, the lexical sets of these SemTypes are bound to feature collocates pertaining to the medical semantic area, such as *lijekovi* (English, *meds*), *terapija* (English, *therapy*), *rana* (English, *wound*), *rak* (English, *cancer*), *bolest* (English, *illness*). If users could access a SemType search menu able to query the pattern inventory, they could easily retrieve all the verbs featuring the desired SemType and all its related terminology.

Moreover, teachers could exploit the fact that each lexical item featuring in a lexical set is adapted to its grammatical context both in terms of morphological inflection and required preposition. For instance, if we look at pattern 3 from



Figure 4, we can see that all the items in the lexical set of the SemType [Illness] bear a genitive singular ending, as required by both the preposition *protiv* and *od*: *od raka*; *protiv virusa HIV-a* (English, *against cancer*; *against the HIV virus*).

### 3.2 Teaching Case Inflection with *Patterns*

Being a Slavic language, Croatian is equipped with a case system consisting of seven different morphological cases, namely nominative, genitive, dative, accusative, vocative, locative and instrumental (Barić et al. 1997: 101).

Since Croatian does not have a fixed word order, in patterns such as the one in Figure 5 portraying the verb *dočekati* (English, *to welcome*), where both arguments are semantically typed as [Human], it is *case* that allows us to understand the grammatical relations between sentence components, i.e. which of the two arguments is the *welcoming* subject expressed by the nominative case and which is the *welcomed* object in the accusative case.

1 [Human = Host | Human Group = Host]<sub>NOMINATIVE</sub> {direktor | suprug | hrvatski narod} dočeka [Human = Guest | Human Group = Guest]<sub>ACCUSATIVE</sub> {goste | nogometaše} ((s | sa) [Activity] (s | sa) {ovacijama})  
[Human] or [Human Group] welcomes, greets [Human] or [Human Group] with [Activity]

Figure 5: Pattern 1 of the verb *dočekati* (English, *to welcome*) in CROATPAS

Even though a resource mainly focusing on verbal polysemy like CROATPAS might be more suitable for intermediate and advanced learners, both the graphical rendering of case markings as bottom-right indexes and the argument colour-coding strategy might be useful devices for teachers to introduce absolute beginners to the concept of case, especially before they have internalized all the different inflectional endings for the different classes of nouns.

### 3.3 Combining *Patterns*, *Concordances* and *Collocations*

In order to become an ever more user-friendly tool for teachers and learners of Croatian as a Foreign Language, CROATPAS could also take example from the Dutch *Woordcombinaties* project<sup>1</sup> (Colman & Tiberius 2018), which – to our knowledge – is the only verb-centred CPA-inspired lexicographic resource equipped with both an inventory of pattern-meaning pairs and some of the features of a SkELL (*Sketch Engine for Language Learning*) interface (Kilgariff et al. 2015). As you can see from Figure 6, *Woordcombinaties* users have access to:

- corpus concordance lines portraying Good Dictionary Examples for each verb, i.e. prototypical *example sentences* (Dutch, *voorbeeldzinnen*) automatically extracted by the GDEX algorithm developed by Kilgariff et al. (2008);
- the verbs' *combination possibilities* (Dutch, *combinatiemogelijkheden*), i.e. a list of the lexical items they are usually found together with divided per word class, sketching their grammatical and collocational behaviour;
- the verbs' different senses encoded in *patterns* (Dutch, *patronen*) based on their valency structures, albeit without overt semantic typing.

Figure 6: Some example sentences, combination possibilities and verb patterns of the Dutch verb *denken* (English, *to think*) taken from the *Woordcombinaties* project interface

In the case of Croatian, it might be interesting to bring this hybrid approach a step further integrating other SkELL-inspired features in the tool interface, such as the possibility to generate word clouds displaying verb collocates divided not only per verb argument but also per specific verb sense, i.e. *pattern*.

<sup>1</sup> <http://woordcombinaties.ivdnt.org/> Website last visited [05/05/2020].



#### 4 Comparing CROATPAS with other Resources for Croatian Verbs

As far as Croatian verbs are concerned, the main currently available resources are: 1) the e-Glava<sup>2</sup> Valency Database of Croatian Verbs (Birtić et al. 2017); 2) the Croatian Valence Lexicon of Verbs<sup>3</sup> (CroVallex, Mikelić Preradović et al. 2009); 3) the Croatian Derivational Lexicon<sup>4</sup> (CroDeriV, Šojat et al. 2014; Filko et al. 2019) and 4) the Nooj Croatian Dictionary of Aspectual Derivatives<sup>5</sup> (Kocijan et al. 2018; Šojat et al. 2018).

The e-Glava project focuses on a sample of 57 psychological verbs, whose senses have been extracted from pre-existing Croatian dictionaries and linked to corpus-based examples providing evidence of their valency patterns. For each verb sense, arguments are described on a morphological, syntactic and semantic level. In terms of morphological and syntactic description, e-Glava and CROATPAS encode similar information. However, the two resources differ most when it comes to semantics. As a matter of fact, e-Glava does not provide English sense descriptions nor Semantic Type labels, but only monolingual sense-based argument periphrases enriched with semantic categories from a small non-hierarchical set. For instance, the nominative subject of the first pattern of the verb *bojati se* (English, *to be afraid*) is described as *onaj koji osjeća strah* (English, *the one who feels fear*) and is enriched by the categories *živo, osoba, skupina ljudi, životinja* (English, *animate, person, human group, animal*).

The first project aimed at building a valency lexicon of Croatian verbs was CroVallex, a resource combining valency theory and frame semantics and containing approximately 1800 high-frequency verbs. In CroVallex, each verb entry is linked to as many valence frames as its number of senses. In each frame, argument slots are labelled with deep role labels (e.g. AGT for Agent, RESL for Result) and morphosyntactic numeric indexes encoding case markings (an idea we borrowed and adapted to the needs of CROATPAS). Overall, these features make CroVallex a highly specialised resource for competent users, primarily aimed at linguistic research – all the more so since it is entirely monolingual.

As for CroDeriV, this resource is centred around Croatian derivational morphology. At the moment, CroDeriV 1.0 consists of an online morphological database containing data about the morphological structure and derivational relatedness of approximately 14,500 Croatian verbs (Šojat et al. 2014). A redesign of the database and its online query interface is currently on-going to include also non-verbal lemmas (Filko et al. 2019). CroDeriV is the perfect tool to learn more about how Croatian verbs can be decomposed into lexical, derivational and inflectional morphemes, and how those verbs belonging to the same derivational families are morphologically connected with one another.

The same could be said for the Nooj Dictionary of Aspectual Derivatives, an on-going project aimed at creating an online database to investigate verb derivation chains and the affixation mechanisms involved in Croatian verb derivation. The database currently contains approximately 4,000 entries and is able to recognise over 377,603 inflected forms.

Overall, we can say that all the above-mentioned resources bear some similarity with the CROATPAS project, but they either tend not to focus on potential accessibility problems by non-native users without a background in linguistics or to concentrate on different linguistic aspects, such as derivational morphology. Therefore, we can say that when compared to these resources CROATPAS stands out for its utmost focus on verb polysemy, its bilingual nature and its compatibility with the other CPA-inspired projects, thus positioning itself as a complementary resource to the existing ones.

#### 5 Concluding Remarks

In this paper, we have introduced CROATPAS (CROATian Typed Predicate Argument Structures), a digital semantic resource currently containing 101 Croatian verb entries linked to 457 corpus-derived semantically labelled valency structures (i.e. *patterns*) and over 22,000 annotated corpus lines (Marini & Ježek 2019 & 2020).

CROATPAS is tailored for investigating verbal polysemy, since each pattern it contains is linked to a different verb sense, but it allows for a variety of different applications. After elaborating on the resource's theoretical underpinnings and providing some pattern examples (see § 2), we focused on its potential as a language teaching tool (see § 3).

As a matter of fact, once equipped with a SemType query option, CROATPAS could already be used in vocabulary teaching activities to access the most important lexical items pertaining to the different semantic areas connected to its SemType inventory (e.g. [Food], [Beverages], [Animals]s, etc.). Moreover, its user-friendly rendering of case markings and argument colour-coding strategy allows for a gentle introduction to the concept of case.

Finally, after comparing it with the other currently available resources for Croatian verbs (see § 4), we can say that CROATPAS is a faceted resource, which stands out for its user-friendly interface, bilingual nature and focus on verb semantics and polysemy. In time, it could become a truly useful tool for the teaching of Croatian as a Foreign Language.

#### 6 References

- Baisa, V., Tiberius, C., Ježek, E., Colman, L., Marini, C. & Romani, E. (2020). Skema: A new tool for corpus-driven lexicography. In *Proceedings of the 19<sup>th</sup> EURALEX International Congress*. Alexandroupolis, Greece.
- Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zenčević, V. & Znika, M. (1997). *Hrvatska Gramatika*. Zagreb: Školska knjiga.
- Baroni, M., Kilgariff, A. (2006). Large Linguistically Processed Web Corpora for Multiple Languages. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy.
- Birtić, M., Brač, I. & Runjaić, S. (2017). The Main Features of the e-Glava Online Valency Dictionary. In *Proceedings of*

<sup>2</sup> <http://valencije.ihj.hr/> Website last visited [05/05/2020].

<sup>3</sup> <http://theta.ffzg.hr/crovallex/> Website last visited [05/05/2020].

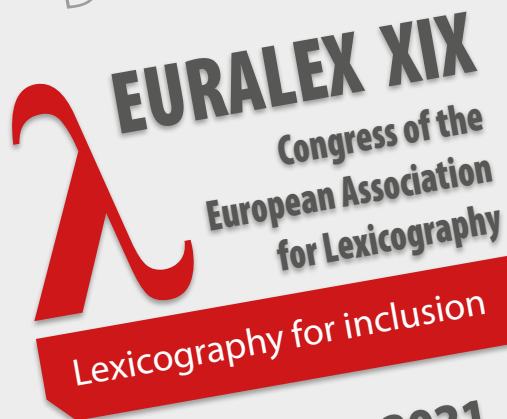
<sup>4</sup> <http://croderiv.ffzg.hr/> Website last visited [05/05/2020].

<sup>5</sup> <https://vidski-parnjaci.herokuapp.com/> Website last visited [05/05/2020].



- the 5th eLex conference - *Electronic lexicography in the 21st century*. Leiden, Netherlands.
- Brown, H.D., Lee, H. (2015). *Teaching by Principles: An Interactive Approach to Language Pedagogy*. Pearson.
- Chan, T., Liou, H. (2005). Effects of Web-based Concordancing Instruction on EFL Students' Learning of Verb – Noun Collocations. In *Computer Assisted Language Learning*, 10(3), pp. 231-250.
- Colman, L., Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In *Proceedings of the 8th EURALEX International Congress*. Ljubljana, Slovenia.
- Filko, M., Šojat, K. & Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In *Proceedings of the 2nd International Workshop on Resources and Tools for Derivational Morphology*. Prague, Czech Republic.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.
- Hanks, P., Ježek, E. (2008). Shimmering lexical sets. In *Proceedings of the 13th EURALEX International Congress*.
- Hanks, P., Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. In *Revue française de linguistique appliquée*, 10 (2), pp. 63-82.
- Hudeček, L., Mihaljević, M. (2017). The Croatian Web dictionary project - Mrežnik. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Brno, Czech Republic; Leiden, Netherlands.
- Ježek, E. (2016). *The lexicon: An introduction*. Oxford: Oxford University Press.
- Ježek, E. (2019). Sweetening Ontologies Cont'd: Aligning bottom-up with top-down ontologies. In *Proceedings of CREOL 2019*. Graz, Austria.
- Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the 9th conference on International Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.
- Kilgariff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*. Barcelona, Spain.
- Kilgariff, A., Baisa, V., Busta, J., Jakubicek, M., Kovár, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. In *Lexicography*, 1(1), pp. 7-36.
- Kilgariff, A., Marcowitz, F., Smith, S. & Thomas, J. (2015). Corpora and Language Learning with the Sketch Engine and SKELL. In *Revue française de linguistique appliquée*, 20(1), pp. 61-80.
- Kocijan, K., Šojat, K. & Poljak, D. (2018). Designing a Croatian Aspectual Derivatives Dictionary: Preliminary Stages. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*. Santa Fe, USA.
- Lee, H., Warschauer, M. & Lee, J.H. (2018). The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-analysis. In *Applied Linguistics*, pp. 1-34.
- Liu, D., Jiang, P. (2009) Using a Corpus-Based Lexicogrammatical Approach to Grammar Instruction in EFL and ESL Contexts. In *The Modern Language Journal*, pp. 61-78.
- Ljubešić, N., Erjavec, T. (2011). hrWaC and slWaC: Compiling Web Corpora for Croatian and Slovene. In *Text, Speech and Dialogue*, Lecture Notes in Computer Science, Springer.
- Marini, C., Ježek, E. (2019) CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian. In *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it)*. Bari, Italy.
- Marini, C., Ježek, E. (2020). Annotating Croatian Semantic Type Coercions in CROATPAS. In *Proceedings of the Sixteenth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. France, Marseille.
- Mikelić Preradović, N., Boras, D. & Kišiček, S. (2009). CROVALLEX: Croatian Verb Valence Lexicon. In: *Proceedings of the 31st International Conference on Information Technology Interfaces*. Zagreb, Croatia.
- Mikelić Preradović, N., Posavec, K. & Unić, D. (2019). Corpus-Supported Foreign Language Teaching of Less Commonly Taught Languages. In *International Journal of Instruction*, 12(4), pp. 335-352.
- Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. In *Applied Linguistics*, 24(2), pp. 223-242.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge: The MIT Press.
- Pustejovsky, J. (1998). The semantics of lexical underspecification. In: *Folia Linguistica* (32).
- Pustejovsky, J., Ježek, E. (2008). Semantic Coercion in Language: Beyond Distributional Analysis. In *Italian Journal of Linguistics*, 20, pp. 181-214.
- Šojat, K., Srebačić, M., Tadić, M. & Pavelić, T. (2014). CroDeriV: a new resource for processing Croatian morphology. *Proceedings of the 9th conference on International Language Resources and Evaluation (LREC)*. Reykjavik, Iceland.
- Šojat, K., Kocijan, K. & Filko, M. (2018). Processing Croatian Aspectual Derivatives. In *Formalizing Natural Languages with NooJ 2018 and Its Natural Language Processing Applications. Communications in Computer and Information Science*.
- Tadić, M., Brozović-Rončević, D. & Kapetanović, A. (2012). Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age. In: *META-NET White Paper Series*, Rehm G. & Uszkoreit, H. (eds.), Springer: Heidelberg, New York, Dordrecht, London.
- Ward, M. (2016). CALL and less commonly taught languages – still a way to go. In *CALL communities and culture – short papers from EUROCALL 2016*.
- Croatian Valence Lexicon of Verbs. Accessed at: <http://theta.ffzg.hr/crovallex/> [05/05/2020].
- Croatian Derivational Lexicon. Accessed at: <http://croderiv.ffzg.hr/> [05/05/2020].
- E-Glava Valency Database of Croatian Verbs. Accessed at: <http://valencije.ihjj.hr/> [05/05/2020].
- NooJ Croatian Dictionary of Aspectual Derivatives. Accessed at: <https://vidski-parnjaci.herokuapp.com/> [05/05/2020].
- Woordcombinaties. Accessed at: <http://woordcombinaties.ivdnt.org/> [05/05/2020].





**7-9 September 2021**  
Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

**Index of Authors**



**Roman****A**

Abel A.	133
Ahmadi S.	63
Aldea M.	275
Andrisani A.	515
Arhar Holdt Š.	41

**B**

Baisa V.	523
Bajčetić L.	73
Barbu A.M.	363
Barrios M. A.	13
Beatrice B.	515
Bilińska-Brynk J.	465
Bobrova, O.B.	247
Bronikowska R.	471
Bueno Ruiz P.J.	333
Buxton C.	503

**C**

Cabezas-García M.	405
Caruso V.	515
Colman L.	523
Contento F.	515

**D**

Dalpanagioti Th.	439
Declerck T.	73
De Tommaso Z.	515
DiMuccio-Failla P.	285
Dolar K.	23
Ducassé M.	81

**F**

Farina A.	371
Fernández-Pampillón A.M <sup>a</sup>	193
Ferrara F.	515
Flinz C.	371
Fournier P.	343

**G**

Galleron I.	393
Gasparini N.	23
Gavriilidou Z.	351
Giacomini L.	285
Giouli V.	91, 263
González M.	109, 503
Grønvik O.	321
Grosse J.	101, 109

**H**

Hajič J.	387
Heid U.	227

**J**

Ježek E.	523, 529
Ježovnik J.	31

**K**

Kallas J.	215
Karasimos A.	305
Kenda-Jež K.	31
Køhler Simonsen H.	183, 509
Koppel K.	215

Kosem I.	41
Kouvara E.	109
Kruse T.	227
Kudashev I.S.	235

**L**

Langemets M.	215
Lanzi E.	285
Latrache R.	343
Lazić Konjik I.	479
León-Araúz P.	405
Leroyer P.	183
L'Homme M.-C.	415
Logar N.	41
Lohk A.	119
Lupu I.	363

**M**

Majdak M.	471
Malli M.	295
Manolessou I.	305
Margalitadze T.	255
Marini C.	523, 529
Markantonatou S.	163, 295, 493
Márquez Cruz M.	193
McCrae J.	63
McGillivray B.	51
Melissaropoulou D.	305
Menniti A.	515
Mexa M.	163
Mikulová M.	387
Milenković A.	479
Minos P.	493
Mitits L.	351
Miyata R.	171
Miyoshi K.	315
Monti J.	515
Motsiou E.	449

**N**

Nesi H.	51
Nimb S.	63

**O**

Ore C.-E.	321
-----------	-----

**P**

Paulsen G.	119
Pavlidis G.	493
Petersson S.	381
Pilitsidou V.	263
Pori E.	41

**R**

Rodek E.	465
Romani E.	523
Rundell M.	51

**S**

Saurí R.	101, 109, 503
Semenova O.V.	235
Sidiropoulos N.F.	91
Şkirmante K.	457
Škofic J.	31
Sköldberg E.	381
Sørensen N.	63



Stamou V.	295
Steffens M.	23
Štěpánková B.	387
Stincone C.	393
Stoica-Dinu O.	363
Sugino H.	171
Süle K.	51
Sviške S.	457

**T**

Takorou P.	295
Tavast A.	215
Teleoacă D.L.	363
Tenieshvili A.	485
Tiberius C.	523
Toraki K.	493
Toroipan T.	363

**V**

Vacalopoulou A.	427, 493
Vainik E.	119

**W**

Wieczorek A.	471
Williams G.	393

**X**

Xylogianni A.	295
---------------	-----

**Z**

Żółtak M.	471
-----------	-----

**Greek****I**

Ιορδανίδου Α.	151
---------------	-----

**Ξ**

Ξυδόπουλος Ι. Γ.	141
------------------	-----

**P**

Ροντογιάννη Α.	203
----------------	-----

**X**

Χριστοπούλου Κ.	141
-----------------	-----













**EURALEX XIX**

Congress of the  
European Association  
for Lexicography

Lexicography for inclusion

**7-11 September 2021**

Virtual

[www.euralex2020.gr](http://www.euralex2020.gr)

Index

Form

Definition

Graphemics

Context

Library

Understand

Translate

Spoken

Semantics

Order

Orthography

Publishing

Spelling

Ex

Sylla

Pragmatics

Origin

gmentation

Reference