Speech Idioms Etymology Glossary NLP Lemma Meaning Dictionary Word Lexicon Definition Corpora Pronounciation Headword amples Entry Lexicology Dictionary Use Lexical Resources cography tain ions ologisms gin ble

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Proceedings Book
Volume 2

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-2-2

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2021 Edition

In Memory of
Alain Rey
(1928-2020)

SPONSORS

+ Μόρ Φω Σn

Το κόσμημα είναι ασημένιο νόμισμα Μαρώνειας 4ου - 3ου π.Χ. αιώνα

ΕΤΑΙΡΕΙΑ ΑΞΙΟΠΟΙΗΣΗΣ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗΣ ΠΕΡΙΟΥΣΙΑΣ
ΔΗΜΟΚΡΙΤΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΘΡΑΚΗΣ
Α.Φ.Μ. 094195820

ΔΠΜΣ Εξειδίκευση στις ΤΠΕ και Ειδική Αγωγή:
Ψυχοπαιδαγωγική της ένταξης

Hornby
A. S. Hornby Educational Trust

Programme Committee

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

Reviewers

Andrea **Abel**
Arleta **Adamska-Sałaciak**
Anna **Anastassiadis**
Battaner **Arias**
Xavier **Banco**
Gilles-Maurice **de Schryver**
Janet **DeCesaris**
Ioannis **Deligiannis**
Anna **Dziemianko**
Angeliki **Efthymiou**
Asimakis **Fliatouras**
Thierry **Fontenelle**
Angeliki **Fotopoulou**
Zoe **Gavriilidou**
Alexander **Geyken**
Rufus **Gouws**
Sylviane **Granger**
Oddrun **Grønvik**
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 **Marello**
Tinatin **Margalitadze**
George **Mikros**
Maria **Mitsiaki**
Rosamund **Moon**
Argyro **Moustaki**
Magali **Paquot**
Stellios **Piperidis**
Natascia **Ralli**
Michael **Rundell**
Max **Silbertzein**
Elsabe **Taljard**
Carole **Tiberius**
Lars **Trap-Jensen**
Anna **Vacalopoulou**
Geoffrey **Williams**
George **Xydopoulos**
Tanara **Zingano Kuhn**

Table of Contents

Foreword

Given the unprecedented circumstances caused by COVID-19, the XIX Euralex, Congress organized by SynMorPhoSe Laboratory of the Department of Greek of Democritus University of Thrace was initially postponed and finally held online from the 7th till the 9th September 2021.

The motto of Euralex XIX was *Lexicography for Special Needs.* Even though the number of disabled people is rapidly increasing worldwide, modern lexicography has just started to address the needs of disabled or functionally diverse people in order to increase their accessibility and e-inclusion in lexicographic products. This congress aimed to make a strong statement and incite lexicographers to start thinking about how to provide more accessible dictionaries.

The second volume of Euralex XIX Proceedings is dedicated, as a mark of esteem for his scientific career in lexicography, to the memory of Alain Rey, who passed away in October 2020. Alain Rey was the editor-in-chief at French dictionary publisher *Dictionnaires Le Robert* from 1967 until his death. He supervised the publication of many dictionaries under the Le Robert trademark: the *Petit Robert* (1967); the *Micro Robert*, a pocket dictionary; the *Petit Robert des noms propres* (1974), a guide to proper names; the *Dictionnaire des expressions et locutions* (1979), a dictionary of phrases and expressions; the *Grand Robert de la langue française*, a nine-volume work (1985); an updated version, the *Nouveau Petit Robert de la langue française* (1993), and the *Dictionnaire historique de la langue française* (1992). In 2005, he published the *Dictionnaire culturel en langue française*.

This volume includes the papers of four Invited Speakers, 17 papers, 2 poster papers and 1 software demonstration. All submissions have been blind-reviewed by two independent reviewers. In case of doubt, a third independent opinion was asked. As was the case with the 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

The Organizing Committee would like to thank all authors for submitting their papers in order to be included in this volume and all the esteemed colleagues who accepted to review the papers.

As the Chair of the Congress, I would like to acknowledge the precious work done by the members of the Organizing Committee, Elina Chadjipapa, Chryssa Dourou and Stavroula Mavrommatidou, who joined efforts with me to make this second volume of EURALEX 2020 Proceedings possible.

Zoe Gavriilidou

Chair of Euralex XIX International Conference,

August 2021

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Invited Speakers

Pour un Dictionnaire de Familles d'unités (sous-)lexicales*

Anastassiadis-Symeonidis A.

Université Aristote de Thessaloniki
ansym@lit.auth.gr

Résumé

Notre recherche concerne la notion de famille de mots dans une perspective de rédaction d'un *Dictionnaire de Familles d'unités lexicales et de leurs parts* en ligne. Pour ce faire, nous procédons à une définition de la notion de famille plus contrainte, fondée sur la cohérence sémantique, partagée par tous ses membres, en suivant la théorie de Morphologie Constructionnelle de Corbin (1987 et 1999). Dans ce but, nous passons en revue les notions d'homonymie, de polysémie et, avant tout, de transparence sémantique, et nous présentons la forme à six champs que devrait avoir chaque article de ce dictionnaire : 1. les préfixés, 2. les suffixés, 3. les composés monolexicaux, 4. les composés polylexicaux et les phrases figées, 5. les conversés, 6. les mots en relation formelle et sémantique mais pas constructionnelle. La notion de famille de mots ainsi définie est utile pour la linguistique théorique, la psycholinguistique, la terminologie, la lexicographie et la didactique d'une langue comme langue maternelle, seconde ou étrangère, ou langue d'héritage.

Mots-clés : Étymologie (populaire); derivation; composition; conversion; figement; transparence sémantique; homonymie; polysémie

1. Introduction

1.1 Généralités

Le terme *famille de mots* est une notion familière véhiculant pourtant des côtés obscurs tant en ce qui concerne le contenu que l'utilité. Selon Triantafyllidis *et al*. (1941/1978 § 253, note 1) la notion du terme *famille de mots* n'est pas définie de manière rigoureuse, et c'est pour cette raison qu'il nous a paru nécessaire de procéder à son investigation. Tout d'abord, rien que l'emploi du terme de *famille de mots*, qui est un hyponyme du concept des champs lexicaux,[1] est une preuve que le lexique est structuré, notion tant contestée par d'autres[2]. Selon Corbin (1987: 88-89), « la morphologie historique a légué à la description synchronique la notion métaphorique de 'familles' de mots. On entend par là un ensemble de termes liés par l'étymologie et, plus accessoirement, le sens. […] Fondée sur l'étymologie, elle a le tort de rassembler, en synchronie, soit des termes formellement très éloignés qui gardent une vague parenté sémantique […], soit des termes formellement proches qui n'ont plus du tout de parenté sémantique synchronique. […] On se contente de dire que des mots sont 'de la même famille', sans tenir compte des diverses relations de 'parenté' à l'intérieur de la famille ». Par ailleurs, le terme de *famille de mots* se rapporte, en général, au nombre des membres d'un ensemble lexical liés entre eux par des relations morphologiques et sémantiques, parce qu'ils contiennent un morphème lexical commun. Nous constatons, donc, une ambigüité, puisque le terme de famille de mots tantôt est rattaché à la linguistique diachronique et tantôt à la linguistique synchronique. Il n'est, donc, pas étonnant de trouver, dans la bibliographie, le terme de *famille étymologique de mots* (Picoche 1977: 68 ; Dorbarakis 1993/1999). Sur ce point, il est à noter que notre analyse se place en synchronie, et elle est fondée sur des critères lexico-sémantiques : Par conséquent, nous sommes d'accord avec Triantafyllidis *et al.* (1941/1978 § 253)[3] qui soutiennent que, dans la famille de l'unité lexicale καράβι 'bateau', le mot καραβίδα 'écrevisse' ne trouvera pas de place, bien que le second entretienne des relations étymologiques avec le premier. Nous sommes d'accord aussi avec Bybee (1985; 1988: 127; 1995: 428) qui soutient que, dans le lexique, les mots sont reliés à d'autres mots à travers des ensembles de connexions entre des caractéristiques phonologiques et sémantiques identiques.

Dans la première partie, nous allons présenter à grands traits les dictionnaires existants et le dictionnaire que nous proposons ; dans la deuxième partie, nous allons examiner le contenu d'une famille en donnant des exemples précis ; dans la troisième partie, nous essaierons de répondre à la question comment sélectionner l'entrée lexicale, et, à la dernière partie, nous nous poserons la question concernant l'utilité d'un dictionnaire tel qu'il est proposé ici.

1.2 Les Dictionnaires Existants

* Je remercie Madeleine Voga, maître de conférences à l'Université Paul Valéry Montpellier III, pour la lecture attentive de mon texte.

[1] En suivant Picoche (1977: 68), nous ne sommes pas d'accord avec le point de vue selon lequel il n'y a pas de différence entre les champs lexicaux et les champs sémantiques (Xydopoulos 2008: 303), puisque, d'une part, il y a des champs lexicaux non sémantiques, ex. un ensemble de mots ou lexèmes qui riment, et, de l'autre, des champs sémantiques non lexicaux, ex. un paradigme flexionnel, où le sens lexical du thème flexionnel n'est pas pris en compte.

[2] Selon la doxa, la Grammaire d'une langue est structurée, contrairement au Lexique, qui est le lieu des exceptions.

[3] Par contre, la famille de καράβι 'bateau' comprendra des mots comme καραβάκι 'batelet', καραβιά 'batelée', καραβίσιος 'de navire', καραβόγατος 'chat de bateau', καραβοκύρης 'capitaine', καραβόπανο 'toile pour bateau', καραβόσκοινο 'amarre', καραβόσκυλος 'chien de bateau', καραβοτσακισμένος 'échoué' (Triantafyllidis *et al*. 1941/1978 § 253), σαπιοκάραβο 'rafiot'.

Dans ce travail, nous proposerons la confection d'un *Dictionnaire de Familles d'unités (sous-)lexicales*, puisque nous avons constaté qu'un tel outil si utile est absent pour le grec moderne (GM). Nous allons, toutefois, présenter brièvement cinq dictionnaires qui s'en rapprochent.

Pour le GM, il y a deux dictionnaires étymologiques avec des familles de mots qui se différencient des autres dictionnaires étymologiques par le fait que l'entrée lexicale est accompagnée de sa famille de mots. Il s'agit du dictionnaire de Dagkitsis (1978-1984)[4] et de Dorbarakis (1993/1999: ια΄). Un exemple de ce dernier : l'entrée θέα 'vue' comprend θέαμα 'spectacle', θεαματικός 'spectaculaire' ; θεατής 'spectateur' ; θέατρο 'théâtre' etc. Pourtant, l'entrée λέγω 'dire' comprend les mots λέξη 'mot', λεκτικός 'verbal' mais pas λόγος 'discours'. De même, l'entrée τρώγω 'manger' comprend les mots τρωκτικά 'rongeur', τρωξαλλίδα 'espèce de sauterelle', mais le mot φαγητό 'nourriture', bien que φαγ- soit le thème supplétif de τρώγω 'manger', constitue une entrée indépendante. De même, φαγάδικος 'qui a tendance à être gros mangeur' avec les mots φαγάνα 'pelleteuse' et φαγανός 'glouton' font partie de l'entrée εσθίω 'manger', qui n'existe plus en GM.

L'*Antilexikon* ou *Onomastikon du grec moderne* (Vostantzoglou 1962) est un dictionnaire onomasiologique, qui comprend aussi les dérivés et composés d'une entrée lexicale. Très riche, il comprend environ 65.000 unités lexicales réparties en 1500 concepts, il est indubitablement un outil très utile de plusieurs points de vue et, en particulier, pour la didactique de la langue. Son intérêt réside aussi dans le fait que son auteur applique le dégroupement des sens, bien que parfois les renvois à des concepts soient en nombre excessif. Par exemple pour le mot κρίσις[5] il prévoit sept concepts : 1. γνώμη 'opinion', 2. απόφανση 'jugement', 3. ευθυκρισία 'rectitude de jugement', 4. δικαστική απόφαση 'décision juridique, verdict', 5. περίοδος δυσκολιών 'crise, période de difficultés', 6. περίοδος κινδύνων 'crise, période dangereuse', 7. ετυμηγορία 'verdict'. Le but du dictionnaire étant différent du nôtre, nous soulignons que les sens 4 et 7 sont identiques, bien qu'ils figurent sous des concepts différents (le sens 4 sous le concept DÉCISION et le sens 7 sous celui de PROCÈS), parce que le verdict est, d'une part, une espèce de jugement et, de l'autre, il appartient aussi au concept de procès devant un tribunal.

Χτίζω λέξεις '*Construire des mots*' (Iordanidou et Pantazara 2010) comprend 6.850 mots construits (dérivés et composés) (p. 7), répartis dans 220 entrées dont 140 sont des premiers constituants (préfixes, formants en première position et premiers composants) et 80 des deuxièmes constituants (suffixes, formants en deuxième position et deuxièmes composants) (p. 5). Il s'agit d'un dictionnaire d'orientation synchronique. Hormis le nombre réduit des mots construits présentés dans ce livre, par ailleurs très utile, il aurait été souhaitable que la polysémie se réduise à ses vraies dimensions,[6] c'est-à-dire, selon nous, il aurait fallu préférer l'homonymie quand le cas se présente, par exemple à l'entrée παρα- (p. 112). Bien qu'il soit dit que cette forme corresponde à deux sens dont chacun provient d'un mot différent, l'entrée est traitée comme polysémique : 1. Forme provenant de l'adverbe πάρα 'très', utilisée pour renforcer le sens, ex. παρα-τρώω 'manger trop, se gaver', 2. Forme provenant de la préposition παρά 'à côté', utilisée pour désigner que (i) qch se passe à côté ou en parallèle, ex. παρα-θαλάσσιος 'au bord de la mer, sur le littoral', (ii) qch est en dehors des limites admissibles, ex. παρακρατικός 'paragouvernemental', et (iii) qch exprime la privation d'une qualité, ex. παράνομος 'illégal'. Dans ces cas-là, nous proposons l'application du dégroupement des sens, étant donné qu'il s'agit d'homonymes. Un problème analogue concerne le segment final d'un mot qui est traité de façon unifiée, bien qu'il puisse correspondre à deux structures différentes du mot construit ; par exemple les mots terminés en -ισμός peuvent être le résultat de deux constructions morphologiques différentes, car la finale -ισμός correspond à deux suffixes bien différents : (i) soit -ισμός '-isme', construisant des substantifs dénominaux, ex. εγω-ισμός 'égo-ïsme', (ii) soit -μός '-ation', construisant des substantifs déverbaux, ex. συλλαβισ-μός 'syllab-ation'. Pourtant, dans ce dictionnaire, les deux mots sont traités comme des dérivés du suffixe -ισμός '-isme'. De même, le segment final -ιστής correspond à deux suffixes bien différents : (i) -ιστής '-iste', ex. βουδ-ιστής 'bouddh-iste', et (ii) -τής '-eur', ex. φροντισ-τής 'répétit-eur'. Pourtant, dans ce dictionnaire, les deux mots sont traités comme des dérivés du suffixe -ιστής '-iste'. Le problème de fond derrière ces cas traités comme des cas de polysémie, qui donne le pas à la forme au détriment du sens, -puisque les exemples donnés ne correspondent pas toujours à l'instruction sémantique du suffixe proposée- consiste à empêcher le locuteur de suivre le fil conducteur sémantique et rend ces descriptions inadéquates en grande partie, tant pour la recherche psycholinguistique que pour la didactique d'une langue.

Enfin, le *Dictionnaire des dérivés et des composés du grec moderne* (Babiniotis 2016) comprend 64.000 dérivés et composés, mais ne trace pas, selon nous, malgré son titre, de façon nette les limites entre diachronie et synchronie, puisqu'il examine (2016: 11) les relations étymologiques des mots, d'une part, avec les ομόρριζα 'mots issus de la même racine' et, de l'autre, avec le champ étymologique d'un mot, constitué des dérivés et composés de sa racine ou de son thème.[7] Cependant, la deuxième catégorie inclut des mots ayant des relations morphologiques en synchronie, ex. les mots présentés dans l'entrée αβγό 'œuf', qui, pourtant, n'existaient pas sous la même forme en grec ancien (GA). Inversement, pour les entrées σοβαρός 'sérieux' et σεπτός 'sacré' il n'y a qu'un seul renvoi à l'entrée σέβομαι 'respecter', avec laquelle le seul lien qui existe en synchronie est d'ordre étymologique. De même, parmi les suffixes construisant des adverbes (p. 26) figurent des suffixes appartenant au système constructionnel du grec de siècles antérieurs qui ne sont plus sentis en synchronie par les locuteurs comme des suffixes, ex. -θεν dans μακρό-θεν 'de loin',[8] -δόν, dans σχε-δόν 'presque', -ί, dans

[4] Date de parution du premier volume A-K 1978 et du deuxième Λ-Π 1984. Malheureusement, l'ouvrage reste inachevé.

[5] Le dictionnaire datant de 1962 est rédigé en langue pure (katharevoussa). La langue actuelle (démotique, du peuple) a été instituée comme langue officielle en 1976.

[6] Par ex., nous sommes d'accord que l'entrée σκυλο- 'de chien' (p. 126) se présente comme polysémique : a. référence au chien, ex. σκυλοτροφή 'nourriture pour chiens', b. intensification du sens du 2^{e} constituant, ex. σκυλο-βρομάω 'puer', c. référence à des boîtes de nuit de bas étage, ex. σκυλο-μάγαζο 'boîte de nuit de bas étage', car il serait aisé de passer du sens propre dans (a) aux sens dérivés (b) et (c).

[7] La référence au sens du mot ne fait qu'accroître la confusion entre diachronie et synchronie.

[8] Comme l'instruction sémantique de -θεν 'mouvement d'un lieu' en GA est perdue en synchronie, le mot μακρόθεν 'de loin' n'apparait plus tout seul, mais il fait partie de la phrase prépositionnelle εκ του μακρόθεν 'de loin', où la notion du mouvement d'un lieu est exprimée par la préposition εκ 'de'.

μαζ-ί 'ensemble'. Enfin, il n'y a pas de partie sémantique concernant le GM, les informations sémantiques étant présentées dans la partie étymologique sous la forme de l'évolution sémantique, ce qui a comme corollaire la présence de sens qui ne sont plus en vigueur en GM, dont cependant dépendent, parfois, les relations constructionnelles.

Par ailleurs, la notion de suffixe n'est pas définie de façon contrainte. Selon Corbin (1987: 458), un segment Y d'un mot complexe X peut être listé parmi les entrées lexicales marquées de la catégorie [Affixe] si et seulement si il sert à construire d'autres mots complexes qui entretiennent avec leur base les mêmes relations catégorielles et sémantiques que X avec la sienne. Ce principe permet de compter au nombre des entrées affixales des segments comme -μα dans βλέμμα 'regard' et d'en exclure des segments comme -αρο dans βλέφαρο 'paupière'. Pour Corbin (1987: 89), il ne suffit pas que des mots soient apparentés formellement et sémantiquement pour qu'ils puissent être dérivés l'un de l'autre : encore faut-il que les relations formelle et sémantique puissent être, de façon conjointe, considérées comme régulières. Sinon, on est conduit

(i) à reconnaitre comme suffixes des formes inexistantes, ex. le mot βλέφ-αρο 'paupière' figure parmi les dérivés de βλέπω 'voir' (Babiniotis 2016: 121), sans qu'un suffixe -αρο[9] soit reconnu comme tel en GM ; de même pour le mot χρυσ-άφι 'or', figurant parmi les dérivés de l'entrée χρυσός 'or', sans qu'il existe un suffixe -άφι en GM,[10] ou pour le mot γκαρ-άζ 'gar-age', sans qu'il y ait de suffixe -άζ[11] en GM ; bien que βλέφαρο 'paupière' et χρυσ-άφι 'or' soient formellement et sémantiquement reliables à βλέπω 'voir' et χρυσός 'or' respectivement, on ne les fera pas dériver de ces derniers, car la relation formelle qui les relie à eux ne peut être considérée comme régulière ; et

(ii) à unifier sous la même entrée des suffixes différents (Tableau 227), erreur due à l'homophonie du segment final : en fait, les mots terminés en -ισμός peuvent être le résultat de deux constructions morphologiques différentes (v. *supra*). Pourtant, dans ce dictionnaire aussi, les deux mots sont traités comme des dérivés du suffixe -ισμός '-isme'. De même, le segment final -ιστής correspond à deux suffixes bien différents (v. *supra*).[12] Pourtant, dans ce dictionnaire aussi, les deux mots sont traités comme des dérivés du suffixe -ιστής '-iste' (Tableau 227, entrée -ιστής). Cette unification erronée peut avoir lieu aussi entre un suffixe et un confixe/formant, ex. le segment final -ώδης (Tableau 239) est présenté comme un seul suffixe, tandis qu'il correspond soit au suffixe -ώδης '-eux' dans ακανθ-ώδης 'épin-eux', soit au formant -ώδης 'qui sent' dans ευ-ώδης 'qui sent bon, odoriférant'. Cette unification erronée peut concerner aussi des lexèmes : par exemple l'entrée πέτρα aurait dû correspondre soit au mot-entrée πέτρα 'pierre' dans πετροβολώ 'lancer des pierres', soit au formant-entrée πέτρ- 'rocher' dans πετρολογία 'pétrologie' et au formant-entrée πετρ- 'pétrole' dans πετροδολάριο 'pétrodollar'.

Le fait que l'instruction sémantique des suffixes et des préfixes et le sens des formants ou des lexèmes n'est pas toujours pris en compte sérieusement réduit la valeur de cet outil par ailleurs précieux, avant tout pour la richesse de l'information ainsi que pour le grand nombre de figures permettant la visualisation du chemin constructionnel des dérivés et composés.

1.3 Le Dictionnaire Proposé

Selon nous, ces erreurs ou incohérences sont dues au fait que ces dictionnaires donnent la primauté à la forme au détriment du sens, c'est-à-dire en subordonnant le sens à la forme, au lieu d'être le produit de l'application d'une théorie constructionnelle qui accorde au sens la contribution qui est la sienne. Pour l'usager l'obscurcissement ou l'opacité du sens qui découle du traitement des homonymes en tant que des polysèmes, enlève à ces dictionnaires, en grande partie, leur valeur d'outil opérationnel pour la recherche en psycholinguistique et la didactique d'une langue. Le *Dictionnaire de Familles d'unités (sous-)lexicales* que nous proposons sera fondé sur la théorie constructionnelle de Danielle Corbin (1987/ 1991, 1999), que nous avons adoptée, parce qu'elle associe sens et forme lors de la construction d'un mot. De cette association forme-sens découlent, de façon contrainte, la compositionnalité sémantique et la transparence constructionnelle du mot construit, lequel est défini de façon restrictive, seul moyen pour mettre en place des règles prédictibles et non seulement descriptives. Notre but consiste à créer un outil dont chaque article sera régi par un sémantisme cohérent. Pour y répondre, nous devons adopter l'opération de dégroupement des sens : chaque article ne regroupera que les mots, construits ou pas, qui partagent le même sens avec le mot vedette placé en entrée (le mot-entrée), indépendamment de leur étymologie. Nous serons ainsi conduits à considérer comme homonymes des mots qui sont décrits comme des polysèmes dans les dictionnaires de langue. D'ailleurs, très souvent, ces homonymes participent à des séries constructionnelles différentes, ex. τέλος 1 'fin', qui comprend dans sa famille τελικός 'final' et τελεσίδικος 'irrévocable', et τέλος 2 'taxe', qui comprend dans sa famille τελωνείο 'douane' et υποτελής 'tributaire'.

Cette homonymisation nous oblige à accompagner chaque mot-entrée d'une description sémantique concise, voire par simple synonymie. De même, les entrées de ce dictionnaire seront numérotées et présentées par ordre alphabétique. Mais, comme chaque article regroupera un grand nombre de mots, nous fournirons à la fin de l'ouvrage, pour en faciliter

[9] D'ailleurs, la forme -αρο ne figure pas dans la liste des suffixes (Babiniotis 2016: Tableau 209), contrairement à -άφι (Babiniotis 2016: Tableau 210). Des critères topographiques, ici la fin du mot, appliqués sans être accompagnés de critères sémantiques, ne peuvent qu'aboutir à une confusion extrême en ce qui concerne la notion de suffixe. Dans le cadre théorique de Corbin (1987/1991: 457), les mots βλέφαρο 'paupière' et χρυσάφι 'or' seraient des mots complexes non construits, car tous les éléments qui les composent ne sont pas des entrées lexicales, ici l'élément droit :
[[χρυσ(ός)]N άφι]N ou [[βλέπ(ω)]V αρο]N contrairement à βλέμμα 'regard' [[βλέπ(ω)]V (-μα)AF]N qui est un mot construit. Nous mettons entre parenthèses les suffixes flexionnels, parce qu'ils ne participent pas à la dérivation.

[10] V. pourtant Babiniotis (2016: Tableau 143 et Tableau 210), qui soutient son existence, en ne mentionnant que la catégorie du nom construit, mais sans donner son instruction sémantique.

[11] La forme -άζ < fr. *-age*, bien qu'elle soit décrite comme terminaison de substantifs de provenance française, est traitée sur un pied d'égalité avec les suffixes (Tableau 204).

[12] Un autre cas qui produit de la confusion est le segment final -ωση, qui peut correspondre (i) soit au suffixe -ωση '-ose', qui construit des termes dénominaux, ex. ίν-ωση 'fibr-ose', (ii) soit au suffixe -ση '-ation', qui construit des noms/termes déverbaux, ex. χλωρίω-ση 'javellis-ation', traités tous les deux comme des dérivés de -ωση '-ose' (Tableau 241).

l'utilisation, la liste de tous les mots du dictionnaire par ordre alphabétique, accompagnés du numéro de l'entrée dans laquelle ils sont traités. Enfin, chaque article comprendra six champs, qui seront décrits ci-après (section 2).

Ensuite, nous essaierons de répondre aux questions suivantes, ayant toujours en vue la confection d'un *Dictionnaire de Familles d'unités (sous-)lexicales* : 1. Que doit comprendre une famille d'unités (sous-)lexicales ? 2. Comment est sélectionnée l'entrée lexicale ? 3. Pourquoi la notion de famille d'unités (sous-)lexicales est-elle importante?

2. Que Doit Comprendre une Famille d'Unités (Sous-)Lexicales?

De ce qui vient d'être dit, il ressort que la famille peut être formée à partir de critères morpho-étymologiques[13] ou morpho-synchroniques. Le dernier aspect est plus récent et lié, de manière intrinsèque, à la didactique des langues et à la recherche psycholinguistique. Un autre point concerne la part de la flexion dans la constitution de la famille. En général, dans la bibliographie grecque et étrangère, les formes flexionnelles d'un mot fléchi ne sont pas comprises dans la famille, constituée avant tout sur la base de critères étymologiques. Pourtant, plus récemment, et dans le cadre d'approches liées à la recherche psycholinguistique (de Jong *et al.* 2000) et didactique (Bauer & Nation 1993: 253), les formes flexionnelles comptent parmi les membres de la famille.[14] Selon Nation (2004: 6) la famille inclut les formes fléchies et dérivées en relation étroite,[15] sans pour autant qu'il soit spécifié en quoi consiste cette relation. Pour Schreuder & Baayen (1997: 133) le fait de ne pas inclure dans la famille les formes fléchies est lié à la régularité et à la prédictibilité de leur sens, mais il pourrait aussi avoir trait, selon nous, à la différence des deux processus qui concernent la construction ou pas d'un nouveau lexème et par conséquent d'un nouveau sens. Plus particulièrement, nous avons soutenu (Anastassiadis-Syméonidis 2004) que, d'une part, la flexion ne constitue pas un processus tout à fait régulier ni prédictible et, de l'autre, que la morphologie constructionnelle ne constitue pas un processus plein d'exceptions et sans prédictibilité. En revanche, la construction d'un nouveau lexème, indépendamment de la régularité ou la prédictibilité de sa construction, est liée à un nouveau référent, ce qui exige l'inscription dans la mémoire à long terme de cette relation, contrairement à la construction d'une nouvelle forme d'un mot fléchi, qui, en général, renvoie au même référent.

Toutefois la flexion n'est pas un phénomène unifié : à part les formes régulières (i), il y a (ii) les formes flexionnelles supplétives et (iii) les formes flexionnelles lexicalisés.

(i) Comme la notion de famille interfère avec des questions de taille de la famille morphologique (Dijkstra *et al.* 2005) et de fréquence de tous ses membres, il nous paraît raisonnable d'inclure aussi dans la famille les formes flexionnelles.[16]

(ii) Le lexème des formes supplétives, ex. τρώ(γ)ω–έφαγα 'manger', se présente sous forme de deux ou plus thèmes, qui peuvent en être très éloignés du point de vue de l'ordre alphabétique. Notre critère principal pour la construction d'une famille étant d'ordre sémantique, nous proposons de traiter les formes supplétives sous le même mot-entrée.

(iii) En ce qui concerne les formes lexicalisées, ex. το είναι 'l'être' (en GA infinitif du v. εἰμὶ 'être'), το Πάτερ ημών 'Notre Père' (en grec hellénistique Πάτερ : vocatif du nom Πατὴρ 'Père' et ἡμῶν : génitif pluriel du pronom personnel ἐγὼ 'moi'), elles pourront faire partie de la famille d'une entrée, à condition qu'elles servent sa cohérence sémantique.

Selon la bibliographie grecque et étrangère, la famille d'unités lexicales comprend les dérivés et les composés d'un lexème non construit, présenté sous la forme d'un seul mot, qui constitue l'entrée lexicale.[17] Triantafyllidis *et al.* (1941/1978 § 251) ont raison d'ajouter que la famille comprend des mots apparentés tant savants que non savants, ex. sous l'entrée άνεμος 'vent' sont inclus les mots ανεμόπτερο 'planeur' (mot savant) et απάνεμος 'à l'abri du vent' (mot non savant).

Cette définition générale doit être complétée selon nous, car, du point de vue morphologique,[18] la famille de mots doit comprendre non seulement les mots préfixés (champ 1), les mots suffixés (champ 2), et les mots composés monolexicaux (champ 3), mais aussi les composés polylexicaux et phrases figées (champ 4),[19] où l'entrée lexicale correspond à l'un des composants, car un grand nombre de composés polylexicaux et de phrases figées posséderaient leurs propres représentations sémantiques dans le lexique mental (Schreuder & Baayen 1997: 136), fonctionnant et s'inscrivant comme un tout dans la mémoire. Plus particulièrement, Hay & Baayen (2005) soutiennent le point de vue selon lequel leur haute fréquence aide à leur inscription et traitement holistique dans la mémoire à long terme. De même, Swinney & Cutler (1979) soutiennent que les phrases figées s'inscrivent dans la mémoire sous forme de mots longs. Aussi, dans une recherche psycholinguistique, Anastassiadis-Syméonidis et Voga (2011) ont-elles trouvé que le temps de réaction est plus rapide dans la catégorie des phrases figées opaques à sens compositionnel, ex. έδεσε το γάιδαρό του [il a attaché son âne] 'il a assuré son avenir', par rapport à la même phrase présentée dans un contexte libre, ce qui est interprété comme un indice fort que les phrases figées sont inscrites de façon holistique dans le lexique mental, conformément au modèle de Swinney et Cutler. Par conséquent, vu le fait que, dans la famille de mots, nous incluons aussi les unités polylexicales, nous n'utiliserons pas le terme de *Dictionnaire de famille de mots*, mais, provisoirement, celui de *Dictionnaire de famille d'unités lexicales*.

À part les dérivés (préfixés et suffixés) et les composés (monolexicaux et polylexicaux), la famille d'unités lexicales doit

[13] Par exemple la BDME, qui fournit une représentation linéaire ou sous forme de graphes les familles et sous-familles des mots de l'espagnol, et qui est fondée sur des critères morpho-étymologiques, va servir à la confection du *Nuevo diccionario histórico del español*.

[14] V. aussi Diependaele, Grainger & Sandra (2012: 319). Pour le GM v. Goutsos (2006: 83) pour qui aussi les formes flexionnelles de l'entrée sont incluses dans l'article lexicographique, puisque la famille de mots est définie comme l'ensemble des éléments constitué de l'entrée, ses formes, ses dérivés et composés.

[15] Selon Nation (2004) exemple de famille : ADD ADDED ADDING ADDITION ADDITIONAL ADDITIVE ADDITIONS ADDS.

[16] Contrairement à ce que nous avons soutenu dans Anastassiadis-Syméonidis (2020).

[17] V. par exemple Schreuder & Baayen (1997: 118) "formations in the morphological family of a given noun (the compounds and derived words in which that noun appears as a constituent)".

[18] Pour la présentation des mots construits, nous suivons la théorie de Morphologie Constructionnelle de Corbin (1987/1991 et 1999).

[19] Pour une proposition similaire pour le français v. Schreuder & Baayen (1997: 135), avec l'argument que le français a un nombre réduit de composés monolexicaux. En effet, la plupart de ses composés sont polylexicaux, ex. *chemin de fer*.

comprendre aussi les mots construits par conversion (champ 5), ex. αγάπη 'amour' → αγαπώ 'aimer'. Ces cinq premiers champs constituent la famille morphologique d'unités lexicales, c-à-d. l'ensemble des unités construites ayant la forme d'un seul mot ou pas, partageant le même morphème lexical.[20] Toutefois, il est nécessaire pour nous que la famille d'unités lexicales comprenne aussi des unités lexicales entretenant en synchronie avec le mot-entrée une relation formelle et sémantique mais pas constructionnelle (champ 6), ex. l'entrée *πολίτης* 'citoyen' doit comprendre le mot πολιτικάντης 'politicard, politicien véreux', ou l'entrée βλέπω 'voir' doit comprendre le mot βλέφαρο 'paupière', car, comme il en résulte de l'examen du *Dictionnaire inverse du grec moderne* (Anastassiadis-Syméonidis 2002), il n'y a pas de suffixe -άντης ou -αρο en GM. Plus particulièrement, dans le champ 6, nous allons inclure (i) des mots qui entretiennent avec le mot-entrée une relation d'étymologie populaire, c'est-à-dire des mots pour lesquels les locuteurs, entraînés par un mouvement psychologique impliquant des processus cognitifs, ont trouvé après coup une relation sémantique qu'ils cherchaient à instaurer à tout prix, ex. πολυθρόνα 'fauteuil', (ii) des mots comportant des suffixes du GA qui ne sont ni actifs ni perçus comme tels en GM, ex. στρωμ-νή 'literie et matelas', et (iii) des emprunts directs se terminant en une séquence qui pourrait être prise pour un suffixe, ex. στρωματσάδα 'couchage par terre' < vénit. *stramazzada*. C'est la raison pour laquelle, il nous paraît évident qu'il ne faut pas utiliser le terme de *famille morphologique d'unités lexicales*, mais le terme plus général de *famille d'unités lexicales*.

En ce qui concerne l'étymologie populaire, qui est bien établie dans le lexique général, les relations de ces unités lexicales avec les autres membres de leur famille, en synchronie, ne sont pas de nature véritablement étymologique -c'est pour cela qu'on parle d'étymologie populaire- et peuvent ne pas être morphologiques, c'est-à-dire qu'elles peuvent être pseudo-constructionnelles ; ce qui ne les empêche pas, pourtant, d'être représentées au sein de la famille d'unités lexicales, à condition qu'un rapport existe sur le plan sémantique.[21]

Examinons les exemples suivants : πολυθρόνα 'fauteuil', κορακιάζω 'crever de soif' et στρωματσάδα 'couchage par terre'. Plus concrètement, bien que πολυθρόνα 'fauteuil' du point de vue étymologique est un emprunt direct à l'italien *poltrona* 'fauteuil' (*DGS* 1998), le mot sera présent dans les entrées πολύς 'beaucoup' et θρόνος 'trône'[22] dans la catégorie (3) des composés monolexicaux. De même, dans l'entrée κοράκι 'corbeau' le verbe κορακιάζω 'crever de soif' sera inscrit à la catégorie des suffixés (2),[23] bien qu'il vienne du turc *kurak* 'sec' (*DGS* 1998). Enfin, en synchronie, le mot στρωματσάδα 'couchage par terre', bien qu'il s'agisse d'un emprunt direct au vénitien *stramazzada* 'couchage sur un lit de plusieurs personnes', donne l'impression qu'il contient le mot στρώμα 'matelas'+ -άδα 'suffixe construisant des noms dénominaux, ex. βαρκάδα 'promenade en bateau'. Ce qui est commun dans les trois cas cités ci-dessus est que le locuteur, entraîné par un mouvement psychologique d'ordre cognitif, parvient à instaurer après coup une relation sémantique à des formes linguistiques qui ne partageaient pas au départ de lien sémantique ni constructionnel. Dans son effort de rendre le sens transparent, l'esprit humain applique ce mécanisme là où il peut y avoir des cas d'opacité, comme dans le cas d'emprunts directs ou de mots savants (Anastassiadis-Syméonidis 1994: 55, 109; 2003: 56-59).

Les unités lexicales des six champs présentés ci-dessus entretiennent obligatoirement des relations lexico-sémantiques avec le mot-entrée, car la transparence sémantique joue un rôle primordial (v. aussi Bertram *et al*. 2000). Les relations constructionnelles viennent en seconde position, puisque, selon nous, la famille peut comprendre aussi des unités lexicales entretenant avec le mot-entrée des relations pseudo-étymologiques et pseudo-constructionnelles. Il est à noter que Feldman et Pastizzo (2003), en examinant l'interaction entre l'amorce (*prime)*, la cible (*target*) et la transparence sémantique de la famille des mots, adoptent le terme de *taille de la famille transparente*,[24] en proposant (2003: 252) que la question de la transparence sémantique ne doit pas se limiter à la relation entre l'amorce et la cible, car la reconnaissance de la cible est influencée par l'ensemble des membres sémantiquement transparents de sa famille de mots.

Par conséquent, c'est grâce à la transparence sémantique, c-à-d. au degré de connexion sémantique d'une unité construite avec ses constituants (Diependaele, Grainger & Sandra 2012: 322) mais aussi avec la structure de sa construction (Corbin 1987/1991), qu'émergent les effets positifs rapportés dans la recherche de Diependaele, Grainger & Sandra (2012: 319; v. 4.2 ici même).

En ce qui concerne la liberté morphémique, l'entrée peut être un morphème libre, ex. γκολ 'but', libérable,[25] ex. παίζ(ω) 'jou(er)' ou lié (formant ou confixe[26]), ex. -σκόπ(ιο) '-scope'. Par conséquent, étant donné que la macrostructure comprend aussi des unités sous-lexicales à sens lexical, le titre définitif du Dictionnaire proposé serait celui-ci : *Dictionnaire de familles d'unités (sous-)lexicales*.

Enfin, il est à noter que la description des familles des unités (sous-)lexicales dans ce dictionnaire ne constitue pas une représentation fidèle des processus morphologiques mis en jeu tels qu'ils résultent de l'application des Règles de Construction de Mots. Comparons les deux représentations des mêmes unités lexicales :

(i) Représentation de la famille de l'unité lexicale τρώ(γ)ω 'manger'
Entrée τρώ(γ)ω-έφαγα 'manger'.

[20] Schreuder & Baayen expliquent (1997: 121) : "We will use the term *morphological family* to denote the set of words derived from a given stem by means of either compounding (*tablespoon, timetable*) or derivation (*tablet, tabular*)". La différence consiste en ce que nous prenons en compte aussi les unités polylexicales et les conversés.

[21] Le statut exact de cette représentation au sein de la famille a besoin d'être davantage spécifié du point de vue expérimental, mais des premiers résultats nous orientent vers cette option (Anastassiadis-Syméonidis et Voga 2011; Anastassiadis-Syméonidis et Voga 2010).

[22] Par étymologie populaire : < πολύς 'grand' + θρόνος 'trône'.

[23] [[κοράκ(ι)]$_N$ (-ιάζ(ω))$_{aff}$]v.

[24] Transparent family size.

[25] Nous empruntons le terme à Martinet (1979) (v. Anastassiadis-Syméonidis 1986).

[26] Nous empruntons le terme à Martinet (1979).

Préfixés : ξανατρώω 'remanger', άφαγος 'à jeun' etc.
Suffixés : φαγάς 'gros mangeur', φαγάκι 'un petit plat', φάγωμα ' de consommation', φαγωμάρα 'dispute' etc.
Composés (et confixés) monolexicaux : τρωγοπίνω 'banqueter', φαγοπότι 'bombance', καλοφάγωτος [vœu pour bien profiter de ce qu'on mange] etc.
Composés polylexicaux (et phrases figées) : γρήγορο φαγητό 'restauration rapide', έτοιμο φαγητό 'plat préparé' etc.
Conversés : φαγητό 'repas', τρωκτικό 'rongeur'
Relation formelle et sémantique mais pas constructionnelle : φαΐ 'plat', φαγανός 'glouton'.[27]

(ii) Représentation des processus morphologiques liés au verbe τρώ(γ)ω 'manger' :[28]
1a Thème [+continu] τρώ(γ)ω 'manger' → ξανατρώω 'remanger' (préfixation) → τρωκτικός 'rongeur' (suffixation) → τρωκτικό 'rongeur' (conversion) → τρωγοπίνω 'banqueter' (composition)
1b. Thème [-continu] i. φαγ- 'manger' → άφαγος 'à jeun' (préfixation), φαγητός 'qui se mange' (suffixation) → φαγητό 'repas' (conversion), → φαγοκύτταρο 'phagocyte', φαγοπότι 'bombance' (composition), → γρήγορο φαγητό 'restauration rapide', έτοιμο φαγητό 'plat préparé' (composition polylexicale), ii. φαγω- 'manger' → φάγωμα 'consommation', φαγωμάρα 'dispute' (suffixation), καλοφάγωτος (composition)
2. φαΐ 'plat'[29] → φαγάκι 'un petit plat', φαγάς 'grand mangeur' (suffixation)
3. φαγανός 'glouton'.

Par la suite, nous donnons sept exemples représentatifs du *Dictionnaire de familles d'unités (sous-)lexicales*:

[1] Entrée πόλη 'ville'
1) préfixés : πρόπολη 'propolis' etc.
2) suffixés : πολίτης 'citoyen' etc.
3) composés (et confixés) monolexicaux : πολεοδόμος 'urbaniste', πολιούχος 'patron d'une ville', νεκρόπολη 'nécropole', παραγκούπολη 'bidonville', μητρόπολη 'métropole', Νεάπολη 'Néapolis' etc.
4) composés polylexicaux (et phrases figées) : πόλη-κράτος 'ville-état', αιώνια πόλη 'ville éternelle', ιερή πόλη 'ville sainte', πόλη του φωτός 'ville des lumières' etc.
5) conversés : Πόλη 'Constantinople'[30]
6) relation formelle et sémantique mais pas constructionnelle : πολίχνη 'bourg'.

[2] Entrée πολίτης 'citoyen'
1) préfixés : απολίτικος 'apolitique' etc.
2) suffixés : πολιτικός 'politique', πολιτικά 'politiquement', πολιτεύομαι 'faire de la politique', πολιτευτής 'politicien', πολίτευμα 'régime politique',[31] πολιτεία 'cité' etc.
3) composés (et confixés) monolexicaux : πολιτοφυλακή 'milice', πολιτικοποιώ 'politiser' etc.
4) composés polylexicaux (et phrases figées) : πολίτης του κόσμου 'citoyen du monde', πολιτικές επιστήμες 'sciences politiques' etc.
5) conversés : πολιτική 'politique' (nom), πολιτικά 'affaires politiques' etc.
6) relation formelle et sémantique mais pas constructionnelle : πολιτικάντης 'politicard'.

[3] Entrée θέατρο 'théâtre'
1) préfixés : αμφιθέατρο 'amphithéâtre' etc.
2) suffixés : θεατράκι 'petit théâtre', θεατρικός 'théâtral', θεατρικά 'théâtralement', θεατρικότητα 'théâtralité', θεατρινισμός 'théâtralisme, affectation', θεατρινίστικος 'théâtral, affecté' etc.
3) composés (et confixés) monolexicaux : θεατράνθρωπος 'homme de théâtre', θεατρολόγος 'théâtrologue', θεατρόφιλος 'amateur de théâtre', θεατρώνης 'producteur de théâtre', κουκλοθέατρο 'théâtre de marionnettes' etc.
4) composés polylexicaux (et phrases figées) : θέατρο σκιών 'théâtre d'ombres', πειραματικό θέατρο 'théâtre expérimental', θέατρο δρόμου 'théâtre de rue', μαύρο θέατρο 'théâtre noir', θέατρο του παραλόγου 'théâtre de l'absurde' etc.
5) conversés : θεατρολογώ 'faire des recherches en études théâtrales' etc.
6) relation formelle et sémantique mais pas constructionnelle : θεατρίνος 'comédien, cabotin'.

[4] Entrée χρυσός 'or' (nom)
1) préfixés : επίχρυσος 'plaqué or' etc.
2) suffixés : χρυσίζω 'tirer sur le doré' etc.

[27] Bien qu'en GA il y ait un suffixe -νός, ex. δεινός 'effrayant' (Oikonomou 1971 : 397,4), qui, en GM, a pris la forme -ινός, ex. θαλασσινός 'de mer', nous considérons qu'en synchronie l'unité φαγανός 'glouton' n'est pas suffixée.

[28] Le processus (préfixation etc.) est noté à la fin.

[29] Malgré la relation étymologique incontestable (φαΐ 'plat' < φαγεῖν 'manger' (verbe) DGS 1998) et la transparence sémantique, en synchronie, il n'y a pas de relation constructionnelle entre τρώ(γ)ω/φαγ- 'manger' et φαΐ 'plat' ou φαγανός 'glouton'.

[30] Ici le changement de catégorie grammaticale concerne la conversion d'un nom commun en un nom propre.

[31] Comme nous venons de dire, cette classification est différente de la représentation des processus morphologiques résultats de l'application des Règles de Construction de Mots. Par exemple l'entrée πολίτης 'citoyen' ne sert pas de nom de base au mot πολίτευμα 'régime politique', qui est un dérivé du verbe πολιτεύομαι 'faire de la politique'. Ce verbe peut aisément faire partie de la famille de πολίτης 'citoyen', grâce à leur relation sémantique transparente.

3) composés (et confixés) monolexicaux : χρυσοποίκιλτος 'brodé or' etc.
4) composés polylexicaux (et phrases figées) : μαύρος χρυσός 'or noir' etc.
5) conversés : χρυσός 'doré' (adj.) etc.
6) relation formelle et sémantique mais pas constructionnelle : χρυσάφι 'or'.[32]

[5] Entrée βλέπω 'voir'
1) préfixés : προβλέπω 'prévoir' etc.
2) suffixés : βλέμμα 'regard' etc.
3) composés (et confixés) monolexicaux : κρυφοβλέπω 'rencontrer en cachette', πρωτοβλέπω 'voir pour la première fois' etc.
4) composés polylexicaux (et phrases figées) : βλέπω το φως 'voir le jour, naître' etc.
5) conversés : απρόβλεπτο 'imprévu' (nom) etc.
6) relation formelle et sémantique mais pas constructionnelle : βλέφαρο 'paupière'.

[6] Entrée στρώνω 'étendre'
1) préfixés : επιστρώνω 'recouvrir', καταστρώνω 'dresser par ex. un plan' etc.
2) suffixés : στρώση 'couche', στρώμα 'matelas' etc.
3) composés (et confixés) monolexicaux : ασφαλτοστρώνω 'asphalter' etc.
4) composés polylexicaux (et phrases figées) : κοινωνικό στρώμα 'couche sociale' etc.
5) conversés : λιθόστρωτο 'pavé de pierres', πλακόστρωτο 'dallage' etc.
6) relation formelle et sémantique mais pas constructionnelle : στρωματσάδα 'couchage par terre', στρωμνή 'literie et matelas'.

[7] Entrée ορα- 'voir'
1) préfixés : ενόραση 'intuition', προόραση 'don de prévoir' etc.
2) suffixés : ορατός 'visible', όραση 'vision', όραμα 'vision, apparition', ενορατικός 'intuitif', προορατικός 'relatif au don de prévoir', οραματίζομαι 'avoir des visions, rêver à' etc.
3) composés (et confixés) monolexicaux : τηλεόραση 'télévision', πανόραμα 'panorama' etc.
4) composés polylexicaux (et phrases figées) : περιφερειακή όραση 'vision périphérique' etc.
5) conversés : ορατά (nom au pluriel) 'monde visible'
6) relation formelle et sémantique mais pas constructionnelle : -

Comment faudrait-il envisager la variation phonologique, ex. χτένα/κτένα 'peigne' ? En suivant Triantafyllidis *et al.* (1941/1978 § 251), les deux formes vont se présenter dans la même entrée, mais aussi avec renvoi de la forme secondaire à la forme principale. Enfin, chaque composant sera présent dans l'entrée adéquate, ex. le composé monolexical κουκλοθέατρο 'théâtre de marionnettes' sera présent à la fois dans l'entrée κούκλα 'poupée' et dans l'entrée θέατρο 'théâtre', le composé polylexical θέατρο δρόμου 'théâtre de rue' sera présent tant dans l'entrée θέατρο 'théâtre' que dans l'entrée δρόμος 'rue'.

3. Comment Sélectionner la Forme Lemmatique?

De ce qui vient d'être présenté, il en résulte, en ce qui concerne la structure morphologique, que la forme lemmatique peut être :

1) un mot simple, c'est-à-dire non construit, ex. πόλη 'ville', χρυσός 'or',
2) un mot construit, ex. πολίτης 'citoyen',
3) une base non autonome, c'est-à-dire un morphème sous-lexical à sens lexical, ex. ορα- 'voir', εαρ- 'printemps'.

Il est évident que le mot simple peut être un emprunt direct, adapté ou pas au système morphophonologique du GM, ex. γκολ 'but' (γκολάρα 'but réussi', γκολκίπερ 'gardien de but', αυτογκόλ 'but contre son camp' etc.), ρεαλισμός 'réalisme'[33] (ρεαλιστής 'réaliste', ρεαλιστικός 'réaliste', νεορεαλισμός 'néo-réalisme', αντιρεαλισμός 'antiréalisme' etc.).
Comment hiérarchiser ces choix ? Le mot simple sera sélectionné comme mot-entrée, s'il en existe, ex. πόλη 'ville'. Dans le cas où le mot simple n'a pas de relation directe et transparente du point de vue sémantique, on sélectionnera un mot dérivé, ex. πολίτης 'citoyen' et non pas πόλη 'ville' pour des unités lexicales comme πολιτεύομαι 'faire de la politique', πολιτικοποιώ 'politiser', qui ont été construites sur la base πολίτης 'citoyen' et entretiennent avec cette unité lexicale une relation sémantique étroite. L'unité lexicale πόλη 'ville' sera l'entrée pour la famille πρόπολη 'propolis', πολίτης 'citoyen', πολιούχος 'patron d'une ville', ακρόπολη 'acropole', μεγαλούπολη 'grande ville', μητρόπολη 'métropole', Νεάπολη 'Néapolis' etc. Mais l'unité lexicale πολίτης 'citoyen', fera partie de la famille sous l'entrée πόλη 'ville', et, en même temps, elle aura le statut d'une entrée à part entière. Autrement dit, nous avons prévu des entrées qui font le pont comme le mot-entrée πολίτης 'citoyen', dont le rôle est, d'une part, de signaler le lien sémantique avec une entrée et, de l'autre, de former autour d'elle une nouvelle famille d'unités lexicales, dont le but est de présenter les sens d'une manière plus cohérente, tout en réduisant la taille des familles. Et parmi les mots de la même famille on sélectionnera celui qui couvre le champ sémantique le plus large ainsi que le plus grand nombre de processus constructionnels, ex. entre κρίνω 'juger' et

[32] Nous considérons χρυσάφι 'or' l'allomorphe [-savant] de l'entrée χρυσός 'or'.
[33] Le locuteur ne peut assigner aucune structure interne au mot ρεαλισμός 'réalisme', car il ne reconnaît pas de base.

κρίση 'jugement' on préférera κρίνω 'juger'. La cohérence sémantique ainsi proposée contribuera à la transparence sémantique, tandis que la consultation des familles d'unités lexicales deviendra plus pratique, vu leur taille réduite.

Enfin, la forme lemmatique peut prendre aussi la forme de base non autonome, c'est-à-dire d'unité sous-lexicale à sens lexical, appelée par d'autres formant ou confixe,[34] mais qui ne peut pas assumer de rôle syntaxique, par ex. celui de sujet, de complément etc., comme εαρ- 'printemps'. Nous considérons qu'il est nécessaire de créer des familles de sous-unités de ce type, à cause de la survie en grand nombre de radicaux d'origine grecque ancienne.

Une autre question qui se poserait concerne (i) l'homonymie et (ii) la polysémie (métaphore, métonymie, catachrèse etc.). Il s'agit d'un sujet épineux, qui n'a, pour le moment, reçu de réponse satisfaisante ni au niveau théorique ni en lexicographie, bien qu'il ait été l'objet d'étude d'un grand nombre de linguistes théoriciens et de sémanticiens. Étant donné que, selon le point de vue traditionnel du lexique, le sujet de la délimitation entre polysémie et homonymie ne se pose pas, les lexicographes des dictionnaires de langue du GM, avant environ 1970, ne s'en préoccupaient pas. Ce n'est que dans la seconde moitié du 20^{e} s. que la problématique a commencé à se développer au sein de théories linguistiques récentes ainsi que par les théories de linguistique computationnelle et de sémantique cognitive. Il est à noter que le point de vue selon lequel la délimitation entre polysémie et homonymie devrait se baser sur des critères exclusivement étymologiques,[35] -une solution appliquée souvent en lexicographie-, est fondé sur la confusion des outils livrés par l'analyse synchronique et diachronique. Cependant, une proposition fondée sur des critères morpho-sémantiques conduit bien des fois à la distinction effective entre les cas d'homonymie et ceux de polysémie. Plus concrètement, Dubois *et al*. dans le *Dictionnaire du Français Contemporain* (1967) ont, pour la première fois en ce qui concerne la lexicographie française, essayé de distinguer les entrées homophones et homographes en tant d'articles différents que les cas où les mots dérivés conservaient le sens de leur base. Nous considérons qu'il s'agit d'un critère fiable, que nous devrions appliquer au *Dictionnaire de familles d'unités (sous-)lexicales* pour distinguer les cas de polysémie de ceux d'homonymie. Un argument en faveur de cette position vient de recherches récentes en psycholinguistique, par ex. Feldman & Pastizzo (2003: 234), selon lesquelles ce qui compte c'est la transparence sémantique, qui n'est valable que dans les cas des dérivés qui conservent le sens de leur base.[36]

3.1 Homonymie

Il y a homonymie si les sens d'une forme ne sont pas dérivables sémantiquement l'un de l'autre (Corbin 1987/1991: 258). Les homonymes et homographes constitueront des entrées différentes, selon nous, puisque la famille des unités lexicales est fondée sur des critères lexico-sémantiques en synchronie,[37] ex. μέλος 1 'membre' (πολυμελής 'à plusieurs membres', μονομελής 'à un seul membre', διαμελίζω 'démembrer') et μέλος 2 (μελωδία 'mélodie', μελωδικός 'mélodieux', μελωδός 'mélode, chanteur', μελομανής 'mélomane', μελόδραμα 'mélodrame', μελοποιώ 'mettre en musique'). Les distinctions proposées ne coïncident pas nécessairement avec celles d'autres dictionnaires, divergences expliquées d'ailleurs par la linguistique cognitive, qui a mis en doute l'existence de sens bien délimités ainsi que la notion de lexème, ex. chez Vostantzoglou (1962: 952), où sous l'entrée polysémique μέλος 'membre',[38] quatre sens sont distingués ; ou avec celui de Iordanidou (2005: 274), où sous l'entrée polysémique κρίση 'jugement, crise' il y a cinq sens ;[39] dans le *DULG* (2014), où sous l'entrée polysémique κρίση 'crise' il y a quatre sens[40] et chez Babiniotis (2002) où l'entrée polysémique κρίση 'crise' a cinq sens ; de même, chez Babiniotis (2002) sous l'entrée polysémique τέλος 'fin, taxe, but', il y a huit sens et dans le *DULG* (2014), où sous l'entrée polysémique τέλος 'fin, taxe' il y a six sens. Plus particulièrement, l'absence de description sémantique en synchronie chez Babiniotis (2016: 577 et Tableau 115) fait que l'entrée τέλος comprend des dérivées et des composés qui sont en relation sémantique 1. avec le sens 'fin', ex. τελικός (nom) 'finale', 2. avec le sens 'taxe', ex. υποτελής 'tributaire' et 3. avec le sens 'but', ex. τελολογία 'téléologie'.

Par ailleurs, il serait intéressant d'examiner l'entrée κρίνω 'juger, critiquer' chez Babiniotis (2016: 306), où, dans la partie étymologique, est présentée son évolution sémantique ('distinguer', 'décider' et puis 'juger') ; en suivant le renvoi à la figure représentant le schéma constructionnel du v. κρίνω 'juger, critiquer' (Tableau 53), le dérivé κρίση 'jugement' aurait dû être présent comme dérivé de κρίνω 'juger' avec les autres dérivés découlant du même sens de κρίνω 'juger', comme κριτήριο 'critère', κριτής 'juge', κριτικός 'critique' et non pas être présenté comme dérivé de κρίνω, qui n'a pas de sens le rattachant à κρίση 'crise'. Autrement dit, tandis que la transformation ο δικαστής κρίνει 'le juge juge' → η κρίση του δικαστή 'le jugement du juge' est possible, η κρίση της κοινωνίας 'la crise de la société' ne peut être le produit d'aucune transformation impliquant le v. κρίνω 'juger'. Nous sommes en présence d'un cas extrêmement intéressant, qui constitue une violation de l'hypothèse de la connectivité due au contact des langues (Georgakopoulos & Polis 2018: 23) : le mot κρίση 'jugement//crise' a la forme d'un nom déverbal (le suffixe -ση est appliqué au v. κρίνω 'juger'), mais du point de vue sémantique il n'est dérivé du v. κρίνω 'juger' que quand il signifie 'jugement'. L'étymologie peut en fournir l'explication : le mot κρίση 'crise' est un emprunt indirect au français (internationalisme) et pas un mot construit en GM, bien qu'il s'agisse d'un emprunt du français au grec via le latin *crisis* 'décision' (réemprunt). C'est-à-dire le mot κρίση 'crise' est un

[34] Voir note 26.

[35] Ex. ρόκα 1 'quenouille' < ital. *rocca*, ρόκα 2 'roquette' < ital. *ruca*.

[36] Évidemment, il s'agit d'un continuum sémantique.

[37] Babiniotis (2016: 353), dans l'entrée μέλος, propose un seul sens composé 'partie du corps - mélodie' et dans la partie étymologique note que le mot μέλος en GA combinait, dès le début, les sens 'partie du corps' et 'mélodie, phrase musicale'. Selon nous, en synchronie, il s'agit d'un cas net d'homonymie.

[38] Μέλος 1 'membre du corps', μέλος 2 'mélodie', μέλος 3 'partenaire, associé', μέλος 4 'membre de la famille'. Dagkitsis (1978-1984) distingue trois entrées μέλος.

[39] 1. point de vue, 2. commentaire, 3. verdict, 4. évaluation, 5. recrudescence.

[40] 1. situation difficile, 2. détérioration de la santé, 3. point de vue, 4. faculté de jugement.

mot formellement reliable à κρίνω et étymologiquement relié à lui, mais dont le sens interdit de le rapporter sémantiquement à κρίνω. Des cas de ce type fournissent un argument fort en faveur de la prise en compte du sens et pas seulement de la forme, qui constitue le seul critère d'analyse dans le cas ci-dessus, car le fait de donner la primauté à la forme au détriment du sens, et par là même de présenter ces unités lexicales soit comme des polysèmes à l'aide de critères étymologiques soit comme des mots construits amène à une présentation lacunaire, voire erronée.
Pour nous, les exemples présentés ci-dessus constituent des cas d'homonymie en synchronie.[41] Voici comment on pourrait les présenter :

[1] entrée κρίση 1 'crise, perturbation', ex. κρίσιμος 'critique', κρισιμότητα 'état critique'
entrée κρίση 2 'jugement', renvoi à l'entrée κρίνω 'juger', qui construit un grand nombre de dérivés et de composés, ex. κριτής 'juge, arbitre', κριτήριο 'critère', κριτική 'critique', κριτικός 'critique', κριτικάρω 'critiquer', άκριτος 'irréfléchi', ακρισία 'manque de discernement', ακριτόμυθος 'qui parle inconsidérément', ευθυκρισία 'rectitude de jugement', δικαιοκρισία 'sûreté de jugement', διακρίνω 'distinguer', εγκρίνω 'approuver', επικρίνω 'reprocher', κατακρίνω 'blâmer', προκρίνω 'préférer'.
[2] entrée τέλος 1 'fin', ex. ατελής 'incomplet', ατέλεια 'imperfection', τελικός (nom) 'finale', συντέλεια 'consommation, fin', τελειώνω 'finir', τελειωμός 'fin', τελευταίος 'dernier', τελειόφοιτος 'en dernière année d'études', τελεσίδικος 'irrévocable', τελευτή 'mort', εντέλει 'en fin de compte'
entrée τέλος 2 'taxe', ex. ατελής 'exonéré', ατελώς 'franco de port', ατέλεια 'franchise', ισοτέλεια 'égalité', τελωνείο 'douane', εκτελωνίζω 'dédouaner', υποτελής 'tributaire'.

Pourtant, en GM on repère un grand nombre de cas qui posent problème et qui ne pourront être résolus qu'après des recherches psycholinguistiques. Le problème ardu de la délimitation de la famille se pose dès le début :
1. Sous l'entrée τέλος 1 'fin' y aura-t-il τέλειος 'parfait' et εντελώς 'tout à fait' ? Il n'y a aucun doute que ces unités ont entre elles des relations étymologiques. Pourtant, puisqu'en synchronie τέλος 1 'fin' évoque le sens de 'fin', tandis que τέλειος évoque celui de 'haut degré', nous proposerions leur autonomisation, c-à-d. une entrée τέλος 'fin' et une entrée τέλειος 'parfait' (τελειομανής 'perfectionniste', τέλεια 'parfaitement', ατελής 'imparfait', τελειοποιώ 'perfectionner', εντελώς 'tout à fait' etc.).
2. Y aura-t-il aussi une autre entrée τέλος 'but',[42] ex. τελικές προτάσεις 'en grammaire, propositions de but', τελ(ε)ολογία 'téléologie' ? Nous proposerions l'introduction d'une entrée τελο-, c-à-d. d'une entrée ayant la forme d'une base non autonome (sous-lexicale), qui construit l'adjectif dérivé τελικός 'de but' et le composé τελ(ε)ολογία 'téléologie'. Ceci pour la raison que la présence du mot τέλος 'but' dans les textes du GM, par exemple dans la définition du terme τελ(ε)ολογία 'téléologie',[43] constitue, dans le cadre d'une définition morpho-étymologique, la mention d'un terme d'une autre synchronie.
3. Sous l'entrée τέλος 2 'taxe' y aura-t-il le mot φιλοτελισμός 'philatélie'? Bien qu'il y ait entre eux un lien étymologique (v. *DGS* 1998), le changement de α en ο (φιλ**α**τελισμός > φιλ**ο**τελισμός) affecte la présence du constituant ατέλεια 'franchise' et pour cette raison nous proposerions de n'inclure le mot φιλοτελισμός 'philatélie' dans l'entrée τέλος 2 'taxe' qu'accompagné de commentaire.
4. Dans un grand nombre de préfixés en GM, la contribution sémantique du préfixe est affaiblie à tel point qu'en synchronie il ne serait pas pertinent, du point de vue sémantique, d'inclure ces formes dans l'entrée de leur base, ex. le locuteur ne reconnaitrait pas de relation sémantique entre αποτελώ 'constituer'[44] et τέλος 1 'fin' ou τελώ 'célébrer'.[45] Par conséquent αποτελώ 'constituer' sera le centre de sa famille, qui comprendra les mots αποτέλεσμα 'résultat', αποτελεσματικός 'efficace', αποτελεσματικά 'efficacement', συναποτελώ 'constituer ensemble' etc. ; de même τελώ 'célébrer' constituera le centre de sa famille, qui comprendra les mots τέλεση 'célébration', τελεστής 'opérateur', τελετάρχης 'ordonnateur', ιεροτελεστία 'rite'.
5. Enfin, vu que, dans notre proposition, nous accordons beaucoup d'importance à la cohérence sémantique des unités (sous-)lexicales de la famille, nous proposons d'enregistrer dans des entrées différentes des unités lexicales et sous-lexicales qui se ressemblent quant à la forme mais qui se différencient du point de vue sémantique pour des raisons diachroniques (Anastassiadis-Syméonidis 2005 ; Fliatouras 2020), ex.

Entrée πέτρα 'pierre', ex. πετροπόλεμος 'bataille à coups de pierres', πετρούλα 'petite pierre', πετραδάκι 'caillou', πετράδι 'pierre précieuse', πετριά 'jet de pierre', πέτρινος 'en pierre', πετροβολώ 'lancer des pierres', πετρώνω 'pétrifier'.
Entrée πετρ- 1. 'rocher', ex. πετρογραφία 'pétrographie', πετρέλαιο 'pétrole', πετρογένεση 'pétrogenèse', πετρόψαρο 'labre', πέτρωμα 'roche'.
Entrée πετρ- 2. 'pétrole', ex. πετρέλαιο 'pétrole', πετρελαιοειδή 'produits pétroliers', πετρελαιοπηγή 'puits de pétrole', πετροδολάριο 'pétrodollar', πετροχημικά 'produits pétrochimiques'.

3.2 Polysémie

Il y a polysémie si les sens d'une unité sont dérivables sémantiquement l'un de l'autre (Corbin 1987: 258). Par exemple la

[41] Solution proposée aussi par le *DGS* (1998).
[42] V. huitième sens de l'entrée τέλος dans le Dictionnaire de Babiniotis (2002).
[43] Selon Leroi (2018: 93), le terme a été inventé en 1728 par le philosophe allemand Christian Wolff, qui s'est basé sur le mot τέλος 'but' du GA. Et, pour le français, *téléologie* est attesté en 1765 dans l'*Encyclopédie* (*Le Robert historique* 1992).
[44] Une recherche psycholinguistique pourrait donner une réponse scientifique sûre.
[45] Dagkitsis (1978-1984) reconnaît deux entrées αποτελώ 'constituer', qui renvoient à τελώ 'célébrer', parce qu'il s'agit d'un dictionnaire étymologique.

relation entre κορυφή 'sommet' et κορυφή 'ce qui domine' est sémantiquement explicable par une métaphore conceptuelle ; il s'agit de deux sens différents qui sont réductibles l'un à l'autre et décrits par la rhétorique classique, la stylistique et, plus récemment, par la linguistique cognitive. Il serait nécessaire de prendre en compte la dimension polysémique du morphème lexical : Les membres d'une unité lexicale polysémique appartiennent à la même famille. Pourtant, nous proposons que la famille comprenne des sous-ensembles pour distinguer par ex. le sens propre du sens figuré, métonymique etc., pour limiter l'ambiguïté morphémique ou lexématique et pour rendre plus clair le caractère systématique du lexique.[46] Par exemple pour l'entrée μέλος 1α 'membre d'un ensemble', ex. πολυμελής 'à plusieurs membres', μονομελής 'à un seul membre' - 1β 'membre, partie du corps', ex. αρτιμελής 'bien formé, entier' ; pour l'entrée κορυφή 'sommet' 1α 'partie la plus élevée', ex. κορυφογραμμή 'ligne de crête', κατακόρυφος 'vertical' – 1β par métaphore 'ce qui domine', ex. κορυφαίος 'éminent', κορυφώνω 'culminer', κορύφωση 'summum', αποκορύφωμα 'point culminant', σύνοδος κορυφής 'sommet'.

De tout ce qui vient d'être dit, il en résulte que, pour la construction d'une famille, la relation qui unit tous les membres est de nature lexico-sémantique. Elle n'est pas que sémantique, car, dans ce cas-là, elle aurait dû contenir aussi les synonymes,[47] les antonymes, les hyperonymes, les hyponymes, les méronymes etc.[48] Elle n'est pas, non plus, que formelle, car l'entrée σκάλα 'échelle' aurait dû comprendre aussi σκαλίζω 'sarcler'.[49] De cette manière, des questions concernant la transparence sémantique et le degré de cohérence sémantique d'une unité lexicale se révèlent d'une importance cruciale.[50]

4. Pourquoi la Notion de Famille est-elle Importante?

De ce qui vient d'être dit, il en résulte que la famille a trois propriétés : hétérogénéité, transparence sémantique et taille. Plus particulièrement, bien que les unités lexicales qui constituent la famille soient hétérogènes en ce qui concerne la liberté des morphèmes et leur statut constructionnel, l'entrée est reliée avec les membres de sa famille par la transparence sémantique. De même, la taille de la famille, une variable paradigmatique, influe quantitativement et qualitativement sur le lexique mental des locuteurs.[51] En incluant dans la même famille pas seulement les mots qui entretiennent une relation sémantique et constructionnelle avec le mot-entrée, mais aussi les mots qui n'entretiennent avec l'entrée qu'une relation sémantique en synchronie, nous délimitons mieux la notion de famille, car sa vraie taille dépend à la fois de processus de productivité lexicale synchroniquement actifs et de relations étymologiques inertes dont le locuteur n'a plus conscience, mais qui peuvent néanmoins jouer un rôle dans ses stratégies de traitement et d'organisation lexicale. La confection d'un dictionnaire de ce type va répondre à ce triple objectif : fournir des matériaux fiables pour la recherche en psycholinguistique, améliorer la description lexicographique et offrir à la didactique du GM un véritable instrument de travail.

Plus spécialement, nous considérons que la notion de *famille d'unités (sous-)lexicales* telle qu'elle a été décrite est importante, car elle est utile :

1) En linguistique théorique pour la classification lexico-sémantique et morphologique des unités lexicales et leurs relations mutuelles ; c'est la raison pour laquelle la lexicologie, par exemple en France avec Picoche (1977), s'est occupée depuis bien des années des différentes sortes de champs.

2) En psycholinguistique, puisqu'à partir de 1997 l'article de Schreuder & Baayen (1997: 129) marque le début de l'examen de l'influence de la Taille de la Famille Morphologique (TFM)[52] durant l'accès lexical, c'est-à-dire durant les premiers stades de l'identification du mot. Il est trouvé que ce facteur (TFM) facilite le traitement[53] et la récupération des membres de la famille et influence les temps de décision lexicale (temps de réaction) de la part du locuteur (Dijkstra et al. 2005), ce qui a des répercussions directes sur l'architecture du lexique mental.[54]

Plus particulièrement, la taille de la famille morphologique, et spécialement la fréquence de type influe de façon positive sur les temps de réaction aux décisions lexicales qu'ont effectuées les locuteurs[55] ainsi que sur les résultats de fréquence subjective des dérivés (Ford *et al*. 2010: 126). De même, elle émerge comme un facteur qui a un effet significatif sur le traitement lexical après l'identification de la forme et sert d'indicateur du degré d'intégration d'un nom dans le réseau des relations sémantiques connectant les concepts dans le lexique mental (Ford *et al*. 2010: 126-127 ; Schreuder & Baayen 1997: 131, 135). Il est toutefois précisé que les effets positifs ne sont valables que dans les cas où la relation sémantique entre la base et ses dérivés est transparente ou semi-transparente (Feldman & Pastizzo 2003).

3) En terminologie, où des distinctions similaires s'appliquent (Kokourek 1982: 162) : par exemple dans l'entrée οξύ

[46] Une conséquence qui irait contre le principe d'économie serait la répétition d'une forme lexicale tant dans le groupe du sens propre que dans le groupe du sens métaphorique. Pourtant, cela est sans importance pour un dictionnaire en ligne comme le nôtre.

[47] Par exemple les unités lexicales όρος 'montagne' et βουνό 'montagne' feraient partie de la même famille.

[48] Comme dans un dictionnaire de synonymes, ex. Iordanidou (2005) ou dans le dictionnaire analogique de Vostantzoglou (1962).

[49] L'exemple est tiré de Ntagkas (2019).

[50] Toutefois, les résultats pourraient varier en fonction du degré de littératie linguistique des locuteurs.

[51] De Jong *et al*. (2000: 359) arrivent à la conclusion que l'effet de la taille de la famille est un effet sémantique, accompagné d'un vrai composant morpho-syntaxique.

[52] Schreuder & Baayen (1997: 118) "the size of the morphological family, i.e., the number of different words in the family, emerged as a substantial factor"; Schreuder & Baayen (1997: 121) "We will refer to the number of different words in the morphological family (excluding from the count the base word itself) as the *morphological family size*".

[53] La grande Taille de la Famille Morphologique facilite la reconnaissance de ses membres, sans pour autant entrer en interaction avec la fréquence de la base ou la productivité du préfixe ou du suffixe (Feldman & Pastizzo 2003 ; Ford, Davis & Marslen-Wilson 2010 ; Diependaele, Grainger & Sandra 2012: 319 ; Xu & Taft 2015).

[54] La fréquence de la famille des unités lexicales d'une entrée présuppose qu'un grand nombre de mots construits ont leur représentation propre dans le lexique mental (Schreuder & Baayen 1997: 136). De même, en ce qui concerne la famille des unités lexicales, elle est située à un niveau supra-lexical (Giraudo & Grainger 2001 ; Voga 2015).

[55] De Jong *et al*. (2000: 343) ont trouvé une corrélation entre la taille de la famille et le temps de réaction.

‘acide’, suffixé όξινος ‘acide’, composé οξυγόνο ‘oxygène’, composé polylexical μαγγανικό οξύ ‘acide manganique’.
4) En lexicographie, où nous constatons l’absence d’un dictionnaire synchronique de familles d’unités lexicales du GM, fondé sur la transparence sémantique, qui illustrera la systématicité du lexique. Ce dictionnaire en ligne pourrait présenter, sous forme de liens, des informations sur la fréquence tant des formes lemmatiques que des unités lexicales de la même famille ou bien des contextes extraits de corpus. Un tel dictionnaire serait utile aux psycholinguistes, aux morphologues, aux métalexicographes et aux enseignants de langue. Les dictionnaires de Dagkitsis (1978-1984) et de Dorbarakis (1993/1999) malgré leur orientation étymologique et celle, analogique, de celui de Vostantzoglou (1962), le livre Χτίζω λέξεις[56] ‘Construire des mots’ (Iordanidou et Pantazara 2010), le *Dictionnaire des dérivés et des composés du grec moderne* (Babiniotis 2016) et les dictionnaires de langue seraient particulièrement utiles à la confection d’un dictionnaire des familles d’unités (sous-)lexicales du GM, tel que nous l’avons décrit.
5) En Didactique du GM comme langue maternelle, seconde, étrangère ou langue d’héritage.[57] Ce n’est pas un hasard si Picoche (1993) commence son livre sur la didactique du vocabulaire de la langue française par les familles des mots. Selon Bauer & Nation (1993), la famille de mots est importante pour une approche systématique de l’enseignement du vocabulaire et pour mesurer la charge lexicale des textes. Et la recherche expérimentale de Morin (2006) est arrivée à la conclusion que l’enseignement explicite de la morphologie constructionnelle comme stratégie pour construire des familles de mots a aidé les apprenants non seulement à approfondir leurs connaissances sur les mots connus mais aussi à appliquer leurs connaissances constructionnelles à des formes nouvelles.
Par conséquent, il nous semble important de noter que le *Dictionnaire de familles d’unités (sous-)lexicales du GM* serait utile à l’enseignement de la stratégie de la segmentation morphologique (Anastassiadis-Syméonidis & Mitsiaki 2010a ; Anastassiadis-Syméonidis 2019), de la stratégie de l’usage du dictionnaire en classe (Anastassiadis-Syméonidis 1997 ; Efthymiou 2013: 142 ; Anastassiadis-Syméonidis & Mitsiaki 2010b) mais aussi de la distinction des paronymes, qui seront affectés à des entrées différentes,[58] ex. κρίσιμος ‘critique’ – κριτικός ‘critique’ (v. *supra* entrée κρίση 1 ‘crise’ – entrée κρίση 2 ‘jugement’), διδάκτορας ‘docteur’ (διδακτορικός ‘doctoral’, διδακτορικό ‘doctorat’, διδακτορία ‘titre de docteur’) – δικτάτορας ‘dictateur’ (δικτατορία ‘dictature’, δικτατορικός ‘dictatorial’).
Si l’enseignant procède à l’initiation de ses élèves au sujet de la famille d’unités (sous-)lexicales et des relations lexico-sémantiques qu’elles entretiennent dans le lexique mental, en moins de temps d’enseignement il apportera plus et de meilleurs résultats en matière de compréhension et de production de discours écrits et oraux. En plus, dans le cas de la didactique du GM comme langue seconde, étrangère ou langue d’héritage, l’enseignant pourrait faire appel aux nombreux cognats et leurs familles, dus à l’emprunt intensif entre le grec et le français ou l’anglais dans les deux sens, qui fait que les mots des deux ou trois langues soient connectés, puisque stockés dans un lexique mental unifié, ex. entrée κρίση 1 ‘crise, perturbation’/fr. *crise*/ang. *crisis* : Famille : κρίσιμος ‘critique’/fr. *critique*, ang. *critical*, υπερκρίσιμος ‘très critique’, κρισιμότητα ‘état critique’.[59] Dans cet exemple, la taille de la famille morphologique grecque étant plus importante (4 unités au lieu de 2 en français et 2 en anglais) elle pourrait conférer à ses membres une plus grande activation dans le lexique, selon les résultats des études expérimentales de Voga, Gardani et Giraudo (2020: 530) et Voga (2020: 41).

5. Conclusion

Nous espérons que l’examen, sous une lumière nouvelle, de la notion de famille de mots peut permettre d’apporter des éléments de réponse aux nombreuses questions concernant avant tout son contenu.
Ce projet est original sur plusieurs points :
La notion de famille est définie de façon plus restrictive en ce qui concerne les relations entre synchronie et diachronie, et, en même temps, plus extensive, puisque le seul critère est la cohérence sémantique entre le mot-entrée et les unités qui lui appartiennent. Bien que les relations constructionnelles entre les membres d’une famille forment la majorité des cas, la famille doit comprendre aussi des unités entretenant avec le mot-entrée des relations pseudo-étymologiques et pseudo-constructionnelles, pourvu qu’elles soient sémantiquement cohérentes avec lui. Ainsi la taille d’une famille peut-elle dépendre à la fois d’opérations constructionnelles actives en synchronie et de relations étymologiques inertes (étymologie populaire), qui peuvent, pourtant, influencer le locuteur dans ses stratégies de traitement et d’organisation lexicale.
Les articles du *Dictionnaire de Famille d’unités (sous-)lexicales* vont présenter une cohérence sémantique. Pour la préserver, nous devons adopter l’homonymisation, en dégroupant les formes identiques qui se différencient au niveau sémantique, ce qui, souvent, va de pair avec la répartition des formes construites, malgré la position de la linguistique cognitive qu’il n’existe pas de sens bien délimités.
En mettant en valeur la cohérence sémantique de la famille d’unités (sous)-lexicales, cet ouvrage sera utile à la linguistique théorique et, avant tout, il va faciliter les recherches des psycholinguistes sur l’architecture du lexique mental ainsi que la tâche de ceux qui travaillent sur la didactique du GM, parce que nous avons tenu compte des acquis récents de la psycholinguistique.

Bibliographie

[56] Surtout pour les entrées d’unités sous-lexicales.
[57] Johnston (1999) essaie de répondre aux questions pourquoi, comment et quand étudier les familles de mots, mais en ne prenant en compte que la phonologie et les rimes. Pour elle *cat, sat, rat, hat, that, mat* forment en anglais une famille. Selon nous, il s’agit plutôt d’un champ lexical.
[58] Avec des renvois, pour faciliter l’apprentissage.
[59] Sur la représentation des cognats dans un modèle d’architecture commune du lexique mental bilingue v. Giraudo & Voga 2013: 104 ; Voga 2020: 55.

Anastassiadis-Syméonidis, A. (1986). La nature et la productivité du formant -ποιώ '-ifier'. In *Studies in Greek Linguistics*, 7, Thessaloniki: Kyriakidis, pp. 49-70. [en grec].

Anastassiadis-Syméonidis, A. (1994). *Emprunt néologique en grec moderne – Analyse morphophonologique des emprunts directs du grec moderne au français et à l'anglo-américain.* Thessaloniki. [en grec].

Anastassiadis-Syméonidis, A. (1997). Éducation et lexicographie. In *Actes du 2e Colloque Panhellénique sur l'enseignement du grec*, Thessaloniki: Kodikas, pp. 149-176. [en grec].

Anastassiadis-Syméonidis, A. (2002). *Dictionnaire inverse du grec moderne*. Thessaloniki: Institut d'Études Néohelléniques (Fondation Manolis Triantafyllidis). [en grec]. Accessible à: https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/reverse/index.html [5/6/2021].

Anastassiadis-Syméonidis, A. (2004). Flexion et dérivation : mythe et vérité. In *Studies in Greek Linguistics*, 24, Thessaloniki, pp. 43-54. [en grec].

Anastassiadis-Syméonidis, A. (2005). Les éléments πετρο- 'petro-' et λιθο- 'litho-' dans la terminologie grecque. In Actes du 5e colloque de terminologie grecque ELETO – *Langue grecque et terminologie*, Athènes, pp. 13-22. [en grec] Accessible à: http://www.eleto.gr/gr/papers.htm#5thPapers [5/6/2021].

Anastassiadis-Syméonidis, A. (2019). La forme étendue de l'approche ÉMILE. In *Philologos,* 174-175, pp. 27-45. [en grec].

Anastassiadis-Syméonidis, A. (2020). La notion de famille d'unités (sous-)lexicales. In *Studies in Greek Linguistics*, 40, Thessaloniki: Institut d'Études Néohelléniques (Fondation Manolis Triantafyllidis), pp. 29-39. [en grec] Accessible à: http://www.ins.web.auth.gr/index.php?option=com_content&view=article&id=1281&Itemid=422&lang=el.

Anastassiadis-Syméonidis, A., Mitsiaki, M. (2010)a. La segmentation morphologique comme stratégie d'enseignement du vocabulaire du grec moderne comme langue seconde et étrangère. *In* A. Psaltou-Joycey, M. Mattheoudakis (éds.), *Actes du 14e colloque international de l'Association Grecque de Linguistique Appliquée,* Thessaloniki, pp. 65-77. [en grec] Accessible à: https://www.enl.auth.gr/gala/14th/Papers/Greek%20papers/Anastasiadi-Symeonidi&Mitsiaki.pdf [5/6/2021].

Anastassiadis-Syméonidis, A., Mitsiaki, M. (2010)b. L'usage des dictionnaires monolingues dans l'enseignement du grec comme langue étrangère : une application didactique. In K. Dinas, A. Hatzipanagiotidi, A. Vakali, T. Kostopoulos, A. Stamou (éds.), Actes du colloque panhellénique avec participation internationale *L'enseignement du grec comme langue première/maternelle, seconde/étrangère*. [en grec] Accessible à: http://linguistics.nured.uowm.gr/Nimfeo2009/praktika/files/down/paraskeui/aithusa2/anastasiadiMitsaki.pdf [5/6/2021].

Anastassiadis-Syméonidis, A., Voga, M. (2010). Le caractère symbolique de quelques lettres du grec moderne. *In Studies in Greek Linguistics* 30, Thessaloniki: Institut d'Études Néohelléniques (Fondation Manolis Triantafyllidis), pp. 79-97. [en grec] Accessible à: http://ins.web.auth.gr/index.php?option=com_content&view=article&id=522&Itemid=179&lang=el [5/6/2021]

Anastassiadis-Syméonidis, A., Voga, M. (2011). Perception en ligne de phrases figées en grec. In C. G. Royo, P. Mogorrón Huerta (éds.) *Estudios y análisis de fraseología contrastiva: Lexicografía, traducción y análysis de corpus,* Alicante: Publicationes de la Universidad de Alicante, pp.15-32.

Babiniotis, G. (2002)2. *Dictionnaire de la langue grecque moderne*. Athènes: Centre de Lexicologie. [en grec].

Babiniotis, G. (2016). *Dictionnaire des dérivés et composés du grec moderne – «Les enfants et petits-enfants» des mots de notre langue*. Athènes: Centre de Lexicologie. [en grec].

Bauer, L., Nation, P. (1993). Word Families. In *International Journal of Lexicography,* 6 (4), pp. 253–279.

BDME TIP Plataforma web para el estudio morfogenético del léxico 1980/2009-2016. Accessible à: https://bdme.iatext.es [5/6/2021].

Bertram, R., Baayen, R.H. & Schreuder, R. (2000). Effects of family size for complex words. In *Journal of Memory and Language*, 42, pp. 390–405.

Bybee, J. (1985). *Morphology: A Study of the Relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. (1988). Morphology as lexical organization. In M. Hammond et M. Nooman (eds.), *Theoretical Morphology. Approaches to modern linguistics*, pp. 119-142. San Diego: Academic Press.

Bybee, J. (1995). Regular morphology and the lexicon. In *Language and Cognitive Processes,* 10(5), pp. 425-455.

Corbin, D. (1987/1991). *Morphologie dérivationnelle et structuration du lexique* (2 τόμοι). Tübingen/ Villeneuve d'Ascq: Max Niemeyer Verlag/ Presses Universitaires de Lille.

Corbin, D. (tapuscrit 1999). *Le lexique construit*.

Dagkitsis, K. (1978-1984). *Dictionnaire étymologique du grec moderne*. 2 vol., Athènes: I. Vassileiou. [en grec].

De Jong N., Schreuder, R. & Baayen R. H. (2000). The morphological family size effect and morphology. In *Language and Cognitive Processes*, 15 (4/5), pp. 329–365.

DGS / Dictionnaire du grec standard. (1998). Thessaloniki: Institut d'Études Néohelléniques. Accessible à: http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/index.html [5/6/2021] [en grec].

Diependaele, K., Grainger, J. & Sandra, D. (2012). Derivational morphology and skilled reading: An empirical overview. In M. J. Spivey, K. McRae & M. F. Joanisse (eds.), *The Cambridge handbook of psycholinguistics*. Cambridge: Cambridge University Press, pp. 311-332.

Dijkstra, T., Moscoso del Prado Martin, F., Schulpen, B., Schreuder, R. & Baayen, R. H. (2005). A roommate in cream: Morphological family size effects on interlingual homograph recognition. In *Language and Cognitive Processes,* 20, pp. 7-41.

Dorbarakis, P. (1993/1999). *Dictionnaire étymologique et sémasiologique du grec moderne – Entrées présentées sous forme de familles étymologiques*. Athènes: Spoudi. [en grec].

Dubois, J., Lagane, R., Niobey, G., Casalis, D., Casalis, J. & Meschonnic, H. (1967). *Dictionnaire du Français Contemporain*. Paris: Larousse.

DULG / Dictionnaire d'usage de la langue grecque. (2014). (Coordinateur et éditeur: Chr. Charalampakis). Athènes: Académie d'Athènes & Imprimerie Nationale. [en grec].

Efthymiou, A. (2013). *L'enseignement du vocabulaire à l'école primaire – Théorie et applications*. Thessaloniki: Epikentro. [en grec].

Feldman, L. B., Pastizzo, M. J. (2003). Morphological facilitation: The role of semantic transparency and family size. In R. H. Baayen & R. Schreuder (eds.) *Morphological structure in language processing.* Berlin/ New York: Mouton de Gruyter, pp. 233-258.

Fliatouras, A. (2020). Vers le besoin de quantification de la recherche étymologique: la distribution statistique étymologique du vocabulaire du grec standard. In *Studies in Greek Linguistics*, 40, pp. 525-535. [en grec]. Accessible à: http://www.ins.web.auth.gr/index.php?option=com_content&view=article&id=1281&Itemid=422&lang=el

Ford, M. A., Davis, M. H. & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. In *Journal of Memory and Language,* 63, pp. 117-130.

Georgakopoulos Th., Polis St. (2018). The semantic map model: State of the art and future avenues for linguistic research. In *Language and Linguistics Compass*. 12:e12270, 33p. Accessible à: https://doi.org/10.1111/lnc3.12270 [10/6/2021].

Giraudo, H., Grainger, J. (2001). Priming complex words: Evidence for supralexical representation of morphology. In *Psychonomic Bulletin and Review,* 8(1), pp. 127-131.

Giraudo, H., Voga, M. (2013). Prefix units in mental lexicon. In N. Hathout, F. Motermini & J. Tseng (eds.) *Morphology in Toulouse – Selected Proceedings of Décembrettes* 7, LINCOM *Studies in Theoretical Linguistics* 51, pp. 91-107.

Goutsos, D. (2006). Développement du lexique: du niveau de base au niveau avancé. In D. Goutsos, M. Sifianou & A. Georgakopoulou, *Le grec comme langue étrangère: Des mots aux textes*. Athènes: Patakis, pp. 13-96. [en grec].

Hay, J. B., Baayen, R. H. (2005). Shifting paradigms: gradient structure in morphology. In *Trends in Cognitive Sciences,* 9(7), pp. 342-348.

Iordanidou, A. (éd.) (2005). Neurosoft – *Thesaurus de synonymes et de contraires du grec moderne.* Athènes: Éditions Patakis. [en grec].

Iordanidou, A., Pantazara, M. (éds) (2010). *Construire des mots*. Athènes: Kondyli. [en grec].

Johnston, F.R. (1999). The timing and teaching of word families. *The Reading Teacher*, 53, pp. 64-75.

Kocourek, R. (1982). *La langue française de la technique et de la science*. Wiesbaden: Brandstetter Verlag.

Le Robert-Dictionnaire historique de la langue française (A. Rey, dir.) (1992). Paris: Dictionnaires le Robert.

Leroi, A. M. (2014). *The Lagoon: How Aristotle invented science*. Traduction en grec: Aim.-Al. Kritikou, Em. Kritikou (trad.), 2018. Thessaloniki: Ropi.

Martinet, A. (1979). *Grammaire fonctionnelle du français*. Paris: Didier – Crédif.

Morin, R. (2006). Building Depth of Spanish L2 Vocabulary by Building and Using Word Families. In *Hispania* 89/1, pp. 170-182. Accessible à: https://www.jstor.org/stable/20063269 [5/6/2021].

Nation, P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards, B. Laufer (éds.) *Vocabulary in a Second Language - Selection, acquisition, and testing*. Amsterdam/Philadelphia: John Benjamins, pp. 3-13.

Ntagkas, N. (2019). Protocole expérimental d'amorçage morphologique. In *Studies in Greek Linguistics* 39, pp. 1125-1142. [en grec] Accessible à: http://ins.web.auth.gr/index.php?option=com_content&view=article&id=1214&Itemid=420&lang=el [5/6/2021]. Στο κυρίως κείμενο αναφέρεται Dagkas

Oikonomou, M. (1971). *Grammaire du grec ancien*. Ministère de l'Éducation nationale et des cultes – Institut Pédagogique. Athènes: OEDV. [en grec].

Picoche, J. (1977). *Précis de lexicologie française*. Paris: Nathan.

Picoche, J. (1993). *Didactique du vocabulaire français*. Paris: Nathan.

Schreuder, R. & Baayen, R. H. (1997). How complex simplex words can be. In *Journal of Memory and Language* 37, pp. 118-139.

Swinney, D. A., Cutler, A. (1979). The access and processing of idiomatic expressions. In *Journal of Verbal Learning and Verbal Behavior* 18, pp. 523-534.

Triantafyllidis, M. et al. (1941/1978). *Grammaire du grec moderne (Démotique).* Thessaloniki: Institut d'Études Néohelléniques (Fondation Manolis Triantafyllidis). [en grec].

Voga, M. (2015). Vers une représentation supra-lexicale de la morphologie dans le lexique mental bilingue: Données de cognats grec-français. In *Studies in Greek Linguistics*, 35, pp. 106-130. [en grec]. Accessible à: http://ins.web.auth.gr/images/MEG_PLIRI/MEG_35_106_130.pdf [5/6/2021].

Voga, M. (2020). *Représentation morphologique et transferts inter-langues dans le lexique mental. De la perception au sens de la construction langagière - Synthèse des travaux de recherche (2004-2020)*, Université Paul-Valéry Montpellier III.

Voga, M., Gardani, F. & Giraudo, H. (2020). Multilingualism and the Mental Lexicon. Insights from language processing, diachrony, and language contact. In V. Pirelli, I. Plag & W. Dressler (eds.), *Word knowledge and word usage: A Cross-Disciplinary Guide to the Mental Lexicon*. Series: Trends in Linguistics. Studies and Monographs [TiLSM], 337. Berlin: Mouton De Gruyter, pp. 506-552. Accessible à: https://doi.org/10.1515/9783110440577 [5/6/2021].

Vostantzoglou, Th. (1962)^2. *Antilexikon* ou *Onomastikon du grec moderne*. Athènes: Patris. [en grec].

Xu, J., Taft, M. (2015). The effects of semantic transparency and base frequency on the recognition of English complex words. In *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 41(3), pp. 904-910.

Xydopoulos, G. (2008). *Lexicologie – Introduction à l'analyse du mot et du dictionnaire*. Athènes: Éditions Patakis. [en grec].

Combating linguistic myths and stereotypes: The contribution of the *Practical Dictionary of Modern Greek* of the Academy of Athens

Charalambakis Ch.

National and Kapodistrian University of Athens, Greece
ccharala@phil.uoa.gr

Abstract

The aim of this study is to highlight the innovations introduced by the *Practical Dictionary of Modern Greek* that significantly differentiate it from comparable modern printed dictionaries. The main focus is on language myths and stereotypes that are reproduced in various dictionaries. The view that Modern Greek is declining, as shown by the poor vocabulary of young people and the invasion of foreign words, is refuted by the simple browsing of the Practical Dictionary. Modern Greek adapts with great flexibility to modern challenges by enriching its vocabulary with a variety of alternatives. Translation loans perform well, as they mask foreign influences. Anglicisms are found in most European languages, which relativizes the criticism that native speakers of Modern Greek do not mind the present status and the future of their language. The myths of the single correct spelling and etymology of each word are refuted with indisputable evidence. Ethnocentrism and sexism have been eliminated from the *Practical Dictionary of Modern Greek*. The necessity of electronic dictionaries, which constitute the future of lexicography, and the establishment of a 'Language Observatory', the findings of which will be used in real time mainly for the needs of teaching Greek as a second/ foreign language, are stressed.

Keywords: Modern Greek, lexicography, linguistic myths, stereotypes, anglicisms, etymology, ethnocentrism, sexism

1. Introduction

The *Practical Dictionary of Modern Greek* of the Academy of Athens (hereinafter PDAA), of which I had the general responsibility of planning, drafting, and editing, was treated with disbelief from the moment its drafting and publication was announced. The superficial reason was that at that time (2003), two recent dictionaries of Modern Greek (DB, DTR) were in circulation, and there was no need to compile a third one. However, the Academy of Athens took the opposite view. Its 42 regular members unanimously ruled that the existing Modern Greek dictionaries did not satisfactorily meet the modern communication needs of native speakers. On the other hand, the Constitutional Decision 'On the Organization of the Academy of Athens' of March 18, 1926, explicitly states that the obligation of the Academy is "to study and regulate the issues related to our national language, to prepare and compile and publish its Grammar, Syntax and Dictionaries."

Almost all entries were drafted in their first form, with certain specifications and instructions, by two different compilers. Based on these data, I reconstructed and revised all lemmata, comparing them with the well-known Modern Greek dictionaries and selectively with the most valid dictionaries of English, French, German, Italian, and Spanish. It is obvious that the responsibility for the whole project, and for the errors that are left, lies solely with me.

Lexicographic works, such as the PDAA, cannot fully meet the objectives they set from the beginning. Almost all professional lexicographers feel like they are walking on quicksand and are in danger of sinking at any time. The writing of a dictionary is never actually completed, because the language is constantly evolving and enriched with new elements. The Academy of Athens does not regulate the language, but systematically and responsibly monitors and records the linguistic reality in an extensive database from which other individual dictionaries can emerge. It proposes solutions based primarily on usage at a purely synchronous level.

The main objective of PDAA is, on the one part, to highlight the richness and expressive completeness of today's language, without ignoring the diachronic aspect, and on the other, to provide all the necessary lexicographic information for its acquisition by native speakers and its learning by foreigners so that they are able to fully understand and produce with the necessary quality written and oral texts, which cover a variety of communication needs.

The optimistic message that follows from the systematic lexicographic surveys of the Academy of Athens is that Modern Greek displays astonishing lexical richness and impressive flexibility and creativity, as shown by the plethora of neologisms and new meanings recorded in this Dictionary, which will be published in electronic form as well.

The Preface of the Secretary General of the Academy of Athens (p. 7) reflects the Academy's positions on Modern Greek. Vasilios Ch. Petrakos, among others, states the following:

"Young people, scholars needed a dictionary that was easy to use, of everyday speech, modern and valid, the preparation of which the Academy had been examining and discussing for years. Fortunately, the finding in 2003 by the then President of the Academy Grigorios Skalkeas of a significant amount of money made possible the realization of the old goal of the Academy, the compilation of the Practical Dictionary of Modern Greek Language. There were already remarkable dictionaries of Modern Greek, but the Academy aimed with its dictionary at achieving a dual result: the publication of a modern Dictionary and the continuous enrichment of its electronic database. At any time, the scholar will be able to use the lexical richness and the expressive vitality and completeness of our language to all their extent known in science.

The Academy with the Practical Dictionary does not aim to regulate the language, which is proven to be unrealistic, at least since the years of Korais. It presents in a systematic way and scientifically verified the real form of Modern Greek, its linguistic richness and its wonderful expressive possibilities; at the same time, the Dictionary provides the necessary lexicographical information for its understanding by Greeks and its learning by foreigners. With the clarity and completeness of the data provided in the Dictionary, the acquisition of the language becomes possible. By using it, Greeks who express their thoughts and ideas in all kinds of texts, orally and in writing, will be able to benefit from the high point that Modern Greek has reached, with efforts and spiritual struggles of centuries and with prose and the poetry that the craftsmen of the word cultivated and created.

Moreover, with the Academy's many years of systematic linguistic and lexicographic research, the amazing richness of Modern Greek and its truly impressive flexibility and plasticity has become apparent, and is at our disposal. The great multitude of neologisms and new meanings contained in the Practical Dictionary is the undeniable testimony of this fact, which shows the creative power of Greeks in the language and disproves those who mourn its decay".

1.1 The Innovations of the Practical Dictionary

The theoretical background on which the PDAA was based corresponds with modern meta-lexicographic research, the findings of which have been used on the basis of the practical benefits that will be gained by the respective dictionary users. The necessary balance between theory and practice was thus maintained. On the other hand, the weaknesses of the circulating Modern Greek dictionaries were identified, and an effort was made to eliminate such weaknesses in the dictionary to be published. Five years before the start of my retirement, two important dictionaries were published simultaneously, of which comparisons and evaluations were expected to be made. The following planned innovations of the PDAA were presented at a linguistic conference to obtain the necessary feedback:[1]

1. Compilation of the entries from extensive printed and electronic databases, starting from scratch.
2. Establishing a double spelling for certain words. These are equivalent *(εταιρεία-εταιρία)* or equally acceptable (*χρεοκοπία* & *χρεωκοπία*) spellings.
3. Deviation from the traditional practice of many centuries of constant termination of verbs in the first-person definitive present and of nouns in the nominal singular: *ρέει* and not *ρέω*; *μακαρόνια* and not *μακαρόνι.*
4. Automatic syllabification of all main entries.
5. Short and substantive definitions.
6. Indication of the scientific names of animals and plants. Most chemical types of compounds and formulas are also recorded.
7. Unification of meanings (*ρεπορτάζ* shows one meaning and not four).
8. Recording of genuine, authentic examples.
9. Detailed presentation of collocations.
10. Precise stylistic definitions of words and expressions.
11. Presentation of stereotypical expressions, collocations, and fixed vocabulary on a new basis at the end of each entry.
12. Quoting the date of the first appearance of many English and French words and stereotypical expressions as a contribution to the history of culture and science.
13. Recording of many words and meanings known mainly from Cyprus, such as *αθλητοπρέπεια* 'behavior and morals appropriate to an athlete', *αιτητής* 'applicant', and *μάππα* 'football'.
14. Presentation of lemmata in three columns for more comfortable reading. Entries, collocations, and phrases are recorded in clear blue letters.
15. Most of the large-scale entries were written and compiled by two different authors, without their own knowledge, to take their final form from the scientific coordinator who proceeded to successive, often radical redrafting at different intervals.

[1] See Charalambakis (2007, cf. 2010). The findings of this preliminary investigation were discussed at length during the conference. In the end, I revised some of my original views, in particular those relating to stylistic-pragmatic indicators, which were further specified and fully applied in the PDAA. From 2020, this study can be found here: https://bit.ly/3f514OQ.

2. Myths about the Role of Dictionaries

Boulanger (1986, 95-101) mentions the following five lexicographic myths (mythes lexicographiques) that apply to all languages, as people, more or less, think in the same way and reproduce the same linguistic stereotypes:

- The myth of permanence (La mythe de la permanence). When one buys a dictionary, one does not feel the need to replace it with a newer one.
- The myth of uniqueness (La mythe de la unicité). The official or established dictionaries act as a regulatory authority.
- The myth of duration (La mythe de la durabilité). Vocabulary remains stable despite social changes. The dictionary records the official language without deviations.
- The myth of objectivity (La mythe de la objectivité). All dictionaries describe language in the same way. In fact, lexicographers reveal their linguistic beliefs in the entries they compose.
- The myth of the norm (La mythe de la norme). Dictionaries standardize the truth. Although their authors claim to describe the language, in many cases they in fact regulate it.

These lexicon myths are essentially linguistic myths that influence users' attitudes toward the general dictionaries they use. Myths about language exist in all peoples and cultures (Bauer, 1998. For the Greek language, see Sarantakos, 2019). Deep down, there is only one lexicon myth: The average dictionary user takes for granted and indisputable the information contained in each article. When, very seldom, one compares the same entry in two or three similar dictionaries and notices the differences they present, one realizes that there are no identical 'photographs' of the entries, and it would be unnatural for there to be, just as photos taken at a given time, and much more those taken many years ago, are never uniform.

3. Stereotypes

The term stereotype is used in various disciplines, such as sociology, linguistics, philosophy of language, etc., but also in everyday language, to indicate the established perceptions that have prevailed in groups of individuals or in the wider community, without there being objectively verifiable evidence to substantiate these attitudes.

Stereotypes are generally negative in nature and could be equated with social prejudices, most of which refer to racial relations (dislike, hatred, or hostility toward other people or racial groups) and sexism (discrimination against women as a rule; and less of men). The sense of social solidarity created by stereotypes, some of which are not entirely arbitrary, contributes to their perpetuation. Even if one presents evidence that dispels these prejudices, there are people who never get rid of preconceived ideas.

3.1 Linguistic Decline

Some consider the decline of Modern Greek to be a given fact, without realizing that this stereotype remains arbitrary, because the verbal diarrhea for linguistic decline is as old as the language itself and is not just a Modern Greek phenomenon. The arguments put forward in support of this view are generally unhistorical and misleading. In essence, it is a covert "linguistic ideology" that reproduces the stereotype of alarmism, thus legitimizing its self-proclaimed patrons. Many do not want to believe that language follows its own independent course, which no native speaker can accurately predict or prescribe. Protecting language from decay and decline presupposes that the language is an "accomplished cultural entity" that we must pass on unscathed to future generations.

3.1.1 'Lexical Poverty'

The allegation that young people have a poor vocabulary is a consistent obsession. Pupils are an easy target, remaining defenseless. Adults in proportion use limited vocabulary, and most importantly, they experience "semantic poverty." The quality of speech depends not solely on the number of words that belong to an individual's active and receptive vocabulary but primarily on knowledge of a multiplicity of words and their combinatorial possibilities. Modern Greek has an impressive vocabulary of about 500,000 words, of which few members of the linguistic community have the privilege of knowing one in ten. Awareness of this should make us all more careful. We think we know our language, but we know only a certain percentage of its vocabulary, with many gaps and weaknesses. Language is not an abstract idealization but the specific linguistic stock of each individual, only part of which remains constant.

3.1.2 Foreign Words

The most important source of language decay is the entrance of foreign words, which "infect" or contaminate the body of Modern Greek. Anyone who systematically studies the mechanisms of operation and contact of languages will understand that the phenomenon of borrowing is much more complex than one could imagine. The Greek language, during its long evolutionary course, borrowed thousands of words from all people and nations with which it came into contact. Since the end of the Second World War, the main source of the influx of foreignisms in Modern Greek, as well as in most languages of the world, has been English, especially American English. (Charalambakis, 2010, 279-284).

3.1.2.1 The Influence of Anglo-American on European Languages

The influx of anglicisms concerns most of the world's languages and should not be seen in the narrow context of common Modern Greek. The average cultured citizen, as well as some scholars, attribute the frequent use of English words to 'linguistic carelessness', 'negligence', 'indifference', 'complacency', etc. of native speakers. They also blame the State and call for various measures to limit the evil. However, things are not that simple. The strong influence of Anglo-American in other languages is an international phenomenon with the same causes everywhere: the dominance of the United States of America in economy, high technology and military equipment, politics, science, and culture in general. Another important reason is the positive attitude of many native speakers, especially young people, toward the American way of life.

Linguists, on the other hand, began to emphasize the positive aspects of foreign words, their contribution to facilitate international communication, and their functionality compared with the uncomfortable or unrealistic translation equivalences in the mother tongue. Failed attempts to limit anglicisms in the recent past, combined with the need to describe rather than regulate language, led researchers to believe that one should let language follow its own path, thus showing in practice not just tolerance but an understanding of linguistic pluralism and multiculturalism. The German philosopher, sociologist, and musicologist Theodor Adorno (1959) undertook the defense of foreign words with a remarkable article.

3.1.2.1.1 Anglicisms in Modern Greek

With the globalization of the economy and knowledge and the development of modern electronic services and telematics (interactive electronic images, electronic access to information, e-mail), as shown by the many new compounds with the morpheme *τηλ(ε)- tel(e)-* (*-αγορές, -εργασία, -ιατρική, -κατάρτιση -markets, -work, -medicine, -training,* etc.), the spread of English has been further strengthened, the validity of which is now indisputable. In the field of scientific terminology, the dominance of English has marginalized almost all other languages (Katsogiannou & Stefanidou 2020). A large number of anglicisms are today widespread internationalisms: AIDS, basketball, cake, dressing, happy end, jackpot.

Loan translations show to what extent and depth the influence of English has penetrated, not only in Modern Greek but also in all European languages. Phrase names, which mask foreign influence and are therefore not in the purview of purists, are today the greatest source of modern language renewal. Some indicative cases that had not been considered until now are *ανοιχτό πανεπιστήμιο* < open university, *αποβιομηχανοποίηση* < deindustrialization, *αυτοεκπλήρωση* (exists as an entry only in the PDAA) < self-fulfillment, *αυτοεκτίμηση* < self-esteem, *αυτόματος πιλότος* < automatic pilot, γραμμή: *γραμμή του κόμματος* (και *κομματική γραμμή*) < party line, *δημοκρατικοποίηση*: διεθνισμός < democratization, *διαψευσιμότητα* < falsifiability, *εγχείρηση ανοιχτής καρδιάς* < open-heart operation, and *πολιτικός γάμος* < civil wedding. In several combinations, the adjective *τεχνητός* is a translation of English artificial: *αναπνοή, γλώσσα, γονιμοποίηση*: respiration, language, insemination. Internationalisms are uncountable: *απομυθοποίηση* <demythologization, French démythification, German Entmythologierung, *κράτος πρόνοιας* < welfare state, German Wohlfahrtsstaat, *ρηματική διακοίνωση* < French note verbal, English verbal note.

The following collocations come from French: *σε αδρές γραμμές* < dans le grandes lignes, *ελαφρύς ύπνος* < sommeil léger, *μια ελαφρά γυναίκα* < une femme légère, *η κατηγορία ελαφρών βαρών* < La catégorie des pods légers, *η μερίδα του λέοντος* < la part du lion, *χαρούμενη νότα* < une note gaie.

In PDAA, many English words and numerous loan translations were lemmatized precisely because they are in widespread use. This describes the current linguistic reality. In many cases, the chronology of a word's appearance or the fixed lexical collocation is recorded for the first time, which illuminates aspects of the history of words and the evolution of science, art, economy, and culture in general. The phrase name *ποιότητα ζωής* denotes the English quality of life, a concept first created in the United States in 1943. In this case, of interest is not only the general meaning 'overall enjoyment of life', but also the specific meaning 'the degree to which a person or group is healthy, comfortable, and able to enjoy the activities of daily living' (Merriam-Webster).

Modern Greek is not at risk from English and Anglo-American loans, for two reasons: They have a low statistical frequency and have not brought about a structural change in the language system.

3.2 The Contempt of Dialects and Other Regional Varieties

The precise definition of the term *official language* presents several difficulties. It denotes at the same time both an established variety and an independent language, a set of linguistic varieties, including dialects. It is necessary for many people to realize that the so-called common Modern Greek cannot be precisely defined, because it consists of many dynamic individual systems and subsystems that are constantly creating new norms. The term 'language of the people' is vague as well, because what one means by the word 'people' must be defined consistently and precisely. The modern language community is not a homogeneous group of people using the same language. It essentially consists of multilingual and multicultural groups united by a complex network of social and professional relationships.

The contribution of Modern Greek dialects and other local and social varieties to the formation of the Modern Greek literary language is also invaluable. Almost all renowned poets and novelists used a multitude of words that were in the mouths of ordinary rural people and in this way enriched common Modern Greek. Particularly, the two Nobel-winning poets Giorgos Seferis and Odysseus Elytis made the best use of unknown or half-forgotten words from Izmir, Cyprus, Crete, and the Aegean islands, thus renewing the poetic discourse.

A representative sample of dialectic words is recorded in PDAA to make the dictionary user aware that the common language has wider dimensions than suspected. From the three great dialects, words that are nationally known were introduced as regular lemmata. Some illustrative examples are as follows: Cretan dialect: *καλτσούνι* ‘sweet cheese pastry’, *κοπέλι* ‘boy, young man’, *πεντοζάλης* ‘kind of dance’, *στιβάνια* ‘tall leather men's boots’. Cypriot dialect: *αγρινό* ‘endemic kind of wild sheep’, *αναρή* ‘kind of soft cheese’, *κουμανταρία* ‘very sweet wine’, *σεφταλιές* 'meatballs'. Pontian dialect: *κεμεντζές* ‘traditional lyre’, *κοτσάκι* 'opposite mating dance’, *περέκ* ‘pie type’, *ωτία* ‘fried ear-shaped sweet’.

3.3 The Myth of a Single Correct Spelling

One of most important innovations of the PDAA is that it debunks spelling in the sense that it negates the stereotype that there is only one correct spelling. When different spellings of a word are widespread, all are recorded: ζήλια & ζήλεια ‘jealousy’, κτίριο & κτήριο ‘building’, ορθοπαιδικός & ορθοπεδικός ‘orthopa(e)dic’. It is worth noting that the Academy of Athens takes a very clear position on this issue. It suggests the spelling that is first in the entry, is even written in a distinct blue colour. The famous dictionaries Duden and Le Petit Robert have served as a model. In the end, the controversial spellings number less than 500 of a total of 75,000 entries. Statistically, this is a negligible percentage. In other words, there is a lot of noise about nothing. The tendency is the predominance of the simplest spelling, while the etymological criterion, which used to be applied in the past with great rigor, does not apply today in many cases.

3.4 The Myth of a Single Correct Etymology

All recent etymological dictionaries, both Greek and foreign ones, overlap, which is to some extent expected when it comes to established etymologies. The search for new etymologies is painful gestation that rarely leads to a happy ending. No serious etymologist believes that he has found the definitive solution to an etymological problem, as there is always the possibility of overturning even the most convincing etymology.

On another occasion, I stressed, many years ago, that the great wronged in teaching Modern Greek is semantics. Since then, things have not changed much in terms of teachers’ perceptions. In the past, language teaching was identified with phonology and morphology, always focusing on spelling, which was and continues to be purely symbolic in nature, in the sense that any spelling simplification is considered by some, at best, a lack of respect for tradition and, at worst, a betrayal of language.

Often, semantic changes are not explained by the history of the Greek language, nor by the evolutionary course of the words themselves. For example, the word *λαγός* ‘hare’ has acquired a second, sports-related meaning: ‘runner who gives fast pace on an endurance race, to help achieve a record by another athlete’. This seemingly inexplicable meaning has its interpretation: it is a loan translation from the French lièvre ‘hare’, which acquired this meaning in 1899. Απασχόληση ‘employment’ comes from the ancient word ἀπασχόλησις ‘distraction’. The modern meaning of ‘paid work for livelihood’ has a different etymological origin. It is a loan translation from the French emploi ‘occupation’ and the English employment.

To show the difference in the way meanings are treated in relation to etymology, I quote the lemma “gazelle” from three comparable Modern Greek dictionaries:

DTR:
γαζέλα η [γazéla] O25**:** είδος μικρής αφρικανικής και ασιατικής αντιλόπης, που είναι περίφημη για τη χάρη των κινήσεών της. || *Γυναίκα σαν ~,* λεπτή, ψηλή και χαριτωμένη. [λόγ. < γαλλ. gazell(e) *-α,* από τα αραβ. (ορθογρ. δαν.)]

DB:
γαζέλα (η) [γαζελών] αντιλόπη τής Αφρικής και τής Ασίας, γνωστή για την ταχύτητα και τη χάρη της. [ΕΤΥΜ. Μεταφορά του γαλλ. gazelle, αραβ. ghazāl].

PDAA:
γαζέλα & (σπάν.) γκαζέλα **1.** ΖΩΟΛ. είδος μικρής αντιλόπης της ασιατικής και της αφρικανικής ηπείρου (γένη Gazella και Procapra), με πυρόξανθο χρώμα, λευκή κοιλιά και μεγάλα κέρατα, το οποίο φημίζεται για τη χάρη, την ταχύτητα και τα μεγάλα άλματά του. **2.** (μτφ.) όμορφη και λυγερόκορμη κοπέλα, συνήθ. μοντέλο, που διακρίνεται για τη χάρη και την κομψότητά της: μαύρη (= Αφροαμερικανίδα)/μελαχρινή ~ του μόντελινγκ/της πασαρέλας. **3.** ΟΙΚΟΝ. (σπανιότ.-μτφ.) ταχέως αναπτυσσόμενη μικρομεσαία επιχείρηση. [< 1, 2: γαλλ. gazelle 3: αμερικ. gazelle (company)].

The third meaning is completely new. None of the English dictionaries (see Onelook. com) record it. There is also a fourth meaning, which is not widely known in Greek: *Τα κέρατα της γαζέλας* 'The horns of gazelle' traditional Moroccan flutes filled with almond paste, almonds, and orange syrup. To the second meaning should be added the expression *μάτια της γαζέλας* (for a woman) < French yeux de gazelle (= big, sweet, and shiny), Italian occhi di gazzella (= big and melancholic).

Most Modern Greek dictionaries, as well as foreign ones, record the word with only one meaning 'a graceful animal'. The relevant entry in PDAA presents three distinct meanings with corresponding etymologies.

In relation to the term semantic polygenesis, I use the term etymological polygenesis to document the legitimacy of double etymologies. According to the theory of semantic polygenesis, a lexical item can appear several times in the history of a language. Each appearance is genetically independent of the others. Multiple etymologies can be interpreted in a similar way, as I suggest, e.g. for the word *γκόμενα* 'chick' (Charalambakis, 2017, 278-279). Until now, the prevailing perception was that there is only one correct etymology of each word.

Perhaps for the first time in the history of etymological research, we have irrefutable evidence that a word can have two etymologies that are equally correct, one older and another newer. The evidence is regarding the adjective *παραολυμπιακός* 'paralympic' *mainly* in the combination of *Παραολυμπιακοί αγώνες* 'Paralympic Games' with synonym *Παραολυμπιάδα* 'Paralympics'. The Oxford English Dictionary (OED) states the following: 'Paralympics: blend of paraplegic and Olympics'. The same etymology repeats Merriam-Webster's, with the date of first appearance in 1953. Petit Robert (see paralympique, first appearance of the word approximately in the year 1960) associates it with paraplégique. On the contrary, Collins etymologizes "parallel + Olympics". Zingarelli's dictionary (see paralimpic & paraolimpico, neologism of 1992), refers to the entry Paralimpiade & Paraolimpiade (neologism of 1988) providing the etymology: "comp. di para- e di (o)limpiade".

The most accurate etymology of the adjective *παραολυμπιακός* is as follows: < English. Paralympics < para(plegic) + (O)lympics, 1953, French paralympique, approx. 1960 & English para(llel) + (O)lympics, 1976.

It should be stressed, however, that contrary to popular belief, etymology does not contribute to the effective use of language, nor does it play a role in communication strategy. It is a highly challenging branch of linguistics that is currently practiced by very few scientists worldwide who experience failure more often than do any other researchers. Toward the end of their careers, they see that very little of their work will survive in the future. The PDAA necessarily succumbed to the etymological stereotype and provides concise etymologies. In fact, it unexpectedly proposes many new etymologies (Charalambakis, 2017), although it is by no means an etymological dictionary.

4. The Stereotype of Ethnocentrism

Ethnocentrism appears in two forms; as a social stereotype, it refers to individuals' or groups' criticism of another culture based on the value system of their own national community, the belief in the uniqueness and/or superiority of the nation to which they belong. This topic is dealt extensively and in a genius way by Fleischer (2020), drawing on examples from DB.

As a linguistic stereotype, ethnocentrism is evident in several etymologies. I would call this phenomenon etymological ethnocentrism. It is well known that in the 18th and especially 19th centuries, some scientists and scholars, in their attempt to prove the unbroken continuity of the Greek language, once reached the limits of hyperbole/exaggeration, insisting on Greek etymologies. Even today, there are some non-specialists who want to eliminate the 'stigma' of foreign words. These 'Greek lovers' could be more careful in expressing their views and less unilateral and absolute if they wanted to deal deeply with the life of words. Almost none of them can imagine that beautiful and well-sounding words, such as *ευκάλυπτος* 'eucalyptus' and *νοσταλγία* 'nostalgia', were not coined by Greeks but passed into Modern Greek through neo-Latin.

I shall confine myself to two typical examples. The exclamation *άντε* is not derived from *άγετε,* imperative (second person plural) of the verb *άγω*, as accepted by DB, following the etymology of G. Hatzidakis. DTR could not unhook itself from this outdated etymology, but it cites as a second possibility the correct etymology from Turkish haydi. For *τσόφλι*, the connection to the hypothetical form *εξώ-φλοιον* (< *έξω* + *φλοιός*) is obviously wrong. This is an etymology of G. Hatzidakis, who suggested a spelling that seems completely strange today, *τσώφλοι(ο)*. More research is needed on the medieval *ceflin*, which is associated with the Arabic *djefl*.

5. The Social Stereotype of Sexism

The international literature on sexism and the way in which the relevant problems are dealt with in dictionaries is overwhelming. For the Greek language, there is not yet an extensive monograph that would cover this major issue. In the past, dictionaries of almost all languages tended to reproduce opinions against women in a completely disparaging way by focusing on the moral side of their personality. For men, on the other hand, there were generally only positive descriptions.[2]

[2] See Charalambakis, 2012, 121-142. Perhaps the relationship between language and sex is examined extensively for the first time in this article.

Three years before I took charge of writing the PDAA, I had read the paper of Encarnación Hidalgo-Tenorio (2000), which impressed me for her penetrating observations on such a sensitive subject in a dictionary that is a milestone in the history of lexicography. The conclusion reached (p.228) also applies mutatis mutandis to the PDAA: 'Therefore, I conclude that this dictionary seems to be an example of what is actually happening in English. Society has developed some stereotypes which language usage itself reinforces; language changes, on the other hand, convey new perspectives in society at the same time, and this dictionary reflects these tendencies sometimes. Whilst it is not committed to eliminating any religious, social, racial, or sexual discrimination, as many could have expected, it aims to introduce new lexical items which no longer allow that distinctiveness to remain'.

Regarding the question of how to put a definitive end to sexist language in the 21st century, there is no clear answer. What is certain is that it is not the language that is to blame for social discrimination, sexism, racism, xenophobia, and so many other prejudices that exist in every society. In the PDAA, an attempt was made to maintain the delicate balance between the actual use of language and the regulatory intervention of the lexicographer, which sometimes reaches the limits of distorting reality. If ordinary words of erotic vocabulary are silenced, as well as swear words and insults that shock much of society, this means that the truth is hidden.[3]

One dictionary, for the entry *ξανθός* ‘blond’ cites as an explanatory example the verse from a folk song ‘some time ago I had an affair with a blonde little girl’ and the expressions: *ανέκδοτα για ξανθές* ‘jokes about blondes’, *οι άντρες προτιμούν τις ξανθές* ‘men prefer blondes’. In the same dictionary, in the entry *μελαχρινός*, we read: *του αρέσουν οι μελαχρινές* ‘he likes brunettes’. Another modern dictionary believes it restores gender equality by quoting the example: *Προτιμά τους μελαχρινούς άντρες* ‘She prefers dark-haired men’. An excellent Modern Greek dictionary records in the entry *κορίτσι* ‘girl’ the example*: Είναι ντροπαλός σαν κορίτσι* ‘He is shy like a girl’. The same example is given in the entry *ντροπαλός*. The meaning of παρθένα‘virgin’ is obsolete and should no longer be mentioned in modern dictionaries, in two of which the following examples are given: *Δεν ήταν κορίτσι όταν παντρεύτηκε* ‘She was not a virgin when she married’ and *Είναι ακόμα κορίτσι* ‘She is still a virgin’. For the expression *είναι κορίτσι από σπίτι* ‘she is a girl from home’ in a very good dictionary the explanation is given: ‘for a moral girl, in good manners, with a good upbringing’. The PDAA uses different wording: ‘with principles, with a good upbringing’. That is, morality, which is unfortunately for many the Achilles’ heel of women, is not mentioned.

In the PDAA, to declare that the adjectives *ξανθός* ‘blonde’ and *μελαχρινός* ‘brunette’ function as nouns, the neutral example is given: *οι ξανθές, οι μελαχρινές*. Of course, it is not right for a dictionary to project the perspective of a man who often sees the woman as a ‘vessel of pleasure’. On the other hand, this phrasal name should not be excluded from a modern dictionary. Its lemmatization could raise awareness of the need for a change of mentality in sensitive social matters. Without seeking the comparison, which would be misplaced anyway, I would say that no one likes environmental pollution, but its recording in dictionaries could raise people’s awareness of ecological issues.

6. The Challenge of Neologisms

The neologisms created each year are innumerable. The PDAA highlights, more than any other dictionary, the great number of neologisms that enrich Modern Greek. (Charalambakis, 2017a). A good dictionary that respects its social, educational, and cultural mission, as well as its users, must regularly revise its lemmata, removing those that have fallen into disuse and adding new ones that have been widely spread. A word or a phrasal name with a new meaning often comes back to the fore. For example, the term *φέικ νιους* (fake news) must also find its place in Modern Greek dictionaries because it does not identify with *ψευδείς ειδήσεις* ‘false news’. It is about false but often sensational information rapidly disseminated mostly through social media. This connotation is evident only in the foreign term. The expression *fake news* might have been witnessed since the late 19th century, but it was described as word of the year by Collins dictionary only in 2017.

Words as meaning bearers carry important messages on their own. By ignoring them, one does not know or realize the upheavals they bring to people's lives. The following words, the examples are indicative, do not exist in any Modern Greek dictionary and should be added to their new editions: αποπαγκοσμιοποίηση, English deglobalization, 2018, French démondialisation, points to the failure of globalization. The word *βάιραλ* ‘viral’ refers to the power of social media. There is an "International Day of Non-Violence" (October 2nd), English nonviolence, 1831, French non-violence, 1921; the term cannot be ignored in Modern Greek dictionaries, even if it is easily understood. Words like *κρυπτονόμισμα* (cryptocurrency, 2009) and *μπιτκόιν* (bitcoin, 2008) came into our lives, and we cannot forget our *μεταμνημονιακές δεσμεύσεις* ‘post-memorandum commitments’. *Μικροϊστορία* ‘microhistory’ opens up new horizons in understanding historical events. *Μικροπλαστικά* (tiny pieces of plastic, less than five millimeters long, that pollute the environment; microplastics, 1990) could awaken ecological consciousness. Overtourism (*υπερτουρισμός*) is beginning to show its negative consequences.

On the occasion of the coronavirus pandemic (*κορονοϊός* < coronavirus, 1969), COVID-19, with its first record in 2020, a series of neologisms emerged (Katsoyannou- Stefanidou, 2020), such as *ασθενής μηδέν* < patient zero, 1987, *ανοσία αγέλης* < herd immunity, 1917*, λοκντάουν* < lockdown. The words *ακίδα* ‘spike’ and *φάκελος* ‘viral/virus envelope’ have acquired new meanings that cannot be ignored.

[3] Lily Thwaites, How do we put an end to sexist language in the 21st century?, theboar.org/2019/10/end-sexist-language-21st-century/

7. Conclusion

The role of dictionaries is crucial for the production and understanding of the language. Having good dictionaries does not automatically make it easier to learn a language. Using dictionaries in teaching is of key importance. At school, students do not practice how to critically evaluate lexicographic information. Two or, even better, three dictionaries are seldom consulted for the same entry. In that case, students will see in practice what lexicographic pluralism means, how each lexicographer illuminates different aspects of the 'life' of words. Dictionaries, like text corpora, are not a panacea. They simply offer useful information, encoded within them, which paves the way for further investigation of words, mainly in terms of their meanings, their stylistic level, and their combinatorial possibilities.

Printed dictionaries must be updated at least every five years. Otherwise, they lose credibility. Words should be not only added but also removed because this is the only way to capture the actual use of language at a given time. The future of lexicography is based on electronic dictionaries. They are updated in real time and widely accessible. Institutional bodies should have the first say because they will guarantee the continuity and consistency of this important work. The Academy of Athens, with the publication of the PDAA, has shown excellent results and can, with State aid, continue its promising lexicographic activities. The Historical Dictionary of the Athens Academy is of paramount importance for Modern Greek dialects. Thus, its scope is de facto limited. Manolis Triandafyllidis Foundation and the Centre for the Greek Language are doing excellent work. However, a 'Language Observatory' is needed to monitor language evolution, especially at the lexical level. It will responsibly inform the general public about language use and facilitate the work of those who teach Modern Greek as a mother tongue and second/foreign language. I wonder why these necessary activities and infrastructure projects, such as updated dictionaries, are not being reinforced for the Greek language, which is simply invoked by some with a rhetorical pomposity to express its greatness as a carrier of national identity and of a brilliant Greek culture.

8. References

Adorno, Th. (1959). Wörter aus der Fremde - Funktion und Gebrauch. In *Akzente*, 6, pp. 176-191.

Bauer, L. and Trudgill, P. (eds.) (1998). *Language myths*. London: Penguin Books.

Boulanger, J.-C. (1986). *Aspects de l'interdiction dans la lexicographie française contemporaine,* Tübingen: Max Niemeyer Verlag.

Charalambakis, C. (2007). Το νέο *Χρηστικό Λεξικό της Νεοελληνικής* της Ακαδημίας Αθηνών. The new Practical Dictionary of Modern Greek of the Academy of Athens. In *Proceedings of the 8th International Conference of Greek Linguistics (ICGL8).* University of Ioannina (CD-Rom Edition), pp. 1263-1282.

Charalambakis, C. (2010). Neohellenic: The present state, in *Greek. A language in evolution*. Essays in honour of Antonios N. Jannaris, C. C. Caragounis (ed), Hildesheim-Zürich-New York: Georg Olms Verlag, pp. 269-292.

Charalambakis, C. (2012). *Νεοελληνικός λόγος. Μελέτες για τη γλώσσα, τη λογοτεχνία και το ύφος. Modern Greek. Studies on language, literature and style.* Athens: M. Romanos.

Charalambakis, C. (2017). Οι ετυμολογικές προτάσεις του Χρηστικού Λεξικού της Νεοελληνικής Γλώσσας της Ακαδημίας Αθηνών. Etymological proposals of Practical Dictionary of Modern Greek Language of the Academy of Athens, In Ch. Tzizilis & G. Papanastasiou (ed.), Greek Etymology, in the series: *Greek Language: Synchrony and Diachrony,* Volume 1, Thessaloniki. Institute of Modern Greek Studies (Manolis Triandafyllidis Foundation), pp. 260-293.

Charalambakis, C. (2017a). Οι νεολογισμοί του Χρηστικού Λεξικού της Νεοελληνικής γλώσσας της Ακαδημίας Αθηνών. The neologisms of the Practical Dictionary of Modern Greek Language of the Academy of Athens. In *4th International Conference of Greek Studies in memory of I.I. Kovaleva*, Moscow, 25-27 April 2017, Summary of Papers, pp. 237-242. Published in *Kathedra of Byzantine and Modern Greek Studies*, 1-2 (3), pp. 274-279.

DB = G. Babiniotis (2019). *Λεξικό της Νέας Ελληνικής Γλώσσας. Dictionary of Modern Greek Language.* 5th ed. Athens: Lexicology Center.

DTR = Λεξικό της κοινής νεοελληνικής. *Dictionary of common Modern Greek*. (1998). Thessaloniki. Institute of Modern Greek Studies (Manolis Triandafyllidis Foundation).

Fleischer, H. (2020). Οι Έλληνες απέναντι στους Άλλους. Εθνικά στερεότυπα και λεξικογραφικές ερμηνείες ταυτότητας. Greeks against the Others. National stereotypes and lexicographic interpretations of identity. In *The Athens Review of Books*, issue 123 (December 2020), pp. 53-62.

Hidalgo-Tenorio, E. (2000). Gender, Sex and Stereotyping in the Collins COBUILD English Language Dictionary. In *Australian Journal of Linguistics*, 20(2), pp. 211-230.

Katsogiannou, M. - Stefanidou, Z. (2020). *Covid 19, Το λεξικό*. *Covid 19, The Dictionary*. Athens: Kavvadia Crew Publications.

PDAA = Academy of Athens*, Practical dictionary of Modern Greek language*. (2014). Athens: National Printing Office.

Sarantakos, N. (2019). *Μύθοι και πλάνες για την ελληνική γλώσσα. Myths and fallacies about the Greek language.* Athens: Publications of PPE.

Acknowledgements

I would like to express my gratitude to the Academy of Athens for continuing support and understanding. I also thank Professor Nikos Lavidas for invaluable comments.

Dictionaries and Morphology

DeCesaris J.

Institute for Applied Linguistics, Universitat Pompeu Fabra
janet.decesaris@upf.edu

Abstract

This paper explores the relationship between word formation and dictionary representation in general purpose monolingual dictionaries of English. The relationship between dictionary representation and morphological structure in languages with inflectional morphology, productive derivation and compounding, and conversion is complex for several reasons and varies across dictionaries. Historically, several important dictionaries of English have chosen to omit words because of their presumed transparent morphological structure. In addition, starting with dictionaries published in the latter half of the 19th century, many dictionaries of English have included affixes and combining forms as headwords, treating these 'partial words' in the dictionary like independent words, yet the information provided in the dictionary about the affix or combining form is often lacking from the standpoint of morphological description. The paper aims to show that while not a frequently discussed topic in current research on lexicography, the relationship between morphological structure and dictionary representation is essential to quality lexicographic products and should be reconsidered in light of digital consultation of dictionaries.

Keywords: Word-formation; inflection; derivation; affixes; compounding; English monolingual dictionaries

1 Introduction

In this paper I consider the relationship between morphology and dictionaries, specifically large-scale monolingual dictionaries. Dictionaries traditionally include definitions and other salient information related to individual words such as pronunciation, etymology, and usage commentary, but many include little or no information on the internal structure of words, how words are structurally related to one another, or how words might combine with other words to produce compounds (in languages in which compounding is productive). The relationship between dictionary representation and morphological structure in languages with inflectional morphology, productive derivation and compounding, and conversion is complex for several reasons and varies across dictionaries.

The impact of morphological structure on dictionary representation has not been a frequent research topic in publications on lexicography in recent years, as evidenced by its very limited presence in important texts such as that by Atkins & Rundell (2008), in which it is afforded only a few pages of discussion in a book over 500 pages long, or by its absence from conference proceedings such as those of EURALEX. Current emphasis, at least in research on dictionaries of English, is on learner's dictionaries, the representation of collocation, and on corpus-based lexicography in general and this has resulted in a tendency to see words as units without internal structure, or at the very least as units the internal structure of which is uninteresting and perhaps even irrelevant to lexicographers. The relative lack of scholarly interest in the relationship between morphological structure and dictionary representation in English, contrasts with the progressive addition of morphological elements like affixes and neoclassical and other combining forms as headwords; these forms play an important role in the morphology of the language yet are not independent words. We also note that many well-respected dictionaries of English have long chosen not to define, or simply to omit, derived words the meaning of which lexicographers assume is known to the dictionary's target users. To the extent that so-called partial words (in Atkins & Rundell's terminology) are now headwords requiring definitions, examples and usage information, and delimiting a dictionary's target audience in the context of digital consultation on the Internet is difficult at best, I submit that it is time to reconsider the role of morphology in dictionaries. I hope to show that this relationship is still of importance and point to how the representation of morphology in dictionaries of English could be improved in quality lexicographic products. In order to do so, I shall consider the general issues at hand and analyse the role morphology has had in a selection of dictionaries of English from the past 150 years.

2 Morphological issues in dictionaries of English

2.1 Overview of English morphology

In order to analyse how morphology interacts with the representation of words in dictionaries of English, it will be useful to first identify which aspects of morphology are particularly relevant to lexicographic projects. Morphology may be divided into two main branches: inflectional morphology and word formation, which, in turn, includes derivational affixation, conversion, compounding, neoclassical compounding, and other, less prominent structures such as blends, initialisms, and acronyms. In the case of English, inflection involves a small number of paradigms and morphemes in comparison with other Indo-European languages. As is well known, English has lost many of its inflections over hundreds of years (Baugh & Cable 1951), and its inflectional morphology has notably fewer forms than that of other

Germanic languages (Putnam & Page 2020). Most inflection in English may be classified as regular and adheres to well-established paradigms and as such is quite straight-forward, although there are a number of frequent forms that are irregular. Word-formation in English, in contrast, is quite complex: there are a large number of word-formation processes involved, with varying degrees of productivity; there are a large number of affixes playing a role in those word-formation processes, and many affixes seem to compete with one another in terms of form but are practically the same in terms of meaning;[1] some affixes are still available to speakers to create new words whereas others are unproductive; conversion, the process by which a word changes its lexical category (for instance, *light*$_{\text{[noun]}}$ → *light*$_{\text{[verb]}}$ or *jump*$_{\text{[verb]}}$ → *jump*$_{\text{[noun]}}$) is extremely productive in English; compounding, especially noun-noun compounding, is difficult to constrain and even describe semantically; blending creates new words based on a combination of phonological and morphological factors and can result in the creation of a new combining form (e.g. *-oholic/-aholic*, arising from words such as *workaholic* or *shopaholic*, created on the model of *alcoholic*); these are just some of the main challenging characteristics facing the analyst of English word-formation..

2.2 Inclusion of inflectional morphology in dictionary entries

Inflectional forms of a word have often been listed as part of the dictionary entries in English, as there are relatively few forms involved. This practice, especially when the inflected form does not fall into the regular pattern, has a long history in English dictionaries. Samuel Johnson, in his *Preface* to *A Dictionary of the English Language*, wrote the following:

Among other derivatives I have been careful to insert and elucidate the anomalous plurals of nouns and preterites of verbs, which in the *Teutonick* dialects are very frequent, and, though familiar to those who have always used them, interrupt and embarrass the learners of our language. (Preface to Johnson 1755: paragraph 21)

Current dictionaries of English often include inflected forms under the headword, regardless of whether they are regular or not. The inclusion of inflected forms in the entries in *The American Heritage® Dictionary of the English Language* (2020)[2] is representative in this respect, and the following inflected forms are listed for the indicated types of words:

- Adjectives: comparative and superlative (if formed by suffixation or suppletion)
- Verbs: simple past, past participle (if different from the simple past), gerund, 3rd person sg. present tense
- Nouns; plural form if irregular; if regular, not expressly listed but often used in examples to display spelling

Dictionaries, of course, are not grammars, but in English they do exercise influence over the standard language, and as a result the inclusion of inflected forms provides valuable information to users who may not know a particular form or who may have doubts concerning the status of a form that they might assume is dialectal. Such variation in English is not uncommon in frequent words; for example, the verbs *dive* and *dream* have two possible preterite forms (*dived*, *dove* and *dreamed*, *dreamt*, respectively), the nouns *index* and *thesaurus* have two possible plural forms (*indexes*, *indices* and *thesauruses*, *thesauri*, respectively), and the debate on whether *toward* or *towards* is correct usage has gone on for more than 150 years (both are correct and common in American English, with use of *toward* being more prevalent; *towards* is more frequent in British English). Providing the standard inflectional form in individual entries in the dictionary never occupied much space in printed volumes because English has few inflections; we note that in dictionaries of languages with many inflected forms, such as Latin or the Romance languages, the dictionary typically identifies the entry as belonging to a specific conjugation or declension and the user must look up the referenced model elsewhere in the printed dictionary. With today's digital consultation, full conjugations and declensions in languages with a significant degree of inflection are often accessed from a click on the landing page, but in English the paucity of forms means that some dictionaries online include inflected words directly on the headword's landing page.

2.3 Inclusion of derivational morphology in dictionary entries

Derived words in dictionaries have been treated in different ways, depending on the degree of lexicalisation of the word. Lexicographers realised early on that regularity in derivational morphology could justify the omission of certain words from the dictionary, thus saving space. Samuel Johnson makes mention of this his Preface, stating that while they are valid words, regular, semantically transparent derivatives such as adjectives ending in *-ish*, adverbs ending in *-ly*, or nouns ending in *-ness* are often omitted from his dictionary because their relationship to the root word is always the same. In fact, however, even for these relatively straight-forward affixes the data are not always so clear. Words like *goodness* or *greatness*, which Johnson lists in his dictionary, display the expected relationship to their stems *good* and *great*, respectively, but have also acquired additional nuances of meaning that should be included in a dictionary (that explains why Johnson did, in fact, define them). He states, "Words arbitrarily formed by a constant and settled analogy […] were less diligently sought, and many sometimes have been omitted […] because their relation to the primitive being always the same, their signification cannot be mistaken" (*Preface* to Johnson 1755: paragraph 34). The fine line between being entirely semantically transparent and only partially so is often blurred. Moreover, most derivational processes in English are not fully productive and have exceptions. In this sense, although lexicographers may state in the dictionary's front

[1] For example, English has many affixes producing nominalizations. For discussion of rivalry in language in general and in morphology specifically, see Štekauer (2018).

[2] The fifth edition of the *American Heritage®Dictionary of the English Language* was published in print form in 2005; entries presented in this paper have been taken from the online version which lists 2020 as its date of publication.

matter that a word's absence from the dictionary does not mean that the word does not exist, many users may not be aware of that proviso, especially when the dictionary is online and the front matter is nowhere to be found.[3]

Another possible way to treat derivational morphology is to list the derived word in the dictionary, but not define it. From the user's point of view, the usefulness of this strategy, which not only saves space but also is a boon to publishers interested in advertising the increased number of entries in the dictionary, depends on both the thoroughness of the list of forms as well as the definitions given for affixes. General purpose dictionaries of English have long afforded affixes, combining forms, and other bound morphemes headword status, but the type of information given for these elements differs greatly from dictionary to dictionary. Some dictionaries classify all word-forming elements as affixes; others discriminate more. Some treat etymologically different sources of a single affixal form as different senses of a single affix, whereas others provide a more detailed—and often etymologically based—analysis. From the viewpoint of morphological description, an affix must include a reference to the morpholexical class of the root or stem to which the affix attaches, a reference to the morpholexical class of the newly created word, some information on the productivity of the affix in current English, and, of course, an explanation of its meaning. It is the nature of many affixes to have abstract meanings because their meaning is both lexical and grammatical: for example, the meaning of a suffix that form nouns from verbs is dependent to some degree on the lexical meaning of the verb, but it is also dependent on the grammatical nature of nouns as opposed to the grammatical nature of verbs. As a result, many dictionary users may not be able to successfully comprehend the abstract definitions of affixes when applied to the definitions of root words.

A special case of derivation that bears directly on the headword list is that of morphological conversion, when a word changes morpholexical class without affixation. The fact that morpholexical classes often display different inflected forms has meant that most dictionaries of English assign words that have undergone conversion to different entries; for example, *jump*$_{\text{[verb]}}$ is a different headword from *jump*$_{\text{[noun]}}$ because some of the inflected forms associated with *jump*$_{\text{[verb]}}$ (*jumped, jumping, jumps*) are different from the inflected form associated with *jump*$_{\text{[noun]}}$ (*jumps*). This approach is taken by the *Merriam-Webster's Collegiate Dictionary* (currently in its 11th edition). Nevertheless, this sort of presentation obscures the obvious close semantic relationship between the two words (in this case, the root verb and the derived noun), and, as a result, has not been adopted by all dictionaries. A competing dictionary, the *American Heritage® Dictionary*, treats the different morpholexical classes as different senses of a single headword.

2.4 Inclusion of compounding in dictionaries

Both regular compounding and neoclassical compounding in English prove particularly problematic for dictionaries because they are highly productive processes and, as Pius ten Hacken has noted:

> In dictionaries for human users, word-formation is usually not seen as a major issue. There is an almost general consensus that can be summarized as follows: it is impossible to achieve completeness because of the productivity of word formation and at the same time unnecessary to aim for it because of the regularity of the new words. Of course, irregular cases should be treated, but there is no need to treat a compound like textbook as being any different from simple words such as textile. (ten Hacken 1998: 157)

This view was expressed by Johnson in the *Preface* to his dictionary (1755, paragraph 33), in which he gives the example that a word like '*woodman'* needs to be defined in the dictionary but a word like '*thieflike'* does not.

The situation is somewhat different for neoclassical compounding, because the constituent parts are not fully independent words in English but rather stems that are used in conjunction with another stem (typically in conjunction with one another). Neoclassical formants are easier to define than affixes because they are based on words with lexical meaning from Latin or Greek. Although lexicographers will not be able to represent all possible neoclassical compounds, there have been some creative attempts to indicate the degree of productivity to users. The *Random House Dictionary of the English Language* (1966) included a novel approach by giving long lists of undefined words with a prefixed neoclassical formant on a divided page. While not exhaustive, the list, which often runs over two or three printed pages and began on the page in which the neoclassical formant was defined, included syllabified words in boldface with prosodic stress marked and morpholexical class (generally *noun* or *adjective*) indicated. This sort of presentation provides users with some insight into the productivity of the formant, and also provides them with standard spelling and pronunciation. In essence, the list of words containing the formant is an extended run-on entry that is alphabetised according to the prefixed combining form.

3 Treatment of morphology in three dictionaries of English

3.1 Dictionaries analysed

Not all dictionaries of English approach the relationship between morphological structure and dictionary representation in the same way. In this section, we discuss issues related to morphology and dictionaries in a selection of three influential general-purpose dictionaries of English published in the past 150 years. Although this survey is necessarily limited in scope and concentrates on dictionaries published in the United States, it will show how different monolingual dictionaries have approached the questions identified in the previous section.

[3] See DeCesaris and Marello (2020) for a discussion of disappearing front matter in some online dictionaries of Spanish and Italian.

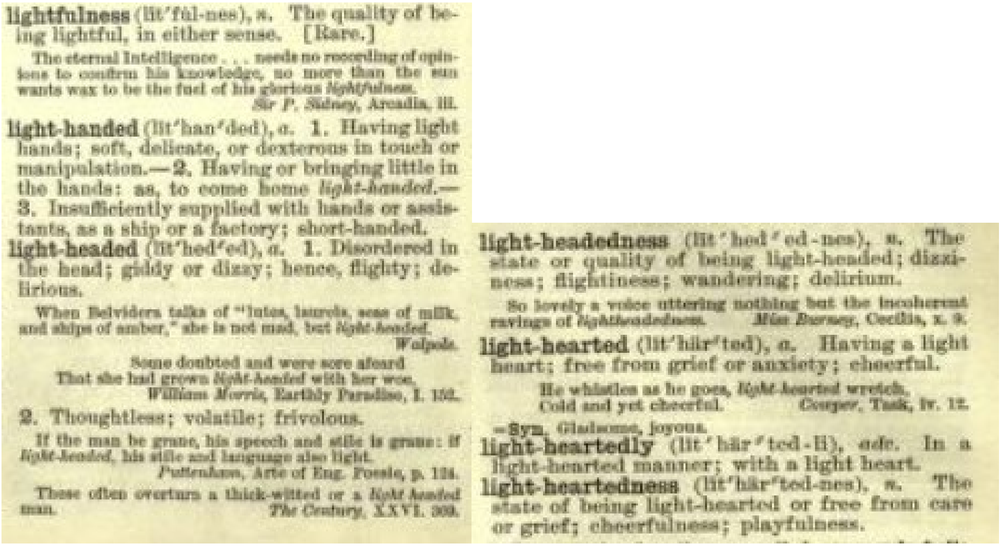
3.1.1 The *Century Dictionary* (1889-1891)

The *Century Dictionary*, edited by the eminent linguist and Sanscrit scholar William Dwight Whitney, is a multi-volume dictionary made on historical principles (Adams 2020). The *Century Dictionary* was based on the *Imperial Dictionary of the English Language* edited by John Ogilvie that had been published in Scotland. It is recognized as one of the great achievements of American lexicography. It is an important dictionary to include in this study not only because of its own influence, but also because it was the basis for two other successful dictionaries, the *American College Dictionary* (1947) and the *Random House Dictionary of the English Language* (1966).

In the preface to the *Century Dictionary*, Whitney gives the following justification for specifically omitting certain types of words from the dictionary:

> No English dictionary, however, can well include every word or every form of a word that has been used by any English writer or speaker. There is a very large number of words and forms discoverable in the literature of all periods of the language, in the various dialects, and in colloquial use, which have no practical claim upon the notice of the lexicographer. A large group not meriting inclusion consists of words used only for the nonce by writers of all periods and of all degrees of authority, and especially by recent writers in newspapers and other ephemeral publications; of words intended by their inventors for wider use in popular or technical speech, but which have not been accepted; and of many special names of things, as of many chemical compounds, of many inventions, of patented commercial articles, and the like. Yet another group is composed of many substantive uses of adjectives, adjective uses of substantives (as of nouns of material), participial adjectives, verbal nouns ending in *-ing*, abstract nouns ending in *-ness*, adverbs ending in *-ly* from adjectives, adjectives ending in *-ish*, regular compounds, etc., which can be used at will in accordance with the established principles of the language, but which are too obvious, both in meaning and formation, and often too occasional in use, to need separate definition. (Preface to the *Century Dictionary*, pg. vi)

In essence, Whitney claims that users of the dictionary are familiar enough with meanings of certain derived words and compounds and their formation that these words need not be included in the dictionary. Given the size and scope of Whitney's dictionary, not including many regularly derived words because they are assumed to be "too obvious" in meaning and formation is an odd decision and contrasts with the practice of listing undefined forms as run-on entries that is adopted in Merriam-Webster dictionaries (as will be shown in §3.1.2). At the very least, providing the written form gives the reader notice that the word exists and is in use, and also establishes the spelling and syllabification. In fact, a cursory look at the dictionary indicates that Whitney did not always follow his own guidelines in this respect, as can be seen in Figure 1.



lightfulness (līt'fůl-nes), *n.* The quality of being lightful, in either sense. [Rare.]

The eternal Intelligence . . . needs no recording of opinions to confirm his knowledge, no more than the sun wants wax to be the fuel of his glorious *lightfulness*. *Sir P. Sidney*, Arcadia, iii.

light-handed (līt'han'ded), *a.* 1. Having light hands; soft, delicate, or dexterous in touch or manipulation.—2. Having or bringing little in the hands: as, to come home *light-handed*.—3. Insufficiently supplied with hands or assistants, as a ship or a factory; short-handed.

light-headed (līt'hed'ed), *a.* 1. Disordered in the head; giddy or dizzy; hence, flighty; delirious.

When Belvidera talks of "lutes, laurels, seas of milk, and ships of amber," she is not mad, but *light-headed*. *Walpole.*

Some doubted and were sore afeard
That she had grown *light-headed* with her woe.
William Morris, Earthly Paradise, I. 152.

2. Thoughtless; volatile; frivolous.

If the man be graue, his speech and stile is graue: if *light-headed*, his stile and language also light. *Puttenham*, Arte of Eng. Poesie, p. 124.

These often overturn a thick-witted or a *light-headed* man. *The Century*, XXVI. 309.

light-headedness (līt'hed'ed-nes), *n.* The state or quality of being light-headed; dizziness; flightiness; wandering; delirium.

So lovely a voice uttering nothing but the incoherent ravings of *lightheadedness*. *Miss Burney*, Cecilia, x. 9.

light-hearted (līt'här'ted), *a.* Having a light heart; free from grief or anxiety; cheerful.

He whistles as he goes, *light-hearted* wretch,
Cold and yet cheerful. *Cowper*, Task, iv. 12.

=Syn. Gladsome, joyous.

light-heartedly (līt'här'ted-li), *adv.* In a light-hearted manner; with a light heart.

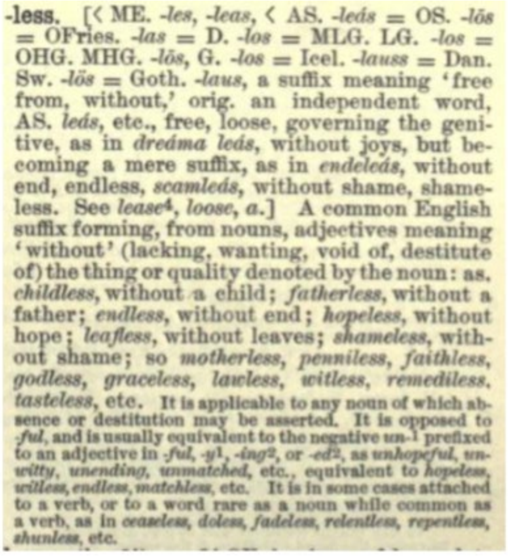
light-heartedness (līt'här'ted-nes), *n.* The state of being light-hearted or free from care or grief; cheerfulness; playfulness.

Figure 1: Entries for compounds and derived words from the *Century Dictionary*, p. 3448.

In this very small excerpt, [4] the dictionary not only lists but also defines an adverb derived with *-ly* (*light-heartedly*), but also three derived nouns with *-ness*, one of which is labelled as rare and the other two of which are, to my mind, formally and semantically "obvious" or transparent (*lightfulness*, *light-headedness*, and *light-heartedness*, respectfully).

The *Century Dictionary* includes many affixes as headwords. The explanation of the affix is quite complete from the standpoint of morphological description: a thorough etymology is given, and the entry identifies the morpholexical class the affix attaches to, the morpholexical class of the newly formed word, and the expected meaning of the newly formed word. Several examples of derived words are listed, and some usage information is usually provided, as can be seen in the entry for *-less* (Figure 2).

[4] Pages in the *Century Dictionary* are printed in three columns. The definition for *light-headed* is at the bottom of the second column on p. 3448 and the entry for *light-headedness* is at the top of the third column on the same page.



-less. [< ME. *-les*, *-leas*, < AS. *-leás* = OS. *-lōs* = OFries. *-las* = D. *-los* = MLG. LG. *-los* = OHG. MHG. *-lōs*, G. *-los* = Icel. *-lauss* = Dan. Sw. *-lös* = Goth. *-laus*, a suffix meaning 'free from, without,' orig. an independent word, AS. *leás*, etc., free, loose, governing the genitive, as in *dreáma leás*, without joys, but becoming a mere suffix, as in *endeleás*, without end, endless, *scamleás*, without shame, shameless. See *lease*^{4}, *loose*, *a.*] A common English suffix forming, from nouns, adjectives meaning 'without' (lacking, wanting, void of, destitute of) the thing or quality denoted by the noun: as, *childless*, without a child; *fatherless*, without a father; *endless*, without end; *hopeless*, without hope; *leafless*, without leaves; *shameless*, without shame; so *motherless*, *penniless*, *faithless*, *godless*, *graceless*, *lawless*, *witless*, *remediless*, *tasteless*, etc. It is applicable to any noun of which absence or destitution may be asserted. It is opposed to *-ful*, and is usually equivalent to the negative *un-*^{1} prefixed to an adjective in *-ful*, *-y*^{1}, *-ing*^{2}, or *-ed*^{2}, as *unhopeful*, *unwitty*, *unending*, *unmatched*, etc., equivalent to *hopeless*, *witless*, *endless*, *matchless*, etc. It is in some cases attached to a verb, or to a word rare as a noun while common as a verb, as in *ceaseless*, *doless*, *fadeless*, *relentless*, *repentless*, *shunless*, etc.

Figure 2: Entry for the suffix *-less* from the *Century Dictionary.*

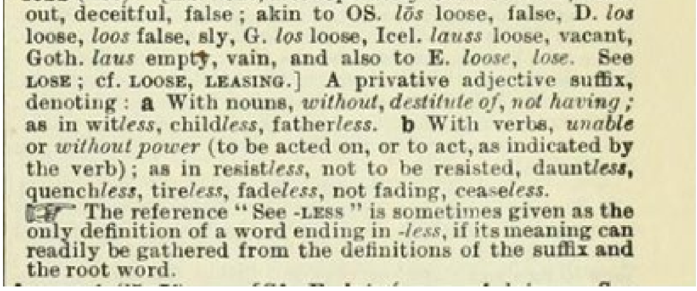
Generally speaking, the definitions for affixes and combining forms in the *Century Dictionary* are the most complete of any general-purpose dictionary of English.

3.1.2 Webster's New International Dictionary (1909)

Noah Webster is the name most associated with American lexicography. The publishing line of dictionaries begun by Webster was continued by the Merriam brothers, who actively took part in the so-called first war of the dictionaries (Adams 2020: 160). Their 1864 edition of *An American Dictionary of the English Language, Royal Quarto Edition*, which incorporated new etymologies by the German scholar C. A. F. Mahn, set the standard for American dictionaries. A completely revised edition of that dictionary was published in 1909 under a new title; the one-volume dictionary had been expanded to 400,000 entries. With so many entries in a large, heavy book, efficient use of space became extremely important. The editors explain the space-saving measures they took in the Preface to the dictionary, and one is directly relevant to morphological structure:

> The third device for saving space is the defining of many purely formal derivatives by references to their prefixes or suffixes. From a primary word or stem, derivatives can be formed, almost at will, by the addition of suffixes like *-hood*, *-ship*, *-ness*, *-ish*, or of such prefixes as *non-*, *anti-*, *contra-*, *infra-*, *super-*, *sub-*, *over-*, *un-*. Any word formed by means of such a general suffix or prefix, although occurring in literature in only one or two of the senses of the main word as modified by the suffix or prefix, might legitimately be used in nearly any other sense appropriate to that of the root word. Great care has been taken to show clearly the meaning of each prefix and suffix in the various combinations in which it may occur, and derivatives have been referred to the proper prefix or suffix, thus leading to an amount of information as to the actual or potential meanings of the derivative that could not possibly be given if each one received independent treatment. By this device the utility of the book has been distinctly increased, and the consulter has also been put in the way of acquiring a knowledge of the force of the formative parts of the English language that might otherwise be overlooked or neglected. (Preface to *Webster's New International Dictionary of the English Language*, p. 6)

The editors assume, in a somewhat cavalier fashion, that most derived words are semantically transparent, and that if affixes are properly defined, users should have no problem in deciphering the meaning of the word at hand. Furthermore, the editors point out that by forcing users to consult both the entry for the affix and that for the root word, users will benefit from becoming more familiar with English word-formation. The definitions of affixes, while quite good, are generally shorter than those given in the *Century Dictionary* because fewer examples are given, and the explanations are less complete. The entry for the suffix *-less* is given in Figure 3.



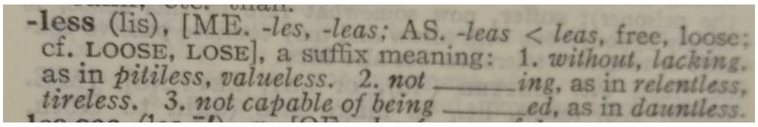
-less (-lĕs). [AS. *-léas*, also separately *léas* free from, without, deceitful, false; akin to OS. *lōs* loose, false, D. *los* loose, *loos* false, sly, G. *los* loose, Icel. *lauss* loose, vacant, Goth. *laus* empty, vain, and also to E. *loose*, *lose*. See LOSE; cf. LOOSE, LEASING.] A privative adjective suffix, denoting: **a** With nouns, *without*, *destitute of*, *not having*; as in wit*less*, child*less*, father*less*. **b** With verbs, *unable* or *without power* (to be acted on, or to act, as indicated by the verb); as in resist*less*, not to be resisted, daunt*less*, quench*less*, tire*less*, fade*less*, not fading, cease*less*.
☞ The reference "See -LESS" is sometimes given as the only definition of a word ending in *-less*, if its meaning can readily be gathered from the definitions of the suffix and the root word.

Figure 3. Entry for the suffix *-less* from *Webster's New International Dictionary* (1909).

The dictionary does not contain run-on entries, a practice which Merriam-Webster adopts in its *Collegiate Dictionary* series.

3.1.3 Webster's New World Dictionary of the American Language (1953)

The period after World War II in the United States was one of great demand for desk-size dictionaries, as servicemen had returned from the war and many were enrolling in colleges and universities across the country, supported by education benefits provided by the federal government. As a result, several successful dictionaries competed at this time for what seemed to be an ever-growing market. Merriam-Webster published the sixth edition of its *Collegiate Dictionary* in 1949, Harper published the *American College Dictionary* (1947) edited by Clarence Barnhart and heavily influenced by studies on vocabulary and reading by Edward Thorndike, and the World Publishing Company headquartered in Cleveland, Ohio published its *Webster's New World Dictionary of the American Language* in 1953 under the direction of David B. Guralnik and Joseph H. Friend. This latter dictionary is particularly interesting with respect to the relationship between morphological structure and dictionary representation for several reasons. First, the editors expressly state that all words entered into the dictionary have full definitions: "Every word entered in this dictionary has been fully defined. Nothing has been left to supposition or guesswork" (*Webster's New World Dictionary of the American Language*, *Guide to the Use of the Dictionary*, p. ix). This dictionary, as opposed to both *Merriam-Webster's Collegiate Dictionary* series and the *American College Dictionary*, contains no run-on entries. The editors, recognizing that many derived words are easily understood from the meanings of their stems and affixes, state in the *Guide to the Use of the Dictionary* that they have omitted such derived words from the dictionary as a space-saving measure, in order to leave more space for words that are not semantically transparent. They justify their stance by stating that their definitions of prefixes and suffixes should allow users "to understand immediately the meanings of such derived words" (*Webster's New World Dictionary of the American Language*, *Guide to the Use of the Dictionary*, p. ix). In essence, they put into practice the guidelines that Whitney had developed for the *Century Dictionary* generations earlier. The definitions of affixes in this dictionary are, in my opinion, quite good in terms of semantics, but lacking in terms of structural information, as information on the morpholexical class of the stem and of the resulting new word is generally missing, as seen in the definition for *-less*, in Figure 4.



-less (lis), [ME. *-les, -leas;* AS. *-leas* < *leas*, free, loose; cf. LOOSE, LOSE], a suffix meaning: 1. *without, lacking*, as in *pitiless, valueless.* 2. *not* ______*ing*, as in *relentless, tireless.* 3. *not capable of being* ______*ed*, as in *dauntless.*

Figure 4. Entry for the suffix *-less*, *Webster's New World Dictionary of the American Language* (1953).

Moreover, the entry would have benefitted from additional examples such as those found in either of the dictionaries previously discussed, but presumably space considerations prevented the editors from including many more examples in this desk-size dictionary.

This dictionary also takes a different approach to the results of morphological conversion. In cases in which the word is not overly polysemous and the semantic relationship between the words belonging to different morpholexical categories is transparent, all categories of a word are defined under a single headword. For example, the word '*broadcast'* is entered as a single headword and the various uses of the word—as a verb, adjective, noun, and adverb—are indicated for each sense (all uses share the same pronunciation). We can contrast that type of representation with that given in *Merriam-Webster's Collegiate Dictionary*, which contains three headwords for '*broadcast'* and lists the use of the word in the derived category of adverb as a run-on entry to the definition of '*broadcast'* as an adjective. As a result, *Webster's New World Dictionary of the American Language* has fewer headwords than the Merriam-Webster dictionary, although the coverage of word meaning in the two dictionaries is quite similar.

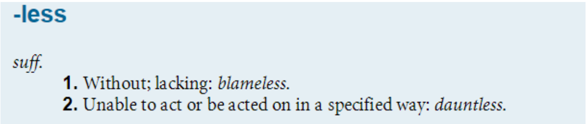
4 Discussion

A brief look at the relationship between morphological structure and dictionary representation in a few dictionaries of English yields a number of observations. First, lexicographers in the past understood dictionaries as reference works to be used in tasks of reading comprehension by native speakers, and as such make assumptions concerning how much information the dictionary's target audience can be expected to know. This assumption has led many dictionaries to either omit words considered to be semantically transparent by a majority of users or enter them into the dictionary without any definition at all, usually at the end of the entry corresponding to the derived word's stem. Compounds do not fare any better, as lexicographers as early as Johnson in the mid-18th century again justified their absence from dictionaries on the basis of semantic transparency. We note that in today's context of dictionary consultation, which generally takes place online and often on a small device like a telephone, the assumption that dictionaries are almost exclusively used in comprehension tasks is outdated. I would also suggest, rather impressionistically, that although dictionaries can certainly be targeted for use by native speakers, the amount of derivational morphological knowledge possessed by speakers may not be as homogeneous across the speech community as assumed. Interestingly, lexicographers never assumed such homogeneous knowledge of inflected forms, which are regularly provided and are seen as complying with the authoritative function of dictionaries.

Second, morphological conversion, which is certainly one of the most salient features of English word-formation, has been treated in several different ways by different lexicographers. Dictionaries which combine forms belonging to different morpholexical categories under a single headword (e.g. *Webster's New World Dictionary* and the *American*

Heritage® Dictionary) are better at displaying the semantic relationship across the forms, but the user must be attentive enough to see that different word classes have been brought together. At least for searches in a digital context, it is probably faster for users to have a drop-down menu at their disposal to choose from among the definitions of the word used as a noun, verb, or adjective, as opposed to scrolling down through a long entry. A larger number of headwords does not necessarily mean that the dictionary covers more meanings.

Third, the entries that dictionaries provide for affixes vary considerably from dictionary to dictionary. The two dictionaries from the 19th century actually provide quite complete descriptions of affixes that are, in fact, better from a linguistic standpoint than the descriptions provided by some current dictionaries. Perhaps the comparison is unfair because both of the 19th century dictionaries were much larger in size and scope than current desk-size dictionaries, but a simple look at the entry for *-less* in the *American Heritage®Dictionary* in Figure 5 shows that it is much less thorough than the definitions provided by either of the older dictionaries, and less informative than the comparably sized *Webster's New World Dictionary* (Figure 4).



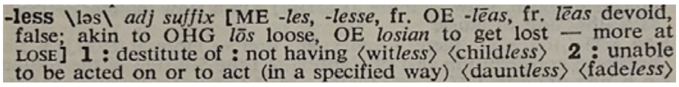
-less

suff.

1. Without; lacking: *blameless.*
2. Unable to act or be acted on in a specified way: *dauntless.*

Figure 5. Entry for the suffix *-less* in the *American Heritage Dictionary* online (2020).

The entry for *-less* in *Merriam-Webster's Collegiate Dictionary* in Figure 6 is a bit more informative in that it indicates that *-less* forms adjectives, but it is still less so than the definition in the comparably sized *Webster's New World Dictionary* (Figure 4).



-less \ləs\ *adj suffix* [ME *-les, -lesse,* fr. OE *-lēas,* fr. *lēas* devoid, false; akin to OHG *lōs* loose, OE *losian* to get lost — more at LOSE] **1 :** destitute of **:** not having ⟨wit*less*⟩ ⟨child*less*⟩ **2 :** unable to be acted on or to act (in a specified way) ⟨daunt*less*⟩ ⟨fade*less*⟩

Figure 6. Entry for the suffix *-less* in *Merriam-Webster's Collegiate Dictionary* (2004).

Finally, the inclusion of a list of undefined words resulting from neoclassical compounding in the *Random House Dictionary* is interesting in that it attempts to deal with the impossible task of representing productivity in a static reference work. Dictionaries are not meant to be grammars and cannot be expected to explain how productive a particular combining form or affix is, but by providing a long list of words containing the combining form or affix users are given insight into the issue (not to mention guidance on pronunciation). This practice from the 1960s could easily be adapted to an online format, with users being able to access a list of forms from the landing page of the definition of the combining form or affix.

5 Conclusion

Dictionaries of English have long afforded affixes, combining forms and other bound morphemes headword status as a space-saving measure. Native speaker users, who have been assumed to know the derivational morphology of the language, are often expected to apply information present elsewhere in the dictionary to words the dictionary has listed, but not defined. This measure was designed to save space in print, but digital users are not necessarily aware of that; in the end, their dictionary look-up may turn out to be a frustrating experience, because the dictionary wants them to supply the definition but they see that—precisely—as the job of the dictionary. Some derived words, and many compounds, were expressly omitted from dictionaries in print because their meaning (and pronunciation) were all assumed to be transparent to speakers. These observations are not meant as criticisms because the dictionaries discussed herein were all published initially before digital consultation was possible. Nevertheless, now that consulting dictionaries online has become the norm as opposed to the exception, we should take the opportunity to reconsider some of the take-aways from our brief analysis with a view to improving our lexicographic products.

Treating 'partial words' which typically both have lexical meaning and play an important role in grammar as if they were independent words is more complicated than just providing information on meaning; in order to process the use of *-less* correctly, it is advisable to know what sort of stem it attaches to and it is essential to know the morpholexical category of the newly created word. This information could surely be added to online dictionary entries.

The practice of omitting words because speakers are assumed to be able to work out their meanings on the basis of their constituent parts developed because space in print was costly. To the extent that that cost factor is no longer applicable in a digital context, it needs to be readdressed. Much current work in corpus lexicography in English is concerned with incorporating collocations into dictionaries, but what about compounds? Are they not individual words worthy of a lexicographer's attention?

The advantages or disadvantages of entries that combine morpholexical categories under a single headword as opposed to positing several independent headwords need to be studied empirically. My initial hypothesis is that online dictionaries that combine entries may be more difficult for users to navigate on a small device, but this needs to be tested with groups of native speakers. Much work has been done in testing how learners of English process the information in online dictionaries with a view to improving those dictionaries, but to my knowledge there has been much less enquiry into how to improve dictionaries for native speakers.

Finally, in a context in which general purpose dictionaries of English must compete with other online resources, lexicographers should rethink the role of the native speaker monolingual dictionary as only a reference tool for text comprehension. Attempting to capture a combining form's productivity, as the *Random House Dictionary* did over fifty years ago, could be a starting point. Learners' lexicography has shown that dictionaries can play an important role in text production, and general-purpose dictionaries should at least consider how the information they already have at their disposal could improve text production by native speakers. As long ago as in 1909, the editors of *Webster's New International Dictionary* stated that by taking advantage of entries for affixes, they could help to inform users of the language's internal morphological structure that may be unknown to them. A noble goal indeed, and one that quality lexicographic projects should embrace.

6 References

Adams, M. (2020). The Making of American English Dictionaries. In S. Ogilvie (ed.) The Cambridge Companion to English Dictionaries. Cambridge: Cambridge University Press, pp.157-169.

American College Dictionary (1947). New York: Harper & Brothers Publishers.

Atkins, S.B.T., Rundell, M. (2008). The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Baugh, A. C., Cable, Th. (1951). A History of the English Language. London, Boston, and Henley: Routledge & Kegan Paul.

Century Dictionary, An Encyclopedic Lexicon of the English Language (1889-1891). New York: The Century Co. Accessed at: www.biodiversitylibrary.org [02/25/2021].

Chisholm, W. (ed.) (1996). Centennial Celebration of the Century Dictionary. In Dictionaries: Journal of the Dictionary Society of North America 17, pp. 1-125.

DeCesaris, J., Marello, C. (2020). Perspectives on front matter in monolingual dictionaries of Spanish and Italian. In Lexicography, 7, pp. 135-149.

Johnson, S. (1755). A Dictionary of the English language. London: J. F. And C. Rivington. 2 volumes. Sixth edition (1785). Accessed at: https://publicdomainreview.org/collection/samuel-johnson-s-dictionary-of-the-english-language-1785 [02/03/2021].

Merriam-Webster's Collegiate Dictionary, 11th edition (2004). Springfield, MA: Merriam-Webster.

Putnam, M. T., Page, B. R. (2020). The Cambridge Handbook of Germanic Linguistics. Cambridge: Cambridge University Press.

Random House Dictionary of the English Language, Unabridged Edition (1966). New York: Random House.

Štekauer, P. (2018). Competition in natural languages. In J. Santana-Lario, S. Valera (eds.) Competing Patterns in English Affixation. Frankfurt: Peter Lang, pp. 14-31.

ten Hacken, P. (1998). Word Formation in Electronic Dictionaries. In Dictionaries: Journal of the Dictionary Society of North America 19, pp. 158-187

The American Heritage® Dictionary of the English Language, Fifth Edition (2020). Boston: Houghton Mifflin Harcourt Publishing Company. Accessed at: www.ahdictionary.com [06/02/2021].

Webster's New International Dictionary of the English Language (1909). Springfield, MA: G. & C. Merriam.

Webster's New World Dictionary of the American Language. (1953). Cleveland and New York: World Publishing Company.

Lexicographic treatment of salient features and challenges in the creation of paper and electronic dictionaries

Prinsloo D.

Department of African Languages, University of Pretoria
danie.prinsloo@up.ac.za

Abstract

This paper focuses on the need for lexicographers to study and to treat the salient features of languages satisfactorily and the challenges faced by lexicographers. The focus is on the challenges facing compilers of African language dictionaries and the lack of dictionaries for these languages. It will be argued that lexicographers are expected to fulfil the role of mediators between complicated grammatical structures, on the one hand, and the target users' needs and expectations, on the other. Dictionaries are expected to be inclusive, e.g., providing for and fulfilling user expectations by giving all the required information in the dictionary in order to reduce the need for consultation of external sources. Expectations for future compilation of paper and electronic dictionaries are discussed. It is expected that paper dictionaries will be used in Africa for many years to come but that paper and electronic dictionaries of high lexicographic quality should be compiled simultaneously. The discussion is presented against the background of the transition of African lexicography from Euro-centred dictionary compilation to Afro-centric compilation. African language dictionaries are continuously compiled in Africa, by Africans for Africans.

Keywords: dictionaries; lexicographic treatment; salient features; challenges; African languages

1 Introduction

Lexicographers must make sure that the salient features of the language or languages treated in their dictionaries are well studied and comprehended by themselves before embarking on the arduous task of lemmatising and treating them. What could be an issue in a specific language might be non-problematic and straight forward in another, or in the other member of the language pair in a bilingual dictionary. Lexicographers, although being mother-tongue speakers of the language(s) treated in their dictionaries, should never assume full knowledge of all the salient features of these languages. Many examples of salient features that were missed in dictionaries were detected. In addition to in-depth knowledge of the grammar of the language(s), the lexicographer should also consider all the relevant external issues and challenges impacting on the compilation of the dictionary. A number of specific lexicographic initiatives in Africa involving community engagement will be discussed.

Lexicographers, unfortunately, do not live in an ideal world. They are faced by a multitude of challenges. In this presentation, African languages, specifically the Bantu language family,[1] will be considered as a case in point, i.e., how their salient features should be detected and treated in terms of the intrinsic challenges pertaining to the language(s) as well as how the challenges posed by the environment or setting that the dictionary has to be compiled within should be handled. Challenges pertaining to the language regard, e.g., morphology, syntax, semantics, and pronunciation, as well as compilation traditions and extra-linguistic factors, such as financial and political issues. The aim of this paper is to give an overview of the salient features and main challenges of dictionary compilation for these languages. The aim is neither to provide a mere listing of problematic issues nor to attempt detailed discussion within the limitations of a conference paper. References to resources where the key issues are discussed in more detail will be given for the interested reader. The main focus will be on the impact of these challenges on dictionary compilation for African languages and to suggest best practices for the lexicographer in order to meet them.

2 The Status of Dictionaries and Lexicographic Initiatives

A study on lexicography in Africa, edited by Hartmann (1990), is taken as a point of departure. Thirty years ago, he

[1] The term 'Bantu' got stigmatized during the Apartheid Era in South Africa. Therefore the term 'African' is preferred in South Africa even in reference to what is internationally referred to as 'Bantu languages'. The discussion in this paper is however focused on the Bantu language family and most of the issues described cannot necessarily be generalized to be applicable to other languages on the continent of Africa. To respect the view of those opposed to the term 'Bantu', it will only be used in cases where a distinction between African languages (languages spoken in Africa) versus a member of the Bantu language family is essential.

conducted a study of lexicography in different regions of Africa, e.g., East, West, South, etc. The opinions echoed by the researchers were that dictionaries for African languages were not of a high lexicographic quality. The main reason given was that existing dictionaries reflected a Euro-centric approach. They were compiled mostly by missionaries from abroad to fulfil their goals, i.e., to assist the missionaries to understand African languages in order to spread the gospel. Such dictionaries were, therefore, not in the first place intended to serve the needs of Africans. In the past decade, the need for dictionaries compiled in Africa by Africans themselves, primarily for speakers of African languages as target users gained momentum, which can thus be called an "Afro-centric approach" to dictionary compilation. Portraying European/western culture instead of African reality is a typical shortcoming in many dictionaries. So, for example, Taljard and Prinsloo (2019: 210) quote an instance where the concept "my house" is represented by an illustration that is unmistakably a typical European dwelling instead of a variant of the houses seen in Africa. In the past decade, the move to Afro-centric dictionary compilation coincided with the decolonisation drive. Several initiatives by entrepreneurs, publishing houses and government-supported agencies were undertaken to give wings to dictionary compilation with a true Afro-centric approach such as IKS (Institute of Kiswahili Studies, https://www.udsm.ac.tz/web/index.php/institutes/iks/the-history-of-the-institute), the Allex Project (http://www.edd.uio.no/allex/aims.html), and the nine national lexicographic units (NLUs) for African languages in South Africa. The requirement for NLUs was the compilation of comprehensive monolingual dictionaries for these languages, which was funded by government. The expectation was that the speech communities of the different languages will eventually take full responsibility for dictionary compilation, including providing the necessary financial resources. There were difficulties faced by these lexicographic initiatives. So, for example, Wolvaardt (2017) reports negatively on the actions of the Pan South African Language Board (PanSALB) responsible for funding and guidance of the South African NLUs.

> […] the national lexicography project, pioneered in the early years of South Africa's democratic transition by some of the country's greatest language activists and academics, […] permitted to degenerate into the scattered efforts of a diminishing band of lexicographers? […] into perpetual begging for adequate funding, the National Lexicography Units (NLUs) hover on the verge of extinction. […] leaves the NLUs where we find them today, desperately trying to justify their existence by producing dictionaries, which, by and large, are based on their feasibility within the constraints of limited funding rather than on any coherent overarching plan. (Wolvaardt 2017: 9)

Financial support from speech communities also did not materialise. The NLUs still rely on PanSALB.

A degree of community engagement, which can be compared to crowdsourcing, materialised where speakers of the language contribute to the extension and updating of the dictionary, e.g., the Xitsonga-English dictionary (https://www.xitsonga.org/dictionary) where the community is involved in correcting dictionary information. Another good example of dedicated community involvement is the Ju|'hoan Children's Picture Dictionary (Jones & Cwi 2014a). In its self-description (Jones & Cwi 2014b), the compilation of this dictionary is described as a collaborative project between the Namibian Ju|'hoan from the Tsumkwe region and academics from various fields. This dictionary clearly indicates an Afro-centric approach to dictionary compilation.

Financial aspects and the fact that African languages are severely under-resourced constitute a major problem in many ways for the lexicographer. In an overview by the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL, 2014), it is stated that "under-resourced languages suffer from a chronic lack of available resources (human-, financial-, time- and data-wise)." This is absolutely applicable to African languages: Compilers of dictionaries for African languages are severely limited in the number of lemmas that can be treated, on the exhaustiveness of treatment of these lemmas, and on the number of pages allowed. This leads to the undesirable situation where the lexicographer must choose between lemmatising, say, 15,000 lemmas with the treatment limited to a few translation equivalents, or 5,000 lemmas with slightly elaborated treatment, still barely fulfilling the need for text reception or dictionary use on demand. See the discussion below regarding dictionaries for text production.

3 Dictionary Compilation for Specific Target Users

Prospective compilers are faced with a situation where dictionaries are required for several thousands of African languages spoken on the content of Africa. Many of these languages do not even have a single dictionary as a reference source. Therefore, the first challenge is to compile, say, a monolingual and a bilingual dictionary for the specific language. Bilingual dictionaries in Africa usually bridge the African language with major languages of the world such as English and French. Lexicographers could depart from the revision of existing dictionaries, where available, or opt for starting afresh with a new compilation.

4 Introspective versus Corpus-based Dictionary Compilation

The advantages and disadvantages of introspective versus corpus-based dictionaries should be carefully considered. It is generally accepted that the utilisation of a corpus can enhance the lexicographic quality of a dictionary on both macrostructural and microstructural levels. In the absence of a corpus, which is the case for most African languages, lexicographers have no option but to compile the dictionary on introspection. The downside of introspective compilation is that words most likely to be looked for by the target users can easily be left out simply because, in the words of Snyman et al. (1990) in *Dikišinare ya Setswana English Afrikaans Dictionary* (DS), they "did not cross the compilers' way". Studies by De Schryver and Prinsloo (2000a) indeed reveal many instances of lemmas most likely to be looked up which are simply not in the dictionary. However, De Schryver and Prinsloo (2000b) also indicate that consistent application of introspection over time can render good quality lemmalists. For many African languages, it is possible to compile corpora albeit relatively small ones, e.g., comparable in size to initial English corpora such as the Brown Corpus consisting of only one million words. Prinsloo (2015) indicated that even such limited corpora can go a long way in assisting the lexicographer with lemmatisation, sense distinction selection of authentic examples, frequency indication in the dictionary, etc.

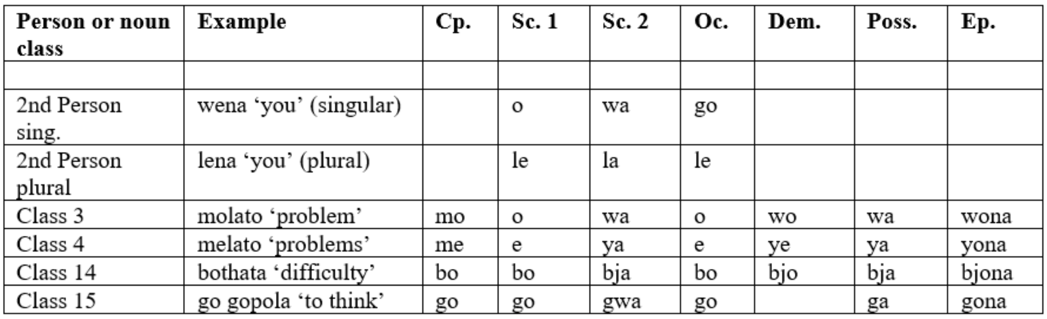
5 Words most Likely to be Looked for and User Expectations

The importance of the user perspective has been echoed several times in the literature emphasising the basic fact that dictionaries are judged as good or bad by their users, cf., Gouws and Prinsloo (2005). Many African language dictionaries can be regarded as examples of linguistic achievement but are not user-friendly. Haas' (1962: 48) remark is still relevant after six decades: "a good dictionary is one in which you can find the information you are looking for — preferably in the very first place you look"; likewise, Barnhart (1962: 161) states that "the function of a popular dictionary [is] to answer the questions that the user of the dictionary asks".

Ideally, dictionaries should be compiled for very specific target users, but when the first dictionary for a language is compiled, the only option for the African language lexicographer is to compile a dictionary that can be used by all users, i.e., an unfortunate attempt towards a one-size-fits-all dictionary. In the compilation of such general dictionaries, the lexicographer should maintain a sound balance between descriptiveness and prescriptiveness. On the one hand, prescriptiveness is required, especially in cases where the language is not fully standardised; on the other hand, the lexicographer should guard against excessive purism, e.g., resisting pressure not to enter any loan words in the dictionary.

6 The Lexicographer as Mediator between the User and Complicated Grammatical Systems

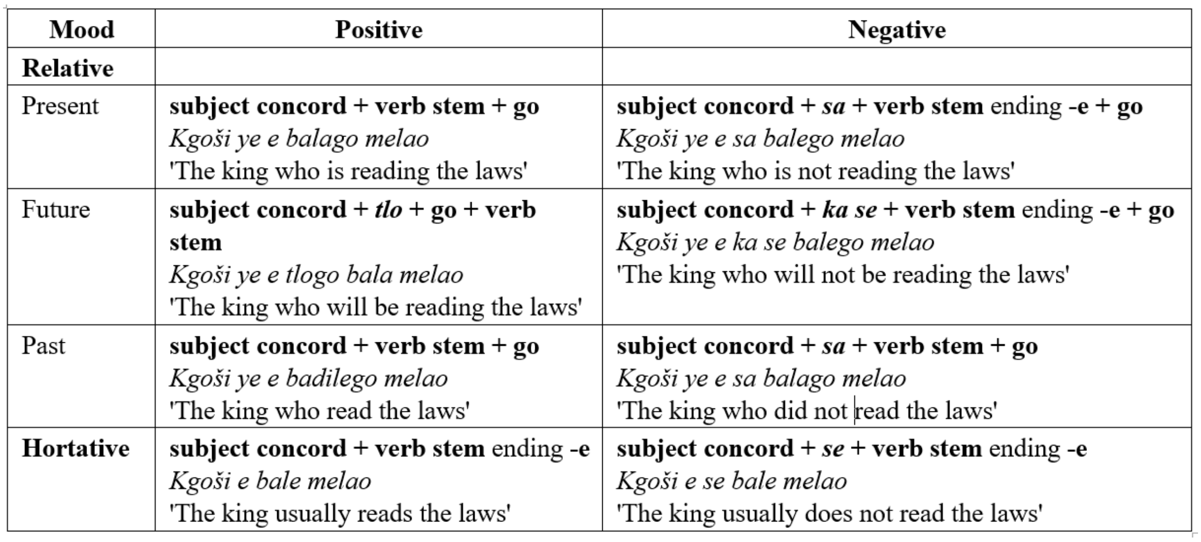
African language lexicographers find themselves in the role of mediators between user expectations and complicated grammatical structures. It is not claimed that African languages are the only languages in the world with complicated grammatical systems; the point is that the lexicographer should be fully acquainted with the core grammatical systems in the language(s) treated. For members of the Bantu language family in particular, these core systems are complicated nominal and verbal systems. Nouns are classified into different classes, each generating different sets of concords and pronouns, which are not interchangeable and are elements required to complete sentences and phrases. Verbs occur in eight moods, see Tables 1 and 2 as well as Prinsloo (2020a and 2020b) for a detailed discussion.



Person or noun class	Example	Cp.	Sc. 1	Sc. 2	Oc.	Dem.	Poss.	Ep.
2nd Person sing.	wena 'you' (singular)		o	wa	go			
2nd Person plural	lena 'you' (plural)		le	la	le			
Class 3	molato 'problem'	mo	o	wa	o	wo	wa	wona
Class 4	melato 'problems'	me	e	ya	e	ye	ya	yona
Class 14	bothata 'difficulty'	bo	bo	bja	bo	bjo	bja	bjona
Class 15	go gopola 'to think'	go	go	gwa	go		ga	gona

Key: Cp. = class prefixes of the noun; Sc. = subject concords; Oc. = object concords; Dem. = demonstratives; Poss. = possessive concords; Ep. = emphatic pronouns

Table 1: Extract from the noun class system in Sepedi.



Mood	Positive	Negative
Relative		
Present	**subject concord + verb stem + go** *Kgoši ye e balago melao* 'The king who is reading the laws'	**subject concord + *sa* + verb stem** ending **-e + go** *Kgoši ye e sa balego melao* 'The king who is not reading the laws'
Future	**subject concord + *tlo* + go + verb stem** *Kgoši ye e tlogo bala melao* 'The king who will be reading the laws'	**subject concord + *ka se* + verb stem** ending **-e + go** *Kgoši ye e ka se balego melao* 'The king who will not be reading the laws'
Past	**subject concord + verb stem + go** *Kgoši ye e badilego melao* 'The king who read the laws'	**subject concord + *sa* + verb stem + go** *Kgoši ye e sa balago melao* 'The king who did not read the laws'
Hortative	**subject concord + verb stem** ending **-e** *Kgoši e bale melao* 'The king usually reads the laws'	**subject concord + *se* + verb stem** ending **-e** *Kgoši e se bale melao* 'The king usually does not read the laws'

Table 2: Extract from the verbal mood system in Sepedi.

Lexicographers should serve the users with lexicographic inclusiveness and present the information in such a way that users can find what they are looking for in and what they need to understand from the dictionary. Users should not be obliged to consult external sources, such as grammatical descriptions of the language, which in most cases do not exist anyway. Of specific importance here is the work of Gilles-Maurice de Schryver (2010) in which he attempts to "revolutionize African language lexicography" as well as the publication of the *Oxford Bilingual School Dictionary: Zulu and English* (OZSD) in which the stem tradition for the lemmatisation of nouns was abandoned — nouns were lemmatised as full orthographic words.

7 Paper versus Electronic Dictionaries

Naturally, dictionary compilation of African languages does not stand in isolation — it is influenced by trends and changes in international lexicography. So, for example, the need to compile and consult corpora became an important and desired aspect of dictionary compilation in Africa. Likewise, the dawn of the electronic era brought new opportunities but also new challenges. Most significant is the resolution of stem identification problems in lemmatisation for conjunctively written African languages in electronic dictionaries (see also below). A major challenge, however, is producing good paper and electronic dictionaries simultaneously for African languages. An extreme approach could be to stick to the compilation of only paper dictionaries until paper dictionaries of high lexicographic quality are available for most African languages. The other extreme is to discontinue paper dictionary compilation and disregard lexicographic traditions and approaches in order to focus only on electronic dictionaries. Such an approach would be in line with the decision announced by Michael Rundell in 2012 that Macmillan decided to discontinue printed dictionaries. Rundell (2012: 74) even said "in an ideal world, we would pulp most of this and start from scratch, producing new resources optimally adapted to digital media". Starting afresh was indeed tempting given the African language lexicographic situation. Such a decision would also "free" the lexicographer from the many restraints and misinterpretations about lemmatisation strategies and alphabetical ordering. One of the biggest frustrations to compilers of dictionaries for African languages is the misconception that stem lemmatisation is more scientific than word lemmatisation. Blindly following this belief resulted in situations where the stem tradition was also followed for languages in which words are disjunctively written, ideal for full-word lemmatisation and for instances in which neither the lexicographer nor the user knows what the stem of the noun is in order to look it up. See Van Wyk (1995) for a detailed discussion. The lexicographer ends up in a minefield of lemmatisation approaches, conjunctive versus disjunctive writing systems, and lexicographic traditions.

Rundell (2015: 303) believes that "in many parts of the world, paper dictionaries still have a healthy future ahead of them. He says that "certain types of dictionary — such as those designed for schools, or special-subject dictionaries, or dictionaries of "smaller" languages — may show a preference for print for some time to come". The reality in Africa is indeed that paper dictionaries are expected to be relevant for many years to come. Phillip Louw, Head of Dictionaries and Dictionary Data, Oxford University Press, Cape Town, South Africa (in email correspondence) emphasises that "the dominance of paper dictionaries in the school dictionary market in Africa [is expected] to continue for at least the next ten years".

Thus, when it comes to paper dictionaries versus electronic dictionaries, the recommended approach would rather be for African language lexicographers to persevere with the improvement and compilation of paper dictionaries but also to embark on the compilation of electronic dictionaries for African languages. They should, however, be careful to avoid the typical pitfalls international lexicography fell prey to in the transition from paper dictionaries to electronic dictionaries. These pitfalls mainly revolve around the presentation of paper dictionaries "on computer" with perhaps only a few added electronic features, such as search functions. Sue Atkins (1996: 515-516) is quite adamant on this issue and bluntly states:

> […] dictionaries of the present […] may even come to you on a CD-ROM rather than in book form, but underneath these superficial modernizations lurks the same old dictionary. […] It is up to us to take up the real challenge of the computer age, by asking not how the computer can help us to produce old-style dictionaries better,

> but how it can help us to create something new.

The important aspect to realise is that it is a new process in which the features enabled by the computer should be maximally utilised. Such features can be called “true electronic features”. Gouws and Tarp (2017: 391) list the following important features applicable to all e-dictionaries:

- Improved search methods and access routes;
- User-based data filtering;
- Less compact article formats with items representing different data categories placed in separate lines;
- Abolition of abbreviations;
- Use of metatexts to introduce sections with specific data categories;
- Use of hidden data, that is data that are not always on display but can be called up when needed;
- Use of pop-up windows and hypermedia to present additional data;
- Inclusion of video and audio options;
- New forms of internal and external linking;
- Interaction between lexicographer and user;
- Continuous updating.

In the same way as OZSD, electronic dictionary designs for Sepedi should include all the required information in the dictionary without the user having to consult external sources. This is, among other things, obtained through a network of pop-up information activated by hovering over or clicking on items for which the user needs more information. Consider a summary of hovering and clicking options available to the user when consulting the Sepedi multi-word lemma *ka se* in Figure 1.

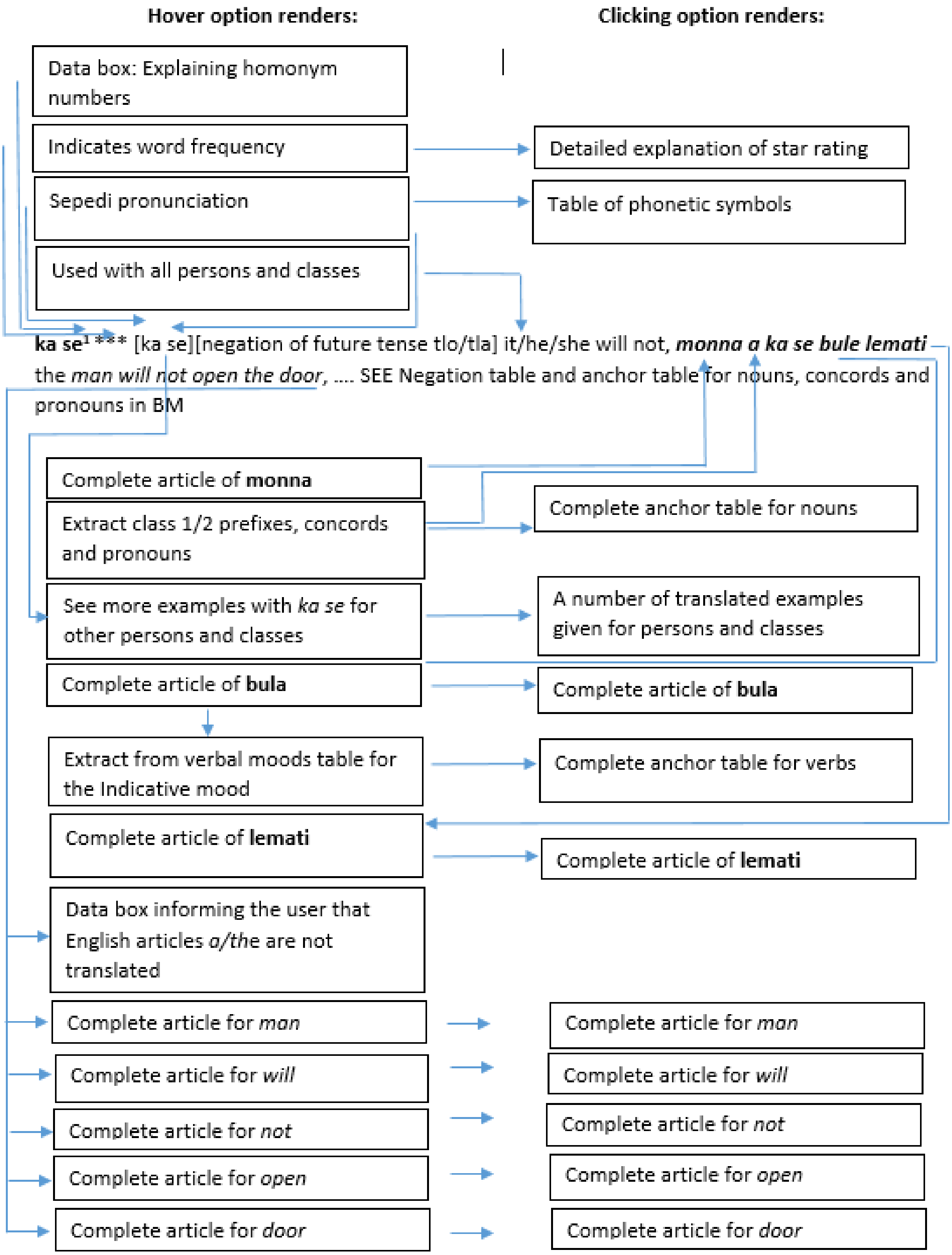

Hover option renders:
Clicking option renders:
Data box: Explaining homonym numbers
Indicates word frequency
Detailed explanation of star rating
Sepedi pronunciation
Table of phonetic symbols
Used with all persons and classes
ka se1 *** [ka se][negation of future tense tlo/tla] it/he/she will not, monna a ka se bule lemati the man will not open the door, ... SEE Negation table and anchor table for nouns, concords and pronouns in BM
Complete article of monna
Extract class 1/2 prefixes, concords and pronouns
Complete anchor table for nouns
See more examples with ka se for other persons and classes
A number of translated examples given for persons and classes
Complete article of bula
Complete article of bula
Extract from verbal moods table for the Indicative mood
Complete anchor table for verbs
Complete article of lemati
Complete article of lemati
Data box informing the user that English articles a/the are not translated
Complete article for man
Complete article for man
Complete article for will
Complete article for will
Complete article for not
Complete article for not
Complete article for open
Complete article for open
Complete article for door
Complete article for door

Figure 1: A design for Sepedi indicating pop-up information obtained through hovering and clicking.

8 Dictionaries Suitable for Text Production

One of the major shortcomings in African language lexicography is the lack of dictionaries for guidance in "text production" situations. Most dictionaries barely fulfil the needs of the users for decoding purposes or the use of dictionaries "on demand". So, for example, negation is a complicated issue in Bantu languages — many negation strategies, e.g., *ga*, *sa*, *se*, *ga se* and *ka se*, are distinguished by Prinsloo (2020b) for Sepedi. These strategies are complicated and non-interchangeable. Most dictionaries do not even fulfil the most basic receptive needs of users, not to mention a lack of guidance on when which negation morpheme can be used. Examples of efforts towards giving guidance in productive dictionary use for Sepedi include a variety of support tools that can be linked to a dictionary such as an assistant for the compilation of isiZulu possessives (Bosch & Faasz 2014) and a sentence constructor, the *Sepedi Helper* for Sepedi.

9 Conclusion

In this paper, it was attempted to give, within the limits of only a few pages, an overview of aspects of the lexicographic treatment of salient features in paper and electronic dictionaries focusing on African language lexicography. This was done in the context of the many challenges faced by the lexicographer in the compilation of dictionaries for African languages.

10 References

Atkins, B.T.S. (1996). Bilingual Dictionaries: Past, Present and Future. In M. Gellerstam, J. Järborg, M. Sven-Göran et al. (eds.) *Euralex '96 Proceedings: Papers Submitted to the Seventh Euralex International Congress on Lexicography*, Göteborg University, pp. 515-546. Göteborg, Sweden.

Barnhart, C.L. (1962). Problems in Editing Commercial Monolingual Dictionaries. In F.W. Householder, S. Saporta (eds.) *Problems in Lexicography,* pp. 161-181. University of Indiana, Bloomington.

Bosch, S.E., Faasz, G. (2014). Towards an Integrated E-Dictionary Application — The Case of an English to Zulu Dictionary of Possessives. In A. Abel, C. Vettori, N. Ralli (eds.) *Proceedings of the 16th Euralex International Congress: The User in Focus 15-19th July 2014*, pp. 739-747. Bolzano, Italy.

CCURL. (2014). Proceedings overview: *Workshop on collaboration and computing for under-resourced languages in the Linked Open Data Era*. Accessed at: http://www.ilc.cnr.it/ccurl2014/ [01/06/2016].

De Schryver, G.M. (2010). Revolutionizing African language lexicography — a Zulu case study. In *Lexikos 20*, pp. 161-201.

De Schryver, G.M., Prinsloo, D.J. (2000a). Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. In *South African Journal of African Languages,* 20(4), pp. 290-309.

De Schryver, G.M., Prinsloo, D.J. (2000b). (In)consistencies and the Miraculous Consistency Ratio of '(x 1.25)4 = x 2.44', A perspective on corpus-based versus non-corpus-based lemma-sign lists. In *Fifth International Conference of AFRILEX*, *July 2000*. University of Stellenbosch.

(DS) Snyman, J.W., Shole, J.S. & Le Roux, J.C. (1990). *Dikišinare ya Setswana English Afrikaans Dictionary*. Pretoria: Via Afrika.

Gouws, R.H., Prinsloo, D.J. (2005). *Principles and practice of South African lexicography*. Stellenbosch: African Sun Media.

Gouws, R.H., Tarp, S. (2017). Information Overload and Data Overload in Lexicography. In *International Journal of Lexicography,* 30(4), pp. 389-415.

Haas, M.R. (1962). What belongs in the bilingual dictionary? In F.W. Householder, S. Saporta, (eds.) *Problems in Lexicography*, pp. 45-50. University of Indiana, Bloomington.

Hartmann, R.R.K. (ed.). (1990). *Lexicography in Africa. Progress Reports from the Dictionary Research Centre Workshop at Exeter, 24-26 March 1989: Exeter Linguistic Studies 15*. Exeter: University of Exeter Press.

Jones, K., Cwi, T.F. (2014a). *Ju|'hoan children's picture dictionary. Interactive disc-gallery*. Pietermaritzburg: University of KwaZulu-Natal Press.

Jones, K., Cwi, T.F. (2014b). *Ju|'hoan children's picture dictionary. Information leaflet*. Pietermaritzburg: University of KwaZulu-Natal Press.

(OZSD) De Schryver, G.M. (ed.). (2010). *Oxford Bilingual School Dictionary: Zulu and English / Isi-chazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi*. OUP Southern Africa, Cape Town.

Prinsloo, D.J. (2015). Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited corpus. In *Lexikos,* 25*,* pp. 285-300.

Prinsloo, D.J. (2020a). Detection and lexicographic treatment of salient features in e-dictionaries for African languages. In *International Journal of Lexicography,* 33(3), pp. 269-287.

Prinsloo, D.J. (2020b). Lexicographic treatment of negation in Sepedi Paper dictionaries. In *Lexikos,* 30*,* pp. 321-345.

Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In R.V. Fjeld, J.M. Torjusen, (eds.) *Proceedings of the 15th Euralex International Congress, 7-11 August 2012*, Oslo, pp. 47-92.

Rundell, M. (2015). From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. In *Lexikos,* 25, pp. 301-322.

Sepedi Helper, Accessed at: (http://sepedihelper.co.za) [20/08/2020].

Taljard, E., Prinsloo, D.J. (2019). African Language Dictionaries for Children — A Neglected Genre. In *Lexikos,* 29, pp. 199-223.

Van Wyk, E.B. (1995). Linguistic assumptions and lexicographical traditions in the African languages. In *Lexikos,* 5, pp. 82-96.

Wolvaardt, J. (2017). South Africa's National Lexicography Units: time for a reboot? In *2nd International Conference of the African Association for Lexicography (Afrilex)*, pp. 9-10. Rhodes University, Grahamstown.

Acknowledgement

This research is supported in part by the South African Centre for Digital Language Resources (SADiLaR). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed are those of the authors.

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Neology

It's a long way to a dictionary: Towards a corpus-based dictionary of neologisms

Afentoulidou V., Christofidou A.

Academy of Athens, Greece, University of Athens, Greece
vafentoul@phil.uoa.gr, christo@academyofathens.gr

Abstract

In this paper we discuss three main different views on the documentation of neologisms supporting the construction of a corpus-based lexicon of neologisms (as a language resource), which will include those new lexical units that enter the *consolidation stage* (according to certain criteria) before their entering into the *establishment stage* (Kerremans 2015). The documentation (collection and monitoring) of those new lexical units will be both linguistically and lexicographically helpful: a. it provides the linguist with a valuable linguistic information tank (morphology, semantics, morphology-text interface etc.) and b. it facilitates the answer to the desideratum of the dictionary inclusion (or not) of neologisms. We focus on the second issue and show that corpus exploration methods and measurements such as peakedness of distributions and lexical dispersion can be operationalized as tangible criteria to conjointly evaluate the frequency profiles of new formations, and that peakedness is a promising indicator of "lexical sustainability". Drawing examples from a 160-million-word sub-corpus of the *Monitor Corpus of Neologisms* compiled for NEOΔHMIA research project at the Academy of Athens, comprising newspaper discourse spanning 5.4 years, we track the frequency development of selected new formations which emerged during the Greek debt crisis and discuss their evolution in time.

Keywords: neology; dictionary inclusion; corpora; consolidation; peakedness; dispersion

1 Introduction

Living in a period overwhelmed by the pandemic vocabulary, the first question which comes to mind would be: how many and which of these new formations will remain? The old question for linguists and lexicographers arises again: is it possible to establish specific indicators of the evolution, the survival, or the life-cycle of new words? In an attempt to give some answers to this question we chose to look back exploring the behaviour and the evolution of the already fading away neological vocabulary of the Greek debt crisis as witnessed by linguistic evidence in the corpus component of NEOΔHMIA (see section 2) and specifically by the data of one Greek newspaper within the timespan 2015-2020.

In the following (section 2), we highlight the main objectives and methodological commitments adopted for the purposes of the Greek Neology project NEOΔHMIA conducted at the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* concerning Modern Greek Neology. In section 3 we present related work on the topic. In section 4 we discuss the different views on the dictionary inclusion and/or documentation of neologisms. In section 5 a corpus-based analysis is conducted (methodology and results) and a discussion of the corpus findings follows in section 6. The paper ends with concluding remarks and a research outlook.

2 NEOΔHMIA at the Research Centre for Scientific Terms and Neologisms

NEOΔHMIA is an ongoing research programme conducted at the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* (2008-), constantly developed to accomplish the tasks of (semi)automated detection and linguistic analysis of Greek neologisms and terminology, the first of its kind concerning Greek neology (Christofidou et al. 2013). The project concerns the development of an integrated databank, comprising four main modules, constantly updated:

1) Neologism text retrieval component: A custom-made crawler is browsing the online versions of selected Greek newspapers (with the largest circulation). Although accurate text extraction and content representation are adapted to the technical and representational demands of newspaper feeds/webpages, the system is flexible enough to trace other kinds of online sources (i.e., non-press). The texts produced from the crawling module are cleaned and pre-processed and the final output is represented in XML and automatically enriched with metadata following the recommendations of the Text Encoding Initiative (TEI P5 Guidelines@tei-c.org). A document model has been defined as a custom-made TEI schema (Afentoulidou & Christofidou 2017) comprising different annotation layers for the representation of basic metadata and text profiling (genre and topic classification), the structural superordinate divisions of newspaper text (document structure), as well as basic grammatical analysis (POS tagging and lemmatization) and is expandable.

2) Neologism extraction component: A dedicated tool uses the output of the text retrieval component (or accepts any kind of XML file conforming to the TEI-schema of the project) and performs automated detection of "new words" (candidate neologisms) through computational techniques, which make use of the well-documented method of exclusion word lists, as well as named entities stoplists (see Christofidou et al. 2013; Afentoulidou & Christofidou 2017 for details on the processing steps, elimination of noise and methods of updating the exclusion procedure). The system identifies one-word

units, although candidate multiword formations are proposed and submitted to human inspection only for n-word grams connected with dashes (Christofidou et al. 2020). Lists of candidate neologisms are produced for manual evaluation. The Neologism extraction component is complemented by a manual selection procedure employing lexicographic criteria, such as (non)occurrence of the candidate neologisms in the reference dictionaries of Modern Greek, as well as the generation of web impact reports (Thelwall 2018) via software[1] or linguistically relevant search engine queries and measurements, Google or Bing-based (*Web as Corpus* methodology, Lüdeling, Evert & Baroni 2007; Christofidou et al. 2013).

3) Neologism classification component: Together with lexical use, morphological (mainly word-formation analysis) and textual information (genre, text type, topic, text structure etc.) is recorded in the Centre's database of neologisms. A system of automatic topic classification of newspaper articles is being developed with the aid of supervised and non-supervised machine learning techniques (see, among others, Hagen 2012) to facilitate contextual analyses of neologisms on a larger scale.

4) Neologism monitoring component: Any further quantitative or qualitative observations regarding the use of the words utterly classified as neologisms make use of the specifically designed corpus of online newspaper discourse mentioned above; the corpus is thus used both for neologism retrieval and monitoring (*Web for Corpus* methodology, see De Schryver 2002). The *Monitor Corpus of Neologisms* (Afentoulidou & Christofidou 2017) nowadays includes more than 400 million words of journalistic discourse and is compiled following international standards (Text Encoding Initiative) to support empirically testable, textually-informed analyses of the morphological tendencies of Modern Greek. Moreover, webometric data (Christofidou et al. 2013; Christofidou, Karasimos & Afentoulidou 2014) are supplied for every neologism for the date of its first recording in the database and on later intervals (on-demand so far, although an annual webometric monitoring is envisaged for all neologisms currently in the database).

The four components offer a dynamic (Renouf 2016; Cartier 2019), unified pipeline of research (although not fully automated yet) and define the Centre's digital infrastructure NEOΔHMIA for tracking and classifying neological formations in Modern Greek.

3 Related Work

As far as research on Greek neologisms/neology is concerned, NEOΔHMIA is active and constantly updated.[2] We should also mention the prominent research of Professor A. Anastassiadis-Symeonidis, which follows a lexicographic, more qualitatively-oriented database approach (Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009).[3]

The study of lexical innovation with the aid of computer technologies is central to numerous European initiatives and related projects concerning neology. Dedicated research environments based on corpus evidence have been developed for various languages. As a common denominator, they share a dynamic, quantitative orientation and a holistic concern for the balanced development of tools and procedures, not only for the challenging and fundamental task of semi-automatic discovery and linguistic classification of new words and/or new meanings (formal and semantic neology), but also for their monitoring across time, space and contexts. For instance, the French platforms *Néoveille* (Cartier 2016) and *Logoscope* (Gérard, Falk & Bernhard 2014), as well as *Néonaute* (a recent extension of *Néoveille* and *Logoscope* in collaboration with the BnF, see Aubry, Cartier & Stirling 2018); the web interfaces of the worldwide neology networks for Catalan and Spanish, coordinated by the Observatori de Neologia – *OBNEO* at the University Pompeu Fabra in Barcelona[4] (*Antenas Neológicas, NEOROM, NEOROC, NEOXOC*); the German web service *Die Wortwarte*[5] under the umbrella of the Berlin-Brandenburgische Akademie der Wissenschaften and the *Neologismenwörterbuch online*[6] at the Institute for the German Language in Mannheim; the *Neocrawler* (Kerremans et al. 2018) and the English Neologisms Research Group at Ludwig-Maximilians-Universität München; the system for neology extraction and monitoring based on the *Norwegian Newspaper Corpus* (Andersen & Hofland 2012); the special dictionary of neologisms *Neologismenwoordenboek*[7] and the neology section[8] of the online, corpus-based *Algemeen Nederlands Woordenboek* of the Instituut voor de Nederlandse Taal; the pioneering service *APRiL* (Renouf 2007a; Renouf 2013) at the RDUES Birmingham City University; all of them forming, to the best of our knowledge, a representative albeit not exhaustive list of neologism trackers, computational tools and infrastructures with solid online presence and/or dissemination of research outcomes in dedicated bulletins and printed series (see Christofidou et al. 2013; Afentoulidou & Christofidou 2017 for supplementary overviews). Irrespective of possible specific design requirements, this line of research necessarily adopts a database (SQL, no-SQL) approach to data management and requires the manual intervention of the expert-linguist to evaluate and enrich the data collected and classified by the machines. Eventually, such systems as products of technology extend their scope beyond the very study of neology and embrace further empirical and applied objectives both in lexicographic and corpus research. More specifically:

(a) They can natively support dictionary compilation (despite the differences in the degree of lexicographic orientation)

[1] Webometric Analyst Web Impact Reports (http://lexiurl.wlv.ac.uk/index.html).

[2] The *Néoveille* platform (Cartier 2016) has the potential to track and monitor Greek neologisms; the public and the guest interface, however, seem currently not updated for Greek.

[3] The *electronic database of Modern Greek Neologisms* is characterized by its creators as a "lexicographic product" (Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009: 419) and was also used to enrich the macrostructure of the *Reverse Index of Modern Greek Vocabulary* (https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/reverse/index.html).

[4] https://www.upf.edu/web/obneo

[5] https://wortwarte.de

[6] https://www.owid.de/docs/neo/start.jsp

[7] http://neologismen.ivdnt.org/search

[8] http://anw.inl.nl/neologismen

by constituting specialized lexicographic resources themselves[9] or by affiliation or contribution to larger dictionary projects and corpus lexicography.[10] In any case, every research infrastructure implements its own inclusion and exclusion criteria in filtering neologisms for final presentation in a dictionary or collection of neologisms.[11]

(b) Such infrastructures utterly make a strong impact in computational corpus-linguistic research, because of their reliance on monitor corpora or at least very large textual sources to automatically extract, classify and systematically monitor neologisms. The domain of the Press (online versions of newspapers or news feeds) has traditionally provided the starting point for neologism tracking for most of the projects[12] for technological and linguistic reasons: RSS protocols permit easier harvesting of data for corpus compilation and journalistic discourse has better chances of representing institutionalized usage, thus the written standard. Despite the limitations in genre representativeness and balance, nowadays, the multifaced character of online journalism encompasses much more genres and topics than in the past and extends to a wider range of documents (supplements, magazines) supporting the web edition of a newspaper, besides the prototypical news categories. At the same time, we can observe a growing tendency to document neologistic use beyond morphological and word-formation taxonomies, with an emphasis on the otherwise neglected (con)textual variables (genres, topics) and the development of relevant annotation schemes, so valuable in corpus-linguistic research. Finally, a reconciliation of the *web as/for corpus* methodologies is witnessed in newest platforms, such as the *Néonaute*, in line with NEOΔHMIA's (see section 2) early commitment in doing both (Afentoulidou & Christofidou 2017), since there are inherent epistemological disadvantages in assessing linguistic usage both through search engine results and big data architectures. In any case, all projects share a mutual concern in developing more intelligent methods for automated neologism detection – for instance by machine learning – as well as corpus resources with richer annotation layers and open science data policies. In the ever-growing digital landscape of linguistic resources and services, the focus in documenting language has now shifted from the lexicographic exclusion principles to inclusion possibilities or rather "prioritization policies" (Connor Martin 2019) and data analytic frameworks,[13] with a parallel concern to eliminate true data noise and balance recall and precision.

NEOΔHMIA, in a parallel line of research (see section 2) embraces all those challenges and methodological commitments in data collection, has a strong orientation towards morphology and textlinguistics with a view to incorporate semantic/pragmatic approaches and computationally unify, in a single digital ecosystem, research on neology with terminological research.[14]

4 Approaches to the Dictionary Inclusion of Neologisms

There are numerous descriptions and definitions of neologisms and/or neology. According to more strict approaches, a neologism is defined as any lexical unit being classified as a member of the active vocabulary of the speech community (institutionalized or lexicalized lexical units, see below, cf. Hohenhaus 2005: 359-365), based on certain criteria, which would allow its inclusion in general-purpose dictionaries (for criteria see among others Teubert 1998: 131ff.; Herberg, Kinne & Steffens 2004: XII; cf. Klosa-Kückelhaus & Wolfer 2019). According to more open approaches a neologism is defined as any new lexical unit occurring in the oral or written discourse of a certain period (see definition in Anastassiadis-Symeonidis, Alexiadou & Nikolaou 2009: 420, cf. Nikolaou & Anastassiadis-Symeonidis 2017: 272). Following an intermediate approach, under the notion of neologism we understand any new lexical unit which exhibits a recent repeatability of use[15] within the speech community, even if it does not meet (yet) the condition of establishment and institutionalization (see below; also, Renouf 2007b). Thus, a decisive factor for the treatment of neology seems to depend on the research objective: a research for the sake of lexicography should follow the first more strict approach while a broader, linguistically driven research should rather follow the other two approaches.

From the three different views above (cf. also the presentation of related work in section 3) it becomes obvious that three respective approaches could be discerned concerning the objectives of neologisms' documentation (see also Guerra 2016):

a. Inclusion of only these new formations which "deserve" it, i.e., they meet certain quantitative and qualitative conditions which ensure that the words are institutionalized or lexicalized[16] (with an indication for the neologicity of the entry), see Klosa-Kückelhaus & Wolfer (2019); Freixa & Torner (2020); Cabré & Nazar (2012); Christofidou (2015); Connor Martin (2019)

[9] For example, the *Neologismenwörterbuch* available at the IDS portal OWID (https://www.owid.de/) and the online dictionary of neologisms of different varieties of Spanish *El Antenario* (https://antenario.wordpress.com/presentacion/).

[10] For example, the so far connection of the late version of *Die Wortwarte* (Lemnitzer 2010) to the enrichment of the *Digitales Wörterbuch der Deutschen Sprache* in the DWDS portal (https://www.dwds.de/), of the *Norwegian Newspaper Corpus* (http://avis.uib.no/) to specific dictionary projects of Norwegian (Andersen 2013), of *Neocrawler* to the Oxford English Dictionary Team, of the *Neologismen Database* to the compilation of the *Algemeen Nederlands Woordenboek* (see section 3).

[11] Cf. the maximalistic all-inclusive *in natu* recording policy of *Die Wortwarte vs.* the conservative, with multiple exclusion criteria, filtering of neologisms to-be-included in the *Neologismenwörterbuch* macrostructure (Klosa & Lüngen 2018).

[12] Cf. the *Neocrawler* system with a different, *web as corpus* approach and *OBNEO's* extended methodology of scanning texts also from magazines and spoken sources to trace neologisms.

[13] A converging trend can be witnessed from the part of lexicographic projects, such as the *OED Oxford Labs Initiative*, with the aim to gain richer insights into language change through the lens of large-scale analysis of the *OED* dataset itself.

[14] The Centre's terminological resources make use of thesaurus building systems and classification schemes based on ontologies.

[15] The criteria to establish repeatability differ according to the specific research objectives.

[16] On institutionalization *vs.* lexicalization and their role to neology there are subtle differentiations among the researchers (see Hohenhaus 2005; Klosa-Kückelhaus & Wolfer 2019; Kerremans 2015: 41).

b. Documentation of any attested new coinage (even nonce formations under certain conditions) in electronic (static/dynamic) databases (among others the German neologisms project *Die Wortwarte*)
c. Documentation of new lexical units (even ephemeral ones) which meet a minimum of – mostly – quantitative conditions for their repeatability in a dynamic (electronic) lexicon of neologisms (and further monitoring)[17]

In the following, we will further discuss the third approach in favour of which we would like to argue: Kerremans (2015: 40, cf. Schmid 2008) proposes a model of the establishment process of a neologism, i.e., she defines three *stages* (creation, consolidation, establishment) which a new formation would undergo within three different *perspectives*: lexicalization (structural perspective), institutionalization / conventionalization (socio-pragmatic perspective) and hypostatization / entrenchment (psycholinguistic perspective). Following partially Kerremans' model (2015) we propose that a new lexical unit should be included in a dictionary (see approach a. above), only if it covers all three *stages* (i.e. creation, consolidation and establishment) from at least the first two *perspectives*, that of lexicalization and institutionalization / conventionalization.[18] As far as linguistic research is concerned, we assume that a new formation should be captured and registered in an electronic dictionary or a dynamic base of neologisms already at the beginning of the *stage of consolidation* (i.e., stabilization of form and meaning) within the first *perspective* of lexicalization (see approach c. above). Moreover, nonce formations (or hapax legomena) are ad hoc formations, which still remain at the first *stage of creation*. To our mind, nonce formations – though linguistically very important – should be treated separately, since ad hoc formations behave dramatically different than formations reaching the *stage of consolidation* (in all three *perspectives*). They often consist in multiword expressions, blends, surface analogy[19] or poetic formations and they are the only formations that can be ungrammatical and exhibit an ad hoc (exclusively context-dependent) meaning (see also Renouf 2007b: 8ff.).

The above proposal seems to be based on two different research views: For a lexicographer it is more important to pursue registering only these new formations which meet the level (*perspective*) of institutionalization / conventionalization, covering all three *stages* from creation to establishment (see above). For a linguist, who investigates the phenomenon of neology (morphology, semantics, language change etc.) it should be crucial to record all new words which show at least a kind of stabilization of form and meaning (*stage of consolidation*). Many new formations, although ephemeral, i.e. not ad hoc but not (yet) dictionarizable, reveal at least the same amount of linguistic information as the successful, thus dictionarizable neologisms, whilst their life-cycle is rather pragmatically and socio-linguistically conditioned (for discussion see Kerremans 2015: 41-43 and Fischer 1998: 178ff.). Thus, they equally provide the linguist with important information on morphological trends (within word families), on semantic evolution, on the text-morphology interface and partially on language change. In this sense the ephemeral, but not (yet) dictionarizable, neologisms constitute a tank of linguistic information, which should be documented for multiple investigation and further monitoring.

Nevertheless, one of the most important contributions of such an approach to the phenomenon of neology per se would be the following: the systematic monitoring, tracking and analysis of the evolution and life-cycle of the majority of new formations, within a specific period in the recent past – regardless of their possible disappearance or success – could provide us with possible estimations for the behaviour of new words in the future, and consequently with suggestions for their dictionary inclusion or not.

In the following sections there will be an attempt, based on data from the corpus component of ΝΕΟΔΗΜΙΑ project (section 2), to show the contribution of the tracking of new lexical units' evolution, according to specific measurements, to the identification of neologisms, either a. for inclusion in a broader dynamic electronic lexicon of neologisms and/or b. for inclusion in a general-purpose dictionary.

5 Corpus-based Development

5.1 Method

5.1.1 Data

In order to monitor the behaviour and the evolution of new lexical units for the purposes described above we decided to track all new formations recorded in the database of ΝΕΟΔΗΜΙΑ, belonging to the Greek debt crisis (2010-2019) vocabulary. Due to our heuristic procedure the only limitation has been a minimum of 100 occurrences in the search engine Google, in order to ensure repeatability and a form-meaning stabilization (checking the context of use for the web occurrences). In addition, we collected all neological lemmas from *The Vocabulary of Crisis* (Varoufakis 2011). This procedure provided us with a list of 32 new lexical units (derivatives, one- and two-word[20] compounds) concerning the Greek debt crisis.

To study the spread and life-cycle of the 32 Greek debt crisis formations we used a sub-part of the *Monitor Corpus of Neologisms* (see section 2). Instead of applying random-sampling techniques to the entirety of newspaper sources crawled, we selected for the purposes of this study the newspaper feed which produced the largest amount of data per year (*Proto Thema*) and made our searches within all collected written content. Since our focus is not on capturing the overall diachronic trend and the fate of those words in Greek society in general (and between newspapers), but our aim is to

[17] Concerning both b. and c.: cf. the Research Programme of the Aristotle University of Thessaloniki, conducted by Prof. Anastassiadis-Symeonidis, presented in Anastassiadis-Symeonidis, Alexiadou & Nikolaou (2009).
[18] The third *perspective* of entrenchment concerns a level of a psycholinguistic approach addressing mostly the perception level.
[19] Such word formation processes also apply to (not ad hoc) neologisms, albeit much rarer.
[20] See Christofidou et al. (2020) on qualitative and quantitative criteria for compoundhood of (new) multiword expressions.

"freeze time" somehow, zoom into their frequency profiles and study how they developed during a specific time span, by prioritizing continuous coverage[21] to source variation, we maximized our chances of providing a representative picture of those novel formations' unique trajectories, of course with a limitation of our observations to the specific newspaper.[22] Moreover, in that way we avoided the thorny issue of having to disentangle from our results topic-related newspaper bias and newspaper-specific coverage, which unavoidably characterises mixed corpora, i.e. of many newspaper sources, and is discussed by Gabrielatos et al. (2012: 162-164) in their study on the peaks and troughs of corpus-based contextual analysis in the UK Press. They witness great differences between the newspapers they study, to the degree of questioning "the utility of examining the development in the number of articles in the corpus as a whole – thus effectively treating British national newspapers as a homogeneous group" (p. 162) and conclude:

> In light of the above, studies of groups of newspapers, taken as a whole, may miss important individual differences. Conversely, studies of individual newspapers can safely generalise only about the particular newspaper. Therefore, if the corpus comprises distinct sub-corpora (in our case, different newspapers), then frequency developments should be examined in those individually as well as in the corpus as a whole. (Gabrielatos et al. 2012: 163-164)

For the purposes of this study, we divided the corpus into monthly sub-corpora, in order to monitor frequency developments over time to a higher level of granularity. As a *terminus post quem* we decided on September 2015, when Greek national elections took place amidst the economic debt crisis, which was then profoundly consolidated in all aspects of life in the country, following the Greek Bailout Referendum of the summer of 2015, when the bailout conditions of the European Union, the IMF and the European Central Bank were rejected. As a *terminus ante quem* we opted for the end of 2020, a year that the spread of the Covid-19 pandemic crisis took over, still reigns supreme – and in any case overshadowed the Greek debt crisis, which has been at the time already softened (the final bailout came to a formal end in the end of 2019). Therefore, a 5.4-year perspective was adopted with a total of 64 successive sampling points (64 months i.e. 64 corpus sections/large XML files) in order to gain a wider scope from the seamless comparison of frequency patterns across time. So the corpus used in this study comprises 552,975 press articles, of various lengths, from the online version of the newspaper *Proto Thema* covering the period from the 1st of September 2015 to the 31st of December 2020, of the total size of about 160 million running words (tokens).

5.1.2 Procedure

Specific (and time-consuming) pre-processing steps were performed semi-automatically to prepare the 160 million tokens corpus for analysis (for instance, article deduplication and removal of repeated content noise, whitespace and non-whitespace character normalization, cleaning of residual CDATA or stripping non-parsable XML entities, conversion to a custom-TEI schema), since the articles were collected (except for the last four months of 2020) using a previous, less automated version of NEOΔHMIA's crawler. Then 32 queries were compiled for each lexical unit, covering all inflectional or spelling variants (a total of 1,055 case-insensitive searches were written, using simple regular expression notation) to be performed with two software packages (Voyant Tools and WordSmith Tools).

The corpus was imported to the Voyant text analysis environment (server edition) using the XML teiCorpus ingest module integrated in Voyant Tools (VTs), with the tokenization parameter set on "Simple Word Boundaries". WordSmith Tools 8.0 (WSTs) produced concordances and enhanced the frequency profiling of the selected neologisms by the computation of dispersion metrics.

Due to time limitations we did not perform manual lemmatization for the lexical units under examination (there were 14,502 occurrences of the search terms in total – see Table 1) and, instead, made use of VTs' search syntax to match items separated with pipes as a single term. Where needed, some spelling variants with hyphens or parentheses, such as *meta-mnimoniakos*, *(meta)mnimoniakos*, *neo-troikanos* were normalized using the TEI element <*reg*> to eliminate noise in the recall of single terms (*mnimoniakos, troikanos*) and make sure that all instances of the 32 formations were correctly retrieved. Finally, the 6 multiword units were annotated and enclosed within the element <*mwu*>, in order to be processed as single items with the WSTs' WordList procedure. For every lexical unit queried with VTs we used the Trends and Terms tools (with the *Relative Frequency* per million, *Peakedness* and *Skew* columns activated besides the default ones – *Count* a.k.a. absolute frequency and *Trends*). The degree of neologism consolidation within a community of discourse receivers (newspaper readers) and producers (journalists, audience commentators) should be captured with the computation of frequency profiles throughout the corpus, as well as the use of time-lined dispersion statistics.

5.2 Corpus Exploration and Analysis

Table 1 presents the quantitative results of the procedure discussed in the previous section. From left to right, *Lemmas* correspond to cumulative searches for each selected new formation. The *Lexical Frequency* values (*Absolute and Relative*) range from one occurrence (two hapaxes*, chreofreno*, *dimokratoria*) to 3,554 hits (22.5 words per million for *mnimoniakos*) and are rather low if we take the size of the corpus into consideration. The *Peakedness* statistic measures

[21] Due to technical reasons that interrupted data collection, 75 days are missing, but the gaps are spread across 1,874 days in total.

[22] According to Alexa's site rankings for Greece (https://www.alexa.com/topsites/countries/GR), the online version of *Proto Thema* (https://www.protothema.gr) was, in 2015, and still is (as of March 2021) first among all other Greek online newspapers ranked by the specific service for overall traffic calculated by the combination of daily visitors and pageviews (thus indirectly measuring degree of readability). Moreover, the newspaper addresses a wider audience, publishes on a diverse range of topics besides the central, typical news categories (great emphasis is given on popular topics such as lifestyle and celebrities, psychology, entertainment, cooking etc.) and produces an overall multi-genre inventory of resources (besides multimedia content, there is a constantly updated blog section with point of view articles, advertorial columns, verticals, recipes, connection with magazines etc.).

how much the relative frequencies of the lemmas are bunched up into peaks, whereas a peak is defined as a region with high values, where the rest are lower.[23] The peaks are formed when there are extreme differences between documents (i.e. corpus periods) and represent significant outliers, that is discourse fluctuations and instability due to topicality. Large spikes denote uneven patterns of sudden rises/falls in usage. In essence, the values in Table 1 provide an overall kurtosis estimation.[24] *Skew* is a statistical measure of the symmetry (skewness) of the relative frequencies. A positive skew is formed when the mass of the distribution is concentrated on the left and the right tail is longer, suggesting that the overall frequency profiles follow a declining path, irrespective of periods of regression. A skew value approaching zero corresponds to usage consistency, whilst negative values would mean that frequencies started low and increased. The *Peakedness* / *Skew* metrics, therefore, holistically evaluate the *shape* of the frequency curve, thus usage intensity.

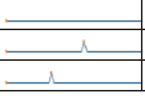
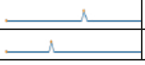
Lemma	Absolute Frequency	Relative Frequency (pmw)	Peakedness	Skew	Sparkline	Dispersion	Kendall's τ coef/ent	Trend [**/* Correlation is significant at the 0.01/0.05 level (2-tailed)]
ftochopiisi	558	3.534884	-0.04406605	0.6568374		0.83	-0.481	Slow decline** ↘
metamnimoniakos	1324	8.387431	0.92924464	1.4513253		0.59	0.209	Slight increase* ↗
pragmatiki_ikonomia	1199	7.595566	1.0510801	1.0915446		0.92	-0.255	Slow decline** ↘
eyroieratio	121	0.766525	2.196963	1.7333602		0.67	-0.434	Slow decline** ↘
ftochopio	286	1.811786	2.415639	1.3975376		0.75	-0.201	Slow decline* ↘
troikanos	114	0.722181	3.4694684	1.809957		0.73	-0.338	Slow decline** ↘
merkelistis	30	0.190048	3.81772	1.9048449		0.79	-0.135	Stable usage (decreasing) →
esoteriki_ypotimisi	90	0.570143	7.9706416	2.4776852		0.67	-0.384	Slow decline** ↘
titlopiisi	422	2.673335	8.054505	2.2520626		0.7	0.503	Moderate increase** ↗
mnimoniakos	3554	22.5143	8.213318	1.9867238		0.78	-0.581	Moderate decline** ↘
ypertamio	1570	9.945821	9.340388	2.8904257		0.77	-0.246	Slow decline** ↘
domimeno_omologo	55	0.34842	9.52648	2.9386773		0.65	-0.316	Slow decline** ↘
trapezokratia	13	0.082354	9.68198	3.1034594		0.58	-0.018	Stable usage (decreasing) →
posotiki_chalarosi	1696	10.74402	9.995694	2.9633954		0.61	-0.352	Slow decline** ↘
ithikos_kindynos	44	0.278736	10.819942	2.7555737		0.71	0.060	Stable usage (increasing) →
eyrofovikos	82	0.519463	11.275036	3.0954626		0.67	-0.119	Stable usage (decreasing) →
apikia_chreoys	43	0.272401	16.617157	3.6415455		0.65	-0.268	Slow decline** ↘
chreokratia	5	0.031675	19.061302	4.3090296		0.47	-0.227	Slow decline* ↘
anakefaleopiisi	1952	12.36576	20.783533	4.4174266		0.78	-0.706	Sharp decline** ↓
ypermnimonio	3	0.019005	24.487204	4.9032397		0	-0.232	Slow decline* ↘
menoymeyropeos	31	0.196382	25.19234	4.7880397		0.67	-0.028	Stable usage (decreasing) →
stasimochreokopia	50	0.316746	25.698536	4.445233		0.7	-0.249	Slow decline** ↘
eyrokratia	2	0.01267	29.37236	5.518538		0.35	0.193	Stable usage (increasing) →
eyroarnitismos	2	0.01267	29.39645	5.5200977		0	-0.201	Slow decline ↘
antimnimonio	102	0.646162	30.239964	4.8551636		0.69	-0.259	Slow decline** ↘
eyroomologo	389	2.464283	37.208927	5.864297		0.78	-0.101	Stable usage (decreasing) →
antimnimoniakos	752	4.763858	40.01884	5.821955		0.73	-0.589	Moderate decline** ↘
merkelismos	3	0.019005	42.524155	6.3934593		0	-0.098	Stable usage (decreasing) →
antimerkelistis	6	0.03801	62.433693	7.8648567		0.17	-0.090	Stable usage (decreasing) →
chreofreno	1	0.006335	64	8		0	0.177	Stable usage (increasing) →
dimokratoria	1	0.006335	64	8		0	-0.003	Stable usage (decreasing) →
germanopio	2	0.01267	64	8		0	-0.076	Stable usage (decreasing) →

Table 1: Frequency distribution / development of the search terms in the corpus during the 64-month period (sorted on *Peakedness*).

Sparkline graphs are generated for each query, namely the concordance hits are visually represented as lines that show the distribution of relative frequencies across the chronologically ordered corpus documents, followed by the *Dispersion* statistic, which is the Juilland's *D* implementation of WSTs and is computed with the WordList tool. Due to the absence of lemmatization, for every lemma, only the highest dispersion value amongst all variants is displayed in Table 1 to represent the degree to which frequencies are evenly spread (maximum value=1, suggesting uniform dispersion | minimum value=0, suggesting burstiness).[25] The last two columns introduce a further abstraction: the detection of trends in the data, by correlating the observed relative frequencies with the sequence of the different temporal stages. Following Hilpert & Gries (2009: 388-390), Kendall's τ correlation coefficients and their *p*-values are produced for each lemma. Values close to 0 indicate the absence of a trend, values approaching -1/+1 indicate sharp decrease/increase.

[23] See VTs Help. WSTs implement a *Peakiness* sorting function of time-lined concordances to graphically display outliers in frequency development within lemmas, where "Peakiness uses the standard deviation of the proportion of hits to word count in each period of a time-line", but the scores per search word are not displayed for a between-lemmas comparison.

[24] High positive values correspond to leptokurtic distributions (extreme fluctuations) lower to mesokurtic and negative to platykurtic.

[25] Aggregated – thus more accurate – dispersion values were also computed with WSTs (generation of time-lined dispersion plots for every lemma but only for the text files in which the search terms appeared, see WSTs Help). After examination of the results, the overall trend was the same, so we opted for the first method of calculation, i.e. using all the files of the corpus.

As the frequency profiles show, overall, there is no considerable pattern of growth for the selected Greek debt crisis new terms at the end of 2020 (especially after August 2019, on close inspection of all the time-graphs), as compared to the beginning of the period of observation (last quarter of 2015). All distributions are positively skewed (with varying degrees of skewness), which means that in principle the words are already in use in 2015, thus seem to undergo a stabilization process (notably *anakefaleopiisi, mnimoniakos, antimnimoniakos, pragmatiki ikonomia, ftochopiisi, troikanos*) but then follow a declining path, most of them fade away with sudden rises and regressions. Some neologisms are visibly attested at the end of the period (*eyroomologo, titlopiisi, chreofreno*) displaying the prototypical exponential curve of neologism diffusion (Cabré & Nazar 2012) but their frequency development, within the window of our observation does not seem to stabilize on a steady trajectory.

Furthermore, when lemmas are sorted on *Peakedness* (see Table 1), we can observe the following pattern: the least peaky neologisms, irrespective of overall frequency, dispersion and direction of change (upwards, downwards) are those with the fuller and more "resilient" life-cycles, whether presently still evolving (thus with the best chance of "survival", such as *ftochopiisi, pragmatiki ikonomia*) or at a time solidly evolved (such as *metamnimoniakos*).[26] On the contrary, as we move up the scale, neologisms with higher *Peakedness* values, thus greater fluctuations and temporal instability, irrespective of overall frequency, dispersion and direction of change (upwards, downwards) are the most transient and their use is purely topical. Thus, in developing practical heuristics regarding the most important quantitative features of successful neologisms undergoing lexicalization (stabilization of form and meaning), the dynamic notion of *Peakedness,* as a predictor that affects "lexical sustainability", seems promising to explore.

Figure 1 plots the Kendall's τ correlation coefficient values of Table 1 in the horizontal axis and the relative frequencies for every lemma a. at the beginning and b. at the end of the period of observation in the vertical axis (see Grieve, Nini & Guo 2016 for a similar methodology to detect emerging word forms in English).

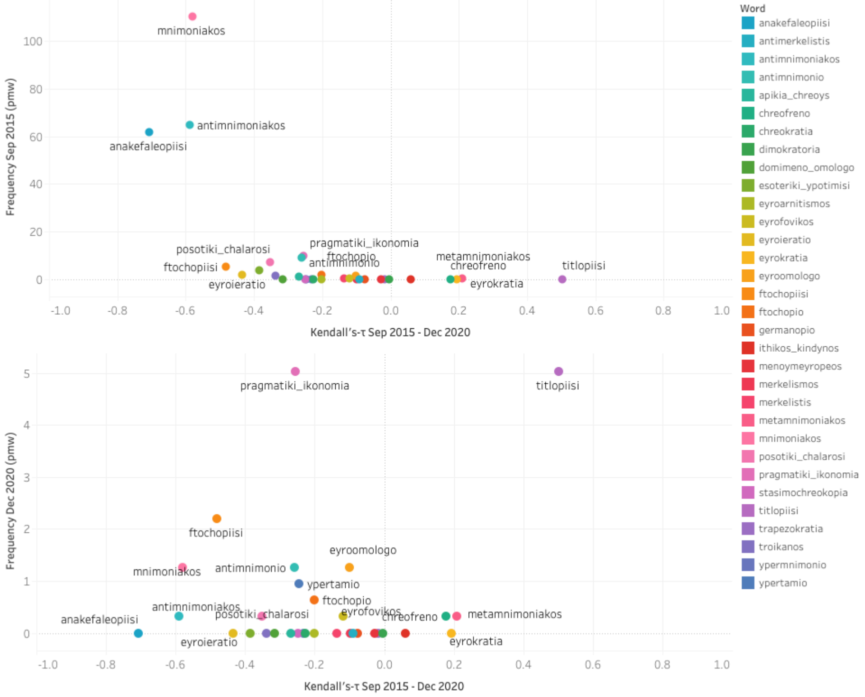


Figure 1: Kendall's τ coefficient *vs.* September 2015 and December 2020 relative frequencies.

The vast majority of the selected formations share a declining trend (see also Table 1). In September 2015, *mnimoniakos, antimnimoniakos, anakefalaiopiisi* were frequently used and kept spreading, but in December 2020 a statistically significant decline in their usage is observed, implying that they did not eventually stabilize in newspaper popular discussions. It is only the least peaky neologisms, *pragmatiki ikonomia* and *ftochopiisi* that reign supreme, remain stable and can be safely considered best candidates "to have marched the long way" towards an *establishment stage*. On the other hand, a few Kendall's τ values are positive, displaying a mild, almost stable upward pattern, but with high *Peakedness* and low *Dispersion* scores (see Table 1). Only *titlopiisi*, as we observe in Figure 1, at the end of 2020 is characterized by rapid growth, namely the formation followed a clear emerging – under way of stabilization – trajectory, as it is also suggested by its middle-range *Peakedness* score. Conversely, *metamnimoniakos* seems from Figure 1 alone, to be "frozen" into an ever-emerging state. Its *Peakedness* values, however, predict otherwise. In fact, *metamnimoniakos* displays the second less peaky frequency development across the period under observation and the mere shape of its

[26] *Eyroierateio,* seems to be a successful, consistent, albeit newspaper-specific preference, since *I Kathimerini,* the second largest newspaper in the corpus disfavors the use of this emotionally-vivid formation. Similar searches were performed for the rest of the sources collected during the specific period (*Ta Nea, To Vima, Ethnos, I Avgi*) and confirm that analysis. A cumulative frequency from both newspapers would then distort this stylistic preference of *Proto Thema*, so we can reliably categorize it as an outlier.

frequency distribution reveals that its usage has not only grown consistently *over a narrower time frame* but also formed a rather stable plateau from, roughly, the summer of 2017 until the summer of 2019, with small regressions until the end of February 2020. At the same time, the *Dispersion* values are rather low, showing uneven distribution across the corpus parts. *Metamnimoniakos* thus seems to have completed a full life-cycle with a shorter life-span than the one we set *a priori* in this study. Had we considered a shorter time-frame of examination, low *Peakedness* would suggest stability and consistency. In other words, a *Peakedness* estimation can also be used *retroactively* to predict subsequent "lexical sustainability" and we argue that the measurement reflects strong on-demand communicative needs (see Discussion).
The candlestick graph in Figure 2 mirrors the lower part of Figure 1 and summarizes frequency development, following Brezina's (2018) adaptation of this type of data visualization used in financial reports to corpus linguistics. The boxes visually represent the y axis of Figure 1 (initial point, September 2015 *vs.* final point, December 2020) and the colour shows when the frequencies were higher (at the beginning – red box – or at the end – blue box). The spikes represent the minimum and the maximum frequencies throughout the 5.4 years. The longer the body of a candle is, the greater is the variation in frequency profiles. The spikes denote frequency fluctuations (when projected outside the box) and, the longer they are in relation to the box and themselves (upper and lower wicks), the less smooth the transitions are. The candles for *pragmatiki ikonomia* and *ftochopiisi* are almost symmetrical if we compare them with the rest.[27] For *metamnimoniakos,* the beginning of its path is also the end and the positive spike in between fits in its whole life-cycle.

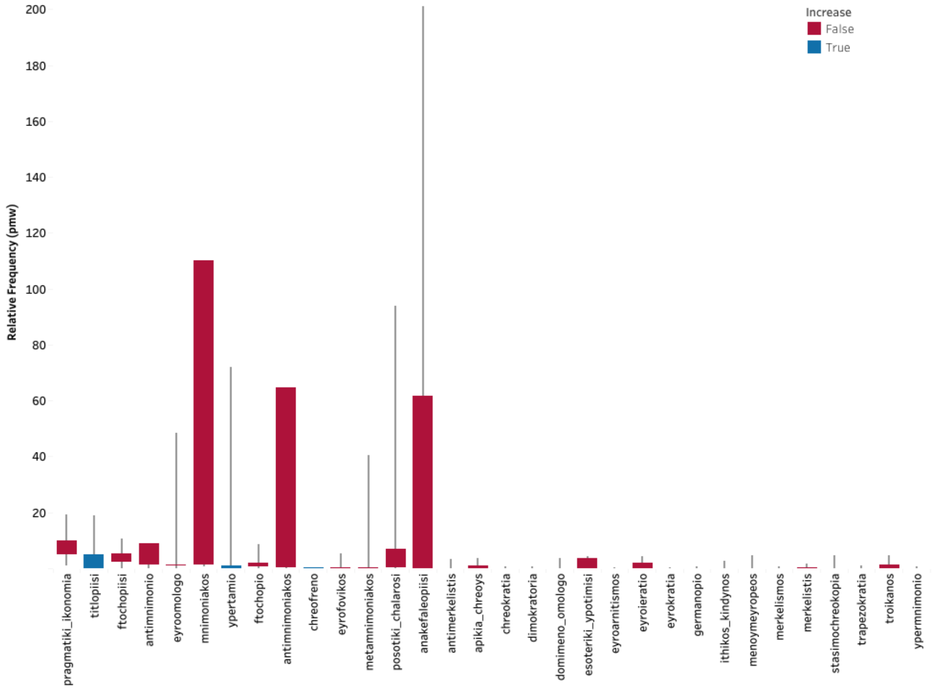


Figure 2: Candlestick graph ordered by declining relative frequencies as of December 2020.

6 Discussion

In the previous section we tracked the frequency development of selected new formations coined during the Greek debt crisis to observe their behavior within a time span of 5.4 years (the end of the crisis) and gain a more fine-grained understanding of their life-cycles within that period. The mere shape of their distribution in the corpus was shown to form a continuum of cases and permitted a glance into the dynamics of the spread process. Although overall, their appearance in journalistic discourse is decaying,[28] the less peaky distributions traced longer paths into the future and characterized the more resilient neologisms. Conversely, high *Peakedness* was an indicator of instability and transient, topical, thus ad hoc uses. These patterns, of course, derive from the different weight of communicative needs that triggered word usage towards the end of the Greek debt crisis and the beginning of the pandemic in 2020. Some new terms were only produced on-demand and never left the *consolidation* (or even the creation) *stage*; high peaks seem to negatively affect the spread of new formations. Others, although slowly diminishing in use, have endured and entered the *establishment stage*; low peaks seem to positively affect the spread of new formations. Kurtosis measures, such as the *Peakedness* values can be used as a top-down, dynamic quantitative filter to monitor "lexical sustainability", together with similar metrics, such as *Dispersion.* As distributional evidence showed (see Table 1), extremely peaky lemmas have very low *Dispersion*, since they not uniformly spread in the corpus. Not all less peaky lemmas, however, are more uniformly spread (see Table 1, *metamnimoniakos*). Moreover, there are peaky lemmas which are indeed uniformly spread (see Table 1, *anakefaleopiisi).* Therefore, *Dispersion* measures partially correlate with *Peakedness* scores in an inverse relationship ($r_s = -0.575$, $p_{\text{2-tailed}}$

[27] They almost resemble the "Spinning Tops" candlestick pattern in financial jargon, representing little movement in the market.
[28] At least for the specific newspaper we selected on theoretical grounds (see section 5.1).

= 0.001). In fact, since *Dispersion* measurements are obviously affected by the duration of the time span (on a horizontal view of the data), they can only be used complementary to *Peakedness* evaluations (an essentially vertical view of the data), for instance as an initial cut-off threshold of under-dispersed, thus ephemeral new formations (for Table 1, see *Dispersion* scores ≤ 0.60).

We nonetheless emphasize the potential of such tangible criteria that corpus-linguistic methods and tools offer and their diagnostic (as if prognostic, for *Peakedness*) value in assessing the "success stories" of different new formations in their way to establishment in a community of language producers / receivers. Once fine-tuned empirically they can contribute to the development of solid prediction models (cf. Jiang et al. 2021) or simply serve as practical heuristics complemented by corpus-based, bottom-up lexicographical assessment.

7 Concluding Remarks and Future Research

In this paper there has been an attempt to conjointly illustrate the importance of quantitative explorations and measurements, like *Peakedness, Dispersion* etc., applied on a newspaper corpus for a selected list of new formations – designating aspects of the Greek debt crisis and covering certain criteria – in order to co-estimate their behaviour and evolution retroactively in time. Specific tools and statistical procedures were used, and it was shown that *Peakedness* was an important indicator for the sustainability of emerging formations. In addition to this, we assume that text type and media diversity should be a second crucial factor for the success of the new lexical units. Thus, the same approach, i.e. a retroactive analysis to lists of new formations of certain periods should be further applied to the entire corpus, including the rest of the newspapers, in order to also detect and evaluate their diffusion to the speech community, thus the beginning of institutionalization, the key notion for dictionary inclusion.

The results of the corpus exploration and analysis can prove both linguistically and lexicographically very profitable. It seems to be important for linguistic research to identify and register new formations – exhibiting a certain repetitive use and a form-meaning stabilization (thus entering the *consolidation stage*) – in a dynamic electronic lexicon of neologisms in order to monitor their behaviour and evolution for certain selected periods. Although many of these registered new formations may not yet be at the *stage of establishment* within the *perspective of institutionalization* (Kerremans 2015: 40, see Schmid 2008), which according to our presentation (see section 4) would signal the step for inclusion in general-purpose dictionaries, their monitoring seems a. to build a valuable linguistic information tank and b. to further facilitate the answer to the desideratum of the inclusion (or not) decision for neologisms.

Our proposal for a dynamic electronic lexicon of neologisms is being supported by the evolution and the enormous possibilities of corpus linguistics and electronic lexicography. Both fields contribute to the investigation, monitoring and recording of a huge amount of data. Taking advantage of these new possibilities the *Research Centre for Scientific Terms and Neologisms of the Academy of Athens* is planning to expand its research area from neologisms for inclusion in general-purpose dictionaries to the construction of a broader dynamic lexicon of (possible) neologisms.

8 References

[Afentoulidou & Christofidou] Αφεντουλίδου, Β. & Χριστοφίδου, Α. (2017 [2018]). Σώμα Εποπτείας Νεολογισμών της Νέας Ελληνικής: Σχεδιασμός και κειμενική ταξινόμηση. Στο Α. Χριστοφίδου (επιμ.) *Όψεις της σωματοκειμενικής γλωσσολογίας: Αρχές, εφαρμογές, προκλήσεις. Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 14, σσ.129-180. Αθήνα: Ακαδημία Αθηνών.

[Anastassiadis-Symeonidis] Αναστασιάδη-Συμεωνίδη, Α. (2003). *Αντίστροφο λεξικό της Νέας Ελληνικής*. Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

[Anastassiadis-Symeonidis, Alexiadou & Nikolaou] Αναστασιάδη-Συμεωνίδη Α., Αλεξιάδου, Χ. & Νικολάου, Γ. (2009). Ηλεκτρονική βάση νεολογισμών της Νέας Ελληνικής. Στο Α. Χριστοφίδου (επιμ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 9-10, σσ. 419-439. Αθήνα: Ακαδημία Αθηνών.

[Christofidou] Χριστοφίδου, Α. (2015). Εισαγωγή. Σχεδιασμός και παρουσίαση της έρευνας. Στο Α. Χριστοφίδου (επιμ.) *Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 13, σσ.11-24. Αθήνα: Ακαδημία Αθηνών.

[Christofidou, Afentoulidou, Karasimos & Dimitropoulou] Χριστοφίδου, Α., Αφεντουλίδου, Β., Καρασίμος, Θ. & Δημητροπούλου, Ε. (2013). Ηλεκτρονικό πρόγραμμα Νεοδημία. Προκλήσεις και δικτυο-λύσεις. Στο Α. Χριστοφίδου (επιμ.) *Δημιουργία και μορφή στη γλώσσα, Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 12, σσ. 198-243. Αθήνα: Ακαδημία Αθηνών.

[Christofidou, Karasimos & Afentoulidou] Χριστοφίδου, Α., Καρασίμος, Θ. & Αφεντουλίδου, Β. (2014). Έλεγχος, παρακολούθηση και ταξινόμηση νεολογισμών με το ηλεκτρονικό πρόγραμμα *Νεοδημία*: Η προσέγγιση των νέων δανείων. Στο G. Kotzoglou, K. Nikolou, E. Karantzola, K. Frantzi, I. Galantomos, M. Georgalidou, V. Kourti-Kazoullis, C. Papadopoulou, E. Vlachou (επιμ.) *Selected Papers of the 11th International Conference on Greek Linguistics,* σσ. 1850-1868. Rhodes: University of the Aegean.

[Nikolaou & Anastassiadis-Symeonidis] Νικολάου, Γ. & Αναστασιάδη-Συμεωνίδη, Α. (2017). Ο ρόλος του παγκόσμιου ιστού στη μελέτη της νεολογίας και της μορφολογικής ανάλυσης: Η περίπτωση των νεολογικών επιθέτων της Κοινής Νεοελληνικής. Στο Α. Χριστοφίδου (επιμ.) *Όψεις της σωματοκειμενικής γλωσσολογίας: Αρχές, εφαρμογές, προκλήσεις. Δελτίο Επιστημονικής Ορολογίας και Νεολογισμών*, 14, σσ. 271-285. Αθήνα: Ακαδημία Αθηνών.

[Varoufakis] Βαρουφάκης, Γ. (2011) *Κρίσης λεξιλόγιο. Οι οικονομικοί όροι που μας καταδυναστεύουν*. Αθήνα: Ποταμός

Andersen, G. & Hofland, K. (2012). Building a large corpus based on newspapers from the web. In G. Andersen (ed.) *Exploring newspaper language.* Amsterdam/Philadelphia: John Benjamins, pp. 1-28.

Aubry, S, Cartier, E. & Stirling, P. (2018). Néonaute: Mining web archives for linguistic analysis. Presentation at the *International Internet Preservation Consortium Web Archiving Conference, 12-15 November 2018*. Wellington, NZ.

Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

Cabré, T. & Nazar, R. (2012). Towards a new approach to the study of neology. In *Neologica,* 6, pp. 63-80.

Cartier, E. (2016). Néoveille, système de repérage et de suivi des néologismes en sept langues. In *Neologica,* 10, pp.101-131.

Cartier, E. (2019). Néoveille, plateforme de détection, de repérage et de suivi des néologismes en corpus dynamique. In *Neologica,* 13, pp. 23-54.

Christofidou, A., Afentoulidou, V., Karasimos, A. & Vassiliadou, R. (2020). Compoundhood: Defining, extracting and monitoring multiword A+N compounds in a database of Greek neologisms. In S. Markantonatou, A. Christofidou (eds.) *Multiword expressions: Drawing on data from Modern Greek and other languages. Bulletin of Scientific Terminology and Neologisms,* 15. Athens: Academy of Athens, pp. 133-192.

Connor Martin, K. (2019). New Words Prioritization Engine: A system for evaluating multiple data inputs to prioritize neologisms for inclusion in dictionary projects. Presentation at *Globalex 2019: Workshop on Lexicography and Neologism, 8 May 2019*. Bloomington Indiana.

De Schryver, G.-M. (2002). Web for/as corpus: A perspective for the African languages. In *Nordic Journal of African Studies,* 11(2), pp. 266-282.

Fischer, R. (1998). *Lexical change in present-day English. A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms*. Tübingen: Gunter Narr Verlag.

Freixa, J. & Torner, S. (2020). Beyond frequency: On the dictionarization of new words in Spanish. In *Dictionaries,* 41(1), pp. 131-153.

Gabrielatos, C., McEnery, T., Diggle, P., Baker, P. & ESRC (Funder). (2012). The peaks and troughs of corpus-based contextual analysis. In *International Journal of Corpus Linguistics*, 17(2), pp. 151-175.

Gérard, C., Falk, I. & Bernhard, D. (2014). Traitement automatisé de la néologie : Pourquoi et comment intégrer l'analyse thématique ? In F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschäfer, S. Prévost (eds.) *Actes du 4e CMLF, Berlin, 19-23 July 2014*, SHS Web of Conferences 8. France: EDP Sciences, pp. 2627-2646.

Grieve, J., Nini, A. & Guo, D. (2016). Analyzing lexical emergence in Modern American English online. In *English Language and Linguistics,* 21(1), pp. 99–127.

Guerra, A. R. (2016). Dictionaries of Neologisms: A review and proposals for its improvement. In *Open Linguistics*, 2, pp. 528-556. Accessed at doi.org/10.1515/opli-2016-0028 [28/02/2021].

Hagen, T. M. (2012). Automatic topic classification of a large newspaper corpus. In G. Andersen (ed.) *Exploring newspaper language.* Amsterdam/Philadelphia: John Benjamins, pp. 111-130.

Herberg, D., Kinne, M. & Steffens, D. (2004). *Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen.* Berlin: de Gruyter.

Hilpert, M. & Gries, S. Th. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. In *Literary and Linguistic Computing*, 24(4), pp. 385-401.

Hohenhaus, P. (2005). Lexicalization and institutionalization. In P. Štekauer, R. Lieber (eds.) *Handbook of word-formation*. Dordrecht: Springer, pp. 353-373.

Jiang, M., Shen, X. Y., Ahrens, K. & Huang, C.-R. (2021). Neologisms are epidemic: Modeling the life cycle of neologisms in China 2008-2016. In *PLoS ONE,* 16(2). Accessed at doi:10.1371/journal.pone.0245984 [28/02/2021].

Kerremans, D. (2015). *A web of new words. A corpus-based study of the conventionalization process of English neologisms*. Frankfurt: Peter Lang Edition.

Kerremans, D., Prokić, J., Würschinger, Q. & Schmid, H.-J. (2018). Using data-mining to identify and study patterns in lexical innovation on the web: The Neo Crawler. In *Pragmatics and Cognition*, 25(1), pp. 174-200.

Klosa-Kückelhaus, A. & Wolfer, S. (2019). Considerations on the acceptance of German neologisms from the 1990s. In *International Journal of Lexicography*, 33(2), pp. 1-18. Accessed at doi:10.1093/ijl/ecz033 [28/02/2021].

Lüdeling, A., Evert, S. & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf, C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam/New York: Rodopi, pp. 7-24.

Renouf, A. (2007a). Corpus development 25 years on: From super-corpus to cyber-corpus. In R. Facchinetti (ed.) *Corpus Linguistics 25 Years on*. Amsterdam: Rodopi, pp. 27-49.

Renouf, A. (2007b). Tracing lexical productivity and creativity in the British Media: 'the Chavs and the Chav-Nots'. In J. Munat (ed.) *Lexical creativity, texts and contexts*. Amsterdam/Philadelphia: John Benjamins, pp. 61–89.

Renouf, A. (2013). A finer definition of neology in English: The life-cycle of a word. In H. Hasselgård, J. Ebeling, S. O. Ebeling (eds.) *Corpus perspectives on patterns of lexis*. Amsterdam/Philadelphia: John Benjamins, pp. 177-208.

Renouf, A. (2016). Big data and its consequences for neology. In *Neologica,* 10, pp. 15-37.

Schmid, H.-J. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. In *Anglia. Zeitschrift für Englische Philologie,* 126(1), pp. 1-36.

Teubert, W. (1998). Korpus und Neologie. In W. Teubert (ed.) *Neologie und Korpus*. Tübingen: Gunter Narr Verlag, pp. 129-170.

Thelwall, M. (2018). *Big data and social web research methods*. University of Wolverhampton. Accessed at http://www.scit.wlv.ac.uk/~cm1993/papers/IntroductionToWebometricsAndSocialWebAnalysis.pdf [28/02/2021].

Acknowledgements

This research is funded by the Research Committee of the Academy of Athens.

EURALEX XIX
Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Bi- and Multilingual Lexicography

Frisian dictionaries, digitized from A to Z

Drenth E.[1], Sijens H.[1], Van de Velde H.[1,2]

[1] *Fryske Akademy, Leeuwarden (NL)*
[2] *Universiteit Utrecht, Utrecht*
edrenth@fryske-akademy.nl; hsijens@fryske-akademy.nl; hvandevelde@fryske-akademy.nl

Abstract

This paper approaches dictionaries as lexical resources with functions for target audiences, which benefit from a strictly defined data format, which means less work and improved interchangeability. Code generation in a reliable automated build process provides validation and documentation. Stable services provide functions that can be realized within the data format. The software can run straight away with a complete docker setup. In this way, creating a dictionary becomes primarily a matter of editing or converting data, for instance with an XML editor that supports editors by means of generated validation and documentation.

Keywords: Frisian; TEI; Universal Dependencies; eXist-db; dictionaries

1 Introduction

1.1 Background

Compiling a Frisian academic dictionary is one of the raisons d'être of the Fryske Akademy (FA). Nowadays, over eighty years since the institute's formation, the creation of Frisian user dictionaries is still a core activity. For the longest time, the process typically spanned several years and resulted in a paper publication. In the early days of digitization (around 1980), the dictionaries that were edited were entered into a database as ASCII text. For coding (for typesetting and other purposes), an at sign was added (e.g., @C = italic). The text was then stored in a full-text database. From this database, the text was generated and converted for printing. Later, the need arose for dictionaries that were digitally accessible. This need was initially met by using scripting languages to query the documents or convert them to queryable formats such as XML or a database. The next step was the introduction of the Fryske Akademy's first 'born digital' dictionary, the Online Dutch Frisian Dictionary (*Online Nederlands-Fries Woordenboek*, ONFW) (Drenth 2017). The ONFW is digital in design. However, the target format is not generic, and this interferes with standardized querying (and editing).

1.2 The next step: Dictionaries as datasets

The Fryske Akademy is now taking the next step in the development of its lexical resources and tools: defining a standardized format and developing generic applications for that format. With this step, dictionaries become lexical resources that can be continually edited and queried, rather than the static end product of an extended lexicographic process. The focus is on developing applications that meet the needs of user groups, and the data format plays a crucial role in this process.

2 Basic principles and preconditions

The major changes in the lexicographic process require a clear formulation of the basic principles and preconditions of the new structure. These are the basic principles and preconditions drawn up by the Fryske Akademy for the lexical infrastructure for the Frisian language.

2.1 Current editing environment

Maintaining the work processes linked to the current editing environment is a precondition. ONFW's editing environment is maintained in collaboration with the *Instituut voor de Nederlandse Taal* (Dutch Language Institute), which secures the lexicographic infrastructure for the Dutch language.

2.2 Serving a variety of end-users

The format and the solutions will be geared to different user groups, such as linguists, professional users (writers, translators, journalists, teachers, civil servants, lawyers), native speakers, and language learners (both L1 and L2 learners). This means that the format should be able to hold information in a variety of formats, handle varying degrees of detail, and make available both comprehensive and simple search options and search results.

2.3 More efficient editing

Lexicographers need to be able to edit lexical information intuitively. In doing so, they require drop-down lists, additional documentation, and validation of their input. A community of language users should be able to contribute in a simple way (citizen science). These contributions must then be identified, edited, and validated by professional lexicographers.

2.4 Quick and easy queries

A data format is a means for querying lexical resources. This requires functions that are stable and clearly defined in terms of input, output, and error handling. The work is greatly simplified if the format's structure and content are strictly defined, so that it is clear where specific information can be found. In addition to online queries, the data should be accessible offline using proprietary tools.

2.5 Ease of conversion

There are many initiatives in the field of lexicography, both open source and proprietary, such as TEI Lex-0 (Salgado 2019), Freedict, grammarly, and wordnet (see Section 4). The option to link up with these initiatives is an important asset and usually entails converting the format, dynamically or otherwise, to that of the target initiative. Such a conversion can be facilitated by (i) the use of a well-defined format, (ii) that is semantically clear and consistent, and (iii) content that does not contain hidden functions (e.g., a # in text with special meaning) to avoid having to carry out an error-prone analysis of optional content.

2.6 Sustainability

The solutions to be set up or developed should be maintainable with the least possible effort: taking no rare technical expertise, little time, and relatively little money. The work is to be carried out with proven methods and technology from the ICT industry. The format co-determines the options: with a leaner and simpler format, it will be easier to maintain the software solutions.

2.7 Adaptability

In the future, the current format may not be suited for new information. In that case, the format may have to be adapted. As a result, the data formats and their functions will start to diverge. A clear procedure must be put in place to avoid problems due to changes and to facilitate switching to later versions.

2.8 Open standards

The significance of open standards is not disputed, so the connection to and use of open standards is a given. The solution itself will also be made available to all, and wider use will be advocated.

2.9 FAIR Principles

The Fryske Akademy is an academic institution and as such adheres to the FAIR principles[1]. The compliance with the FAIR principles and open standards will pay off in terms of interoperability, acceptance, available tools, and the collaboration between parties.

3 Current standards

Based on the principles set out above, we have considered the following standards and infrastructures that could be part of the solution.

3.1 TEI

TEI[2] stands for Text Encoding Initiative; this is a well-established, widely used open guideline for encoding text. It is used primarily for the encoding and online publication of historical manuscripts, and it contains a module for dictionaries and support for linguistic annotations. TEI is comprehensive and designed to support a wide range of scenarios. As a result, there are often many options for encoding and few forced options. To meet the drawbacks of its wide standard, TEI has a powerful mechanism for customization: One Document Does It All (ODD). With ODD, the guidelines can be geared to specific applications by defining which components are used and how they are applied.

3.2 TEI Lex-0

TEI Lex-0 (Bański 2017) is an initiative intended to establish an open dictionary standard based on TEI that is better suited to digital processing. It is an ODD that is the result of a community process; it focuses on limiting the opportunities available in TEI. TEI Lex-0 is primarily intended as a format for existing dictionaries in order to improve interoperability.

3.3 Universal Dependencies

Universal Dependencies[3] (UD) is an open framework for the consistent linguistic encoding of text. UD contains guidelines for word type, morphosyntactic description, treebanks for many languages, and tools such as a part-of-speech tagger. As

[1] https://www.go-fair.org
[2] https://tei-c.org
[3] https://www.universaldependencies.org

such, UD provides a solid foundation for natural language processing.

3.4 Freedict

Freedict[4] is a technical open-source project containing many translation dictionaries in TEI format. Most dictionaries are flat and mainly focus on word-to-word translation. There are multiple applications for Freedict, especially for Android and Linux.

3.5 Ontolex

Ontolex (McCrae 2017) is an open semantic web model used to classify information (language/words) lexicographically. Semantic web technologies allow computers to analyse information and apply a form of artificial intelligence across various information domains. When lexicographic information is converted to, and made available in, the semantic web, it also becomes available for the applications that use artificial intelligence.

3.6 ELEXIS

ELEXIS[5] is not a standard but rather an infrastructure. Internationally speaking, the European Lexicographic Infrastructure (ELEXIS) (elex.is) is of great importance. It is an initiative intended to make linguistic data openly accessible and to make language resources available. There are several ways to connect to ELEXIS, for instance by converting existing material to TEI Lex-0 or Ontolex. Tools available from ELEXIS include Sketch Engine, a large corpus system; Lexonomy, an online editing and publishing environment for dictionaries; and Elexifier, for converting existing dictionaries to TEI Lex-0 or Ontolex.

4 Specifics

In this section, we outline the choices and approaches by which we have arrived at our solutions.

4.1 Choices

We have opted for a solution based on TEI and Universal Dependencies. These two international standards have been used by the Institute and its partners for many years. Both standards are actively maintained, and several matching tools are available, such as the TEI stylesheets, oxygen, udpipe, and teipublisher. The target audience and approach of the Freedict project is clearly different from creating professional dictionaries for more experienced language users and scholars. Therefore, the Fryske Akademy has not selected this project as a foundation for its dictionaries. The availability of Frisian in Freedict dictionaries is relevant, however, as Frisian is a low-resource language that is of little interest to commercial companies. Therefore, we will be supplying data to the Freedict project.

TEI Lex-0 is marketed primarily as a format that facilitates interoperability between lexicographic resources. It can also be used as a basic format for compiling dictionaries. TEI Lex-0 restricts the space allowed by TEI, but still offers a lot of freedom. This makes it less suitable for software development because it is not certain where information can be found within the data structure, nor how information can be recognized. We have decided to set up our own format, very similar to TEI Lex-0, but with more restrictions. This provides a better foundation for software development. It also makes it easier to convert the format from the editing environment. We have developed a conversion to TEI Lex-0 for interoperability with ELEXIS and others.

As a semantic web format, Ontolex is not a suitable format for editing or building a dictionary service. In time, lexicographic data will be published as Linked Open Data.

4.2 Approach

For the TEI customization, an ODD was developed in which the primary objective was to achieve a simple structure that could be properly validated. In the development stage, five different dictionaries were simultaneously converted to the target format. These five dictionaries are the ONFW, the Frysk Hânwurdboek (Duijff 2008), the Dutch-Frisian dictionary (Visser 1985), the Frisian-Dutch dictionary (Zantema 1984) and the legal dictionary (Duijff 2000). In addition, a REST service was developed for querying the target format. By running these development paths in parallel, we were able to test the target format and service in practice. During the development process, we carefully considered which components were generic. These components and their development processes were separately made available as open source, see section 5.

4.3 Components

Figure 1 visualizes current components used in lexicography at the Fryske Akademy.

Figure 1: schematic of components for lexicography. Generic parts are marked in green

[4] https://freedict.org
[5] https://elex.is

4.3.1 ODD with library

An ODD is an XML file in TEI format that can be used to capture a TEI customization. The resulting scheme roughly accommodates word forms with grammatical designations, homonyms, meanings, paradigms, synonyms, variants, examples, collocations, and proverbs. Translations can be included as well. In the meta-information, editors can indicate which of the following functions are supported.

- Formtranslation: translation of words
- Textsearch: search for words in text
- Synonyms: search for synonyms of words
- Variants: search for variants of words
- Compounds: search for compounds of words
- Pronunciation: search for the pronunciation of word forms
- Hyphenation: search for the hyphenation of word forms
- Usage: search for usage information regarding word forms
- Stress: search for emphasis in word forms
- Definition: search for definition of word forms
- Grammar: search for grammar including position regarding words
- Paradigm: search for the paradigm in headwords
- Examples: search for examples of words
- Collocations: search for collocations of words
- Proverbs: search for proverbs containing specific words

These functions are available in the REST service, see section 4.3.3.
Using the TEI stylesheets and some freely available tools, the ODD is translated into a validation file containing documentation. This translation is part of a tightly defined, repeatable Maven[6] build process. Once a version of the format has been approved, it is published in a globally available repository: Maven Central. The published version can be used in dictionary projects, for instance by editors who are aided in their editing environment by drop-down lists, documentation of items to be added, and the option to validate their work, or by ICT staff who write conversions and want to validate the result.

4.3.2 Exist-db

Dictionaries based on our method use XML, which has prompted us to store them in an XML database. We have opted for exist-db[7] because exist-db has a long track record, is open source, standards-based, has an active community, and because the FA and its partners are familiar with exist-db.

4.3.3 REST service

In exist-db, a REST service was developed to query dictionaries. This service provides powerful, lucene-based search functions that can be used to search for translations of words, grammatical properties, pronunciation, examples, etc. (See the list under 'ODD with library'.) The results are presented in a simple standardized manner in Json for further processing, see table 1.

/translate?form=dag&lang=nl	/paradigm?form=wurkje&lang=fry	/synonyms?form=fyts~
	 0: "pronoun.clitic"	

Table 1: example queries and results

4.3.4 Application

In addition to this service, a web application was developed in three languages, English, Dutch, and Frisian, suitable for computers and smartphones. Since user needs and wants may vary, this application is secondary to the REST service as far as we are concerned.

5 Dissemination

5.1 REST services

In accordance with the setup described above, four dictionaries are now available as REST services:

- the Frysk Hânwurdboek
- the Frysk Wurdboek Nederlânsk-Frysk
- the Frysk Wurdboek Frysk-Nederlânsk
- the Juridysk Wurdboek

[6] https://maven.apache.org
[7] https://exist-db.org

These services can be accessed at https://frisian.eu/dictionaries.

5.2 Test applications

For a first impression and testing purposes, web applications for these dictionaries are available through the following links:

- the Frysk Hânwurdboek: https://frisian.eu/dictionaries/fhwbapp
- the Frysk Wurdboek Nederlânsk -Frysk: https://frisian.eu/dictionaries/nfwbapp
- the Frysk Wurdboek Frysk-Nederlânsk: https://frisian.eu/dictionaries/fnwbapp
- the Juridysk Wurdboek: https://frisian.eu/dictionaries/jurwbapp

5.3 Software library

A software library for the solution is freely accessible at https://search.maven.org/search?q=a:teidictionaries. The library can be used to validate whether XML documents conform to the format that we have developed. It can also be used to convert XML into programmable objects and vice versa, and to convert XML into TEI Lex-0 with validation.

5.4 Exist-db extension

During the development phase, the need arose for additional features for exist-db: periodic synchronization of articles with exist-db, the ability to properly configure applications, and some search options. These features were donated to the exist-db community: see https://search.maven.org/search?q=a:exist-db-addons and https://github.com/eXist-db/documentation/pull/549.

6 Next steps

Figure 2 visualizes future developments in lexicography at the Fryske Akademy. The sections below describe this in more detail.

Figure 2: schematic of future developments in lexicography

6.1 Editing environment

The current situation is a good starting point for considering future developments. One of the first things we want to realize is a generic editing environment in which both lexicographers and volunteers can edit material, aided by documentation, drop-down lists, validation, and a simple workflow with the phases ‘under consideration’, ‘review’, and ‘approved’.

6.2 Communities

Lexicography is all about language, and language connects people, so, the FA wants to facilitate communities by developing opportunities for feedback.

6.3 Conversions

As mentioned above, a conversion to TEI Lex-0 has been developed, and a conversion to Ontolex will be developed at a later stage. In this way, we aim to connect with ELEXIS and Clarin.

6.4 Integrations

In the near future, this dictionary functionality of this solution will be integrated with other language solutions such as corpora, lexicons, text translation, and spell checking. A GraphQL-based interface is under development for this purpose.

7 In conclusion

Based on analyses of existing solutions and standards, using proven methods and techniques from ICT, we have realized a robust open solution for lexical resources. The solution greatly reduces the amount of work required for publishing lexical data. Moreover, it is a solid foundation for software development, and promotes integration with other initiatives in lexicography. Our solution is available for all languages, and is of particular interest to low-resource languages that cannot afford commercial support. We hope that our initiative will be applied more widely and that a community will emerge to continue the development of the solution.

8 References

Drenth E., Duijff, P., Sijens, H. (2017). Open Access to Frisian Language Material.
Duijff, P. (2000) Juridysk Wurdboek Nederlânsk-Frysk
Duijff, P., Van der Kuip, F., De Haan, R. and Sijens, H. (2008). Frysk Hânwurdboek
McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P., Cimiano, P. (2017). The OntoLex-Lemon Model: Development and

Applications.

Bański, P., Bowers, J., Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms.

Visser, W. (1985). Frysk Wurdboek Nederlânsk-Frysk

Zantema, J.W. (1984). Frysk Wurdboek Frysk-Nederlânsk

8.1 Websites

Apache lucene query syntax. Accessed at: http://www.lucenetutorial.com/lucene-query-syntax.html [2016].

CLARIN - European Research Infrastructure for Language Resources and Technology. Accessed at: https://www.clarin.eu/ [2018].

Docker. Accessed at: https://www.docker.com [2019].

Dictionary software. Accessed at: https://search.maven.org/search?q=a:teidictionaries [2020].

European lexicographic infrastructure. Accessed at: https://elex.is [2019].

Exist-db Xml database. Accessed at: https://exist-db.org [2018].

Fair principles. Accessed at: https://www.go-fair.org [2021].

Free dictionaries. Accessed at: https://freedict.org [2020].

Maven build and dependency management tool. Accessed at: https://maven.apache.org/ [2016].

R package for Tokenization, Tagging, Lemmatization and Dependency Parsing. Accessed at: https://www.rdocumentation.org/packages/udpipe/ [2020].

TEI guidelines. Accessed at: https://tei-c.org [2018].

TEI publisher. Accessed at: https://teipublisher.com [2018].

Universal Dependencies. Accessed at: https://www.universaldependencies.org [2019].

Acknowledgements

We want to thank our colleagues at the FA for their cooperation and patience. We would also like to thank the communities, especially those around TEI, UD and eXist-db, for thinking along, which improved the solution and made it more interoperable. Finally, we would like to thank the province of *Fryslân* and the Dutch government for providing the finances that made these developments possible via project PF-2017/172874 (ONFW) and 01785576 (*Taalweb*).

Charting A Landscape of Loans. An e-Lexicographical Project on German Lexical Borrowings in Polish Dialects

Meyer P., Hentschel G.

[1] Fryske Akademy, Leeuwarden (NL)
[2] Universiteit Utrecht, Utrecht
edrenth@fryske-akademy.nl; hsijens@fryske-akademy.nl; hvandevelde@fryske-akademy.nl

Abstract

This paper reports on an ongoing international project of compiling a freely accessible online Dictionary of German Loans in Polish Dialects. The dictionary will be the first comprehensive lexicographic compendium of its kind, serving as a complement to existing resources on German lexical loans in the literary or standard language. The empirical results obtained in the project will shed new light on the distribution of German loanwords among different dialects, also in comparison to the well-documented situation in written Polish. The dictionary will have a strong focus on the dialectal distribution of Polish dialectal variants for a given German etymon, accessible through interactive cartographic representations and corresponding search options. The editorial process is realized with dedicated collaborative web tools. The new resource will be published as an integrated part of an online information system for German lexical borrowings in other languages, the *Lehnwortportal Deutsch*, and is therefore highly cross-linked with other loanword dictionaries on Polish as well as Slavic and further European languages.

Keywords: lexical borrowings; Polish dialectology; dialect lexicography; XML database

1 Introduction and Lexicographical Background

This paper reports on an ongoing project of compiling a freely accessible online Dictionary of German Loans in Polish Dialects (henceforth, DGlPd) funded by the German Research Foundation (DFG) and jointly carried out by the Institute of Slavic Studies at the University of Oldenburg and the Leibniz Institute for the German Language (IDS), in cooperation with the Institute of the Polish Language of the Polish Academy of Sciences, Kraków (IJP PAN). The project started in April of 2019 and will end in 2022/23.

Scientific research on German loans in Polish (and other Slavic languages) started in the 19th century. Older studies of the phenomenon did not have the opportunity to take advantage of the enormous progress of historical and dialectal lexicography in Poland (and other Slavic speaking countries, cf. SALDAS for bibliographical references) in the decades after World War II up until today. The DGlPd will be the first comprehensive lexicographic compendium of its kind, serving as a complement to the Dictionary of German Loans in the Polish Written and Standard Language (WDLP).

The empirical results obtained in the project will shed new light on the distribution of German loanwords among different dialects, also in comparison to the well-documented situation in written Polish as documented in the WDLP (cf. Hentschel 2009; 2010). One of the general questions to be asked is to what extent the transfer of German loans was mediated by dialects and to what extent dialects took over loans from the written language. A small quantitative pilot study was carried out in preparation for the project, based on entries in the SGP for the initial letters <A> and <F>, not counting derived forms, totalling to 529 elements. The study indicates that Polish "core dialects" such as those of Lesser Poland, Greater Poland and Mazovia that had almost not at all, or only in some parts, been under long-term German (speaking) rule (Greater Poland, Mazowia) or just for a comparatively limited time (Lesser Poland from the end of the 18th century to World War I, though) show significantly more overlap in German loans with written Polish than dialectal areas with a centuries-long direct contact situation, including bilingualism in large parts of the population. Within the latter group, e.g. Silesian and Kashubian (see below for the linguistic status of the latter) have borrowed extensively from German but hardly had an impact on the development of Literary Polish, while core dialects definitely had. In late medieval until early modern times a broad scale direct contact of Poles with German migrants to Poland was a main source of lexical transfer in large parts of Poland. From the 19th century to the end of World War II, such a situation was given only in Silesia, Kashubia and parts of East Prussia. Nevertheless, there is a considerable amount of German loans dating back to the 19th century, when large parts of Poland (viewed in its contemporary boarders) were under Prussian or Habsburg rule. In the DGlPd, special attention will thus be paid to the amount of overlap of core dialects and non-core dialects in German loans on the one hand and to the overlap of different dialects with written Polish.

There are two theoretical aspects of lexical borrowings that have recently received more attention in the literature and that are challenges to our investigation. The first one concerns the discussion in contact linguistics on code switching and code mixing (cf. Muysken 2000) concerning insertional code switching of single lexical items and so-called spontaneous or nonce borrowings (cf. for example Poplack & Dion 2012), given that adaptation of transferred elements in a recipient language is not always a reliable cue to the problem. In non-core dialects, the speakers of which are at least to some extent bilingual, with German as the dominating "H-variety", the mere presence of a lexical item in the standard dictionary of Polish Dialects, SGP, is no guarantee for acceptance of the item by the speech community as a whole. Here only the number and distribution of witnesses can roughly mirror the degree of acceptance. The second aspect is an issue in typological investigations of borrowing (cf. Haspelmath & Tadmor 2009). Following the lines of Haspelmath (2009), an attempt will be made in the DGlPd to differentiate between “insertions” (note: not in the sense of “insertional code switching / mixing”) into the lexicon (roughly: cultural borrowings, to name new things or concepts), “replacements” (roughly: borrowings replacing an older, typically native Polish word), and – as we would like to add – constellations of “coexistence” involving both the new loanword and an older word, with some type of reorganization of the denotational scope of the borrowed and the native (or at least older) word.

In what follows, we will first present the data sources and scope of the lexicographical project (section 2) before outlining the editorial process (section 3) and the IT tools used in this process (section 4). Section 5 gives an overview of the planned entry microstructure, while relevant details of the online presentation are covered in section 6. #wrapup/conclusion in section 7.

2 Source and Scope of Data

The central, though not exclusive, source of data for the project is the Dictionary of Polish Dialects (SGP), which is being compiled in the Department of Dialectology of the IJP PAN. Begun in 1982, the SGP now counts nine released volumes, up to the lemma *hyżki*. This means that most of the SGP data, roughly three quarters, has to be taken from the card index to the SGP. It should be noted that the SGP does not give any information on the origin of the words described.

In accordance with the conception of the SGP, all traditionally acknowledged dialects of Polish will be taken into account in the dictionary: Silesian, Greater Polish, Lesser Polish, Mazovian, Northern and Southern Kresy, as well as Kashubian, which had been regarded as a Polish dialect in Poland for a long time. This point of view can be found in the two-volume compendium *Języki indoeuropejskie – The Indo-European languages* (Bednarczuk 1986-88), authored by highly acknowledged representatives of Polish linguistics of that time (vol. II, pp. 919f). In 2005, Kashubian was officially given the status of a regional language. Since the Polish language law explicitly rules out that a dialect of the official language of the state (i.e. Polish) can be given the status of a regional language, this means that Kashubian is no longer regarded as a dialect of Polish (USTAWA §19).

Similar to the WDLP, only loanwords that are not themselves loans in German (coming from Italian, French, Latin, to name the most important ones) will be considered. Given the considerable number of German loans with German(ic) etymology in Polish dialects (we estimate about 5,000 elements, without derived forms, compared to some 2,500 in Written Polish according to WDLP), the number of loans from other (“Western”) languages, which have possibly been mediated by German, probably has a similar order of magnitude, which would surpass the practical possibilities of a 3-year project. There is, however, a more qualitative justification for limiting the scope of loanwords. For each possibly mediated loan in Polish it would be necessary to determine whether the loan has in fact been mediated by German or whether it is a direct loan from some other language – a question that, incidentally, in many cases would have to be left open anyway. In the latter case, we would expect these words to have been borrowed into Standard / Cultural Polish and only later taken over by dialects as direct language contact and bilingualism was restricted to the elites. Until now there has been no systematic and comprehensive investigation of such loans in Standard / Cultural Polish, with studies like Walsleben (1997) covering only chronological or thematic parts of the lexicon. In other words, decisive prerequisites for the inclusion of such, possibly mediated, elements are currently lacking, whereas German loans with German(ic) etymology in Standard / Cultural Polish have been fixed and documented by the WDLP.

In one respect, the DGlPd will not follow the conception of the WDLP. The latter does not comprise German loans, even with Germanic etymology, if they were mediated by Czech to Cultural / Written Polish before the beginning of the 17th century (the Czech impact on Polish vanished afterwards). Our different approach is motivated by the possibility that a German loan, though mediated by Czech to Written Polish, has been taken over into some Polish dialects (e.g. Silesian) directly from German. There are, for example, some language islands of dialectal Czech in Upper Silesia, which stand in direct contact with varieties of Polish. If a German loan fulfils the condition of German(ic) etymology, then it will be included in the DGlPd. This means: (i) A German etymon under consideration must not be described as being transferred into German from other languages, mainly Latin and Romance languages such as Italian and French. (ii) In order to categorize a word in a receiving language as being transferred from a donor language the corresponding word in the latter should exhibit a substantial affinity to the potential source word: (a) in expression, taking into account known interlingual sound substitutions, here between German and Polish, as well as intralingual sound relations between dialects of the receiving language and universal phonological processes and (b) in meaning, taking into account semantic processes such as the narrowing of meanings from source word to loan as well as metaphoric and metonymic meanings shifts. Last but not least, for classifying a word as a loan from some other language socio-historical plausibility is an important aspect.

3 Compilation Process

The compilation of the DGlPd proceeds roughly as follows. As a first step, proto-entries are created that each connect phonetically similar Polish dialectal word forms, taken from a comprehensive Polish word index of the SGP (cf. section 4), to a small candidate set of candidate German etyma (due to a considerable amount of phonetic variation of Polish dialectal words, their relation to possible German etyma is in many cases rather opaque). After that, detailed information on the Polish words, in particular on their attested meanings and dialectal distribution, obtained from the published volumes and the card index of the SGP, is inserted into these proto-entries. As a last step, the final entries are constructed by splitting and merging the proto-entries according to an etymological assessment guided by the attested word senses, adding an essential commentary on etymology and word history in general.

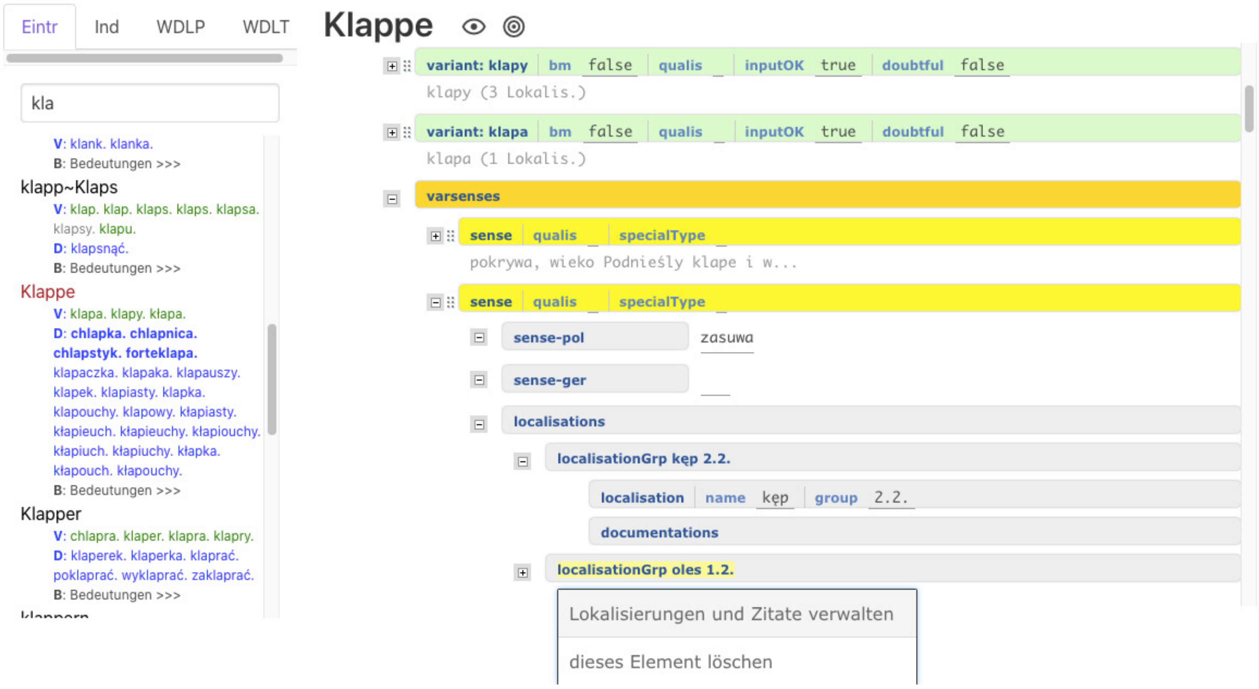
As indicated above, roughly three quarters of the material needed for the DGlPd still only exists in the form of a huge number of cards that, for the time being, have been sorted only on the basis of unlemmatized expressions. This means that the project partners at the IJP PAN in Kraków must manually select all possibly relevant cards for each loan preliminarily selected on the basis of the alphabetical index. This involves (i) checking the card index for expression variants (and there can be many at very different alphabetical positions, cf. Polish dialectal equivalents to German *Hundsfott* as noted in SGP: *hunctwot, huncwót, huncwód, huncót, huncwat, hunzwot, hunsot, huńcot, huńcót, hucwont, hucwont, hacnont, hyncwant, hicnond, uncwot, wuncwot*); (ii) classifying them for different meanings; (iii) fixing the areal distribution of meanings, expression variants and derived words; (iv) selecting citations for meanings and, where possible, for expression variants. With an estimated 5,000 entries of the DGlPd (not counting derivative forms), the card index must be consulted for almost 4,000 entries, meaning that tens of thousands of cards will have to be checked to collect the information needed. As the card index has been collated over the course of many decades, by a large number of people, only the colleagues working at the location in Kraków have the competence to interpret the heterogeneous transcriptions and can thus guarantee the reliability of the notation of the expressions and the citations in the DGlPd.

4 Tooling

For the very specific purposes and requirements of the project, an in-house web application with role-based user management has been developed for collaborative compilation and editorial work in Oldenburg and Kraków. Similar to the LeXmart dictionary development framework (Simões et al. 2019), the software uses an XML database management system (in our case, BaseX, https://basex.org) as its storage backend and features the JavaScript library Xonomy (also used in the well-known Lexonomy system, cf. Měchura 2017) as its browser-based XML entry editing component for the fully bilingual German/Polish user interface. The bare XML editor functionality is supplemented by a growing range of project-specific tools.

- Words and their attested senses must be "localized" according to a scheme that differentiates roughly 350 dialects, subdialects and townships. Frequently, a large number of such localisations has to be selected for a single word or word sense. A dedicated input system speeds up this process with autocomplete functionality, multiple selection within dialect groups, and visual feedback as to whether localisations for words and their word senses match. No manual XML editing is required.
- The 'proto-entries' often comprise a large number of word forms and word senses. In the editing process, the senses have to be assigned to the word forms they are attested for. A generic linking mechanism based on universally unique identifiers is used to track the 'cross-referencing' between words and their senses even across entries. Users edit the presence or absence of connections between form and meaning in a convenient matrix representation, again without having to bother with XML or even UUID minutiae. The general mechanism, which does not encode the links within the entries but as separate XML elements, can later be used to e.g. associate Polish words with their etyma.
- For the final editing phase, a comprehensive tool to split and merge entries, reallocate words and senses to other entries, reclassify expression variants as derivatives and vice versa, etc. is needed. Care has to be taken not to inadvertently sever the association between a word reallocated to a new entry and its senses. The number of proto-entries involved in a single consolidation process can be large, implying that the tool must abstract away from the XML source to give the lexicographer the overview needed.
- Reporting of lexicographical activity and results is done through a powerful XQuery-based interface. During editing, time-stamped snapshots of the current entry are stored in the database automatically in regular intervals, which means that the complete project history can be accessed in database queries such that arbitrary kinds of editing processes can be tracked easily. For each (proto-)entry, all of its editing snapshots are only a mouse click away.
- Entries can be accessed and looked up via their German etyma and the Polish words contained in them. In particular, the entire Polish word index of the SGP (approx. 230.000 words, including phonological and morphological variants and derivatives) is available in the browser, as well as the full XML datasets of two other lexicographical projects related to German loanwords in Polish, viz. the WDLP and the WDLT.
- The XML source code for large entries can quickly get too long for convenient editing. An XSLT-based entry preview function allows lexicographers to easily "jump" from any position in the preview to the corresponding position in the XML editor, just by clicking on the position.
- There is a simple, extensible system of macros to enter special characters without having to change (virtual) keyboards.

Figure 1 shows a screenshot of the basic editor system. The tabs on the left-hand side provide immediate access to the dictionary's (proto-) entries, the Polish word index, and the two Polish dictionaries WDLP and WDLT cross-referenced in the DGlPd entries. Figure 2 shows a screenshot of the localisation editing dialog for the expression variant *kłapa*, with a multiple select input option for localisations within a dialect and colour feedback as to whether, in the case shown, localisations provided for the word also appear in the word senses connected to it and vice versa.



kla

V: klank. klanka.
B: Bedeutungen >>>

klapp~Klaps
V: klap. klap. klaps. klaps. klapsa. klapsy. klapu.
D: klapsnąć.
B: Bedeutungen >>>

Klappe
V: klapa. klapy. kłapa.
D: **chlapka. chlapnica. chlapstyk. forteklapa.** klapaczka. klapaka. klapauszy. klapek. klapiasty. klapka. klapouchy. klapowy. kłapiasty. kłapieuch. kłapieuchy. kłapiouchy. kłapiuch. kłapiuchy. kłapka. kłapouch. kłapouchy.
B: Bedeutungen >>>

Klapper
V: chlapra. klaper. klapra. klapry.
D: klaperek. klaperka. klaprać. poklaprać. wyklaprać. zaklaprać.
B: Bedeutungen >>>

Klappe

variant: klapy | bm false | qualis _ | inputOK true | doubtful false
klapy (3 Lokalis.)

variant: klapa | bm false | qualis _ | inputOK true | doubtful false
klapa (1 Lokalis.)

varsenses

sense | qualis _ | specialType _
pokrywa, wieko Podnieśly klape i w...

sense | qualis _ | specialType _

sense-pol zasuwa

sense-ger _

localisations

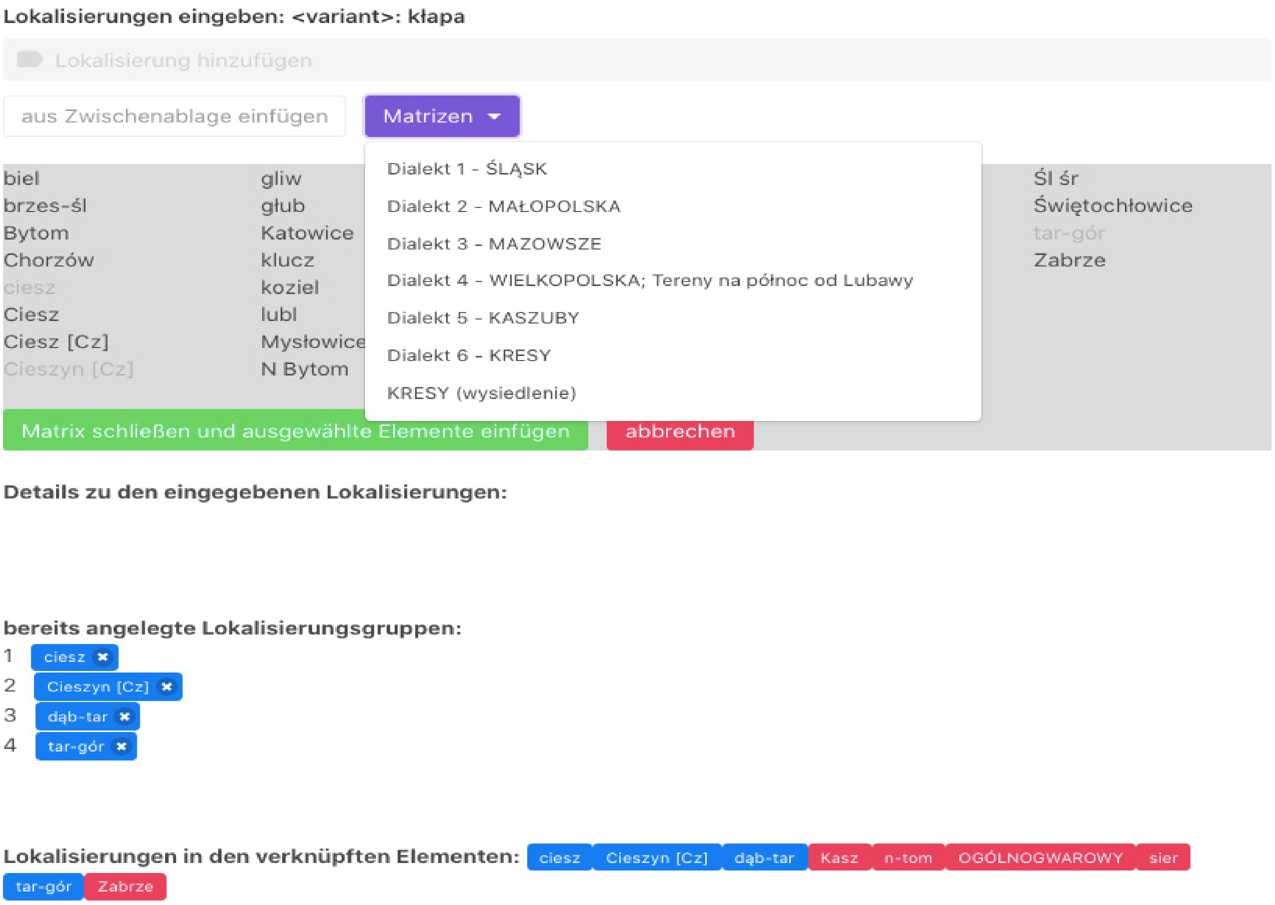
localisationGrp kęp 2.2.

localisation | name kęp | group 2.2.

documentations

localisationGrp oles 1.2.

Figure 1: Basic XML editor screen.



Lokalisierungen eingeben: <variant>: kłapa

Lokalisierung hinzufügen
aus Zwischenablage einfügen
Matrizen

Dialekt 1 - ŚLĄSK
Dialekt 2 - MAŁOPOLSKA
Dialekt 3 - MAZOWSZE
Dialekt 4 - WIELKOPOLSKA; Tereny na północ od Lubawy
Dialekt 5 - KASZUBY
Dialekt 6 - KRESY
KRESY (wysiedlenie)

biel
brzes-śl
Bytom
Chorzów
ciesz
Ciesz
Ciesz [Cz]
Cieszyn [Cz]

gliw
głub
Katowice
klucz
koziel
lubl
Mysłowice
N Bytom

Śl śr
Świętochłowice
tar-gór
Zabrze

Matrix schließen und ausgewählte Elemente einfügen
abbrechen

Details zu den eingegebenen Lokalisierungen:

bereits angelegte Lokalisierungsgruppen:

1 ciesz ✖
2 Cieszyn [Cz] ✖
3 dąb-tar ✖
4 tar-gór ✖

Lokalisierungen in den verknüpften Elementen: ciesz Cieszyn [Cz] dąb-tar Kasz n-tom OGÓLNOGWAROWY sier tar-gór Zabrze

Figure 2: Localisation editor.

5 Entry Microstructure

The estimated 5000 entries of the DGlPd group Polish dialectal words together with their word senses according to their common German etymon in a nest structure. The currently planned main lexicographical indications are enumerated below. A detailed lexicographic protocol will depend on the outcome of the compilation process for the 'proto-entries' as explained above in section 3.

- The German **etymon** is provided with grammatical information and relevant word senses. Frequently, multiple etyma, usually with close diasystemic ties to each other, must be given.
- All attested Polish **dialectal variants** for a given German etymon are listed, together with grammatical information. The depth of the differentiation of dialectal "phonic" – i.e. phonological or phonetic – expression variation will be the same as in the SGP. The variants will be offered in a normalized graphic form, again based on the rules applied in the SGP. Each expression variant will receive a rough classification as to its degree of deviation from the expression of the German etymon, namely (i) no or minor deviation, (ii) medium deviation, or (iii) strong deviation. An "approximation algorithm" will be used for the classification, conceptually similar to an (informal) variant of the Levenshtein distance and based on the number of phonological segments affected by the transfer of expression material, the number and the quality of the phonological processes that could be seen behind phonic substitutions and the fact that certain observed phonic substitutions cannot be explained by known types of phonological processes (cf. Stachowski 2011). Such a classification will at least have a heuristic value and will, as one of many search options, offer valuable information for linguistic research on mechanisms of material transfer of expressions.
- The entries will provide the **word senses** attested for the expression variants and map them to a semantic field classification roughly along the lines of Haspelmath/Tadmor 2009.
- **Primary derivatives** will be listed with their forms and word senses.
- Information will be provided on **equivalents in Eastern Yiddish**, taking advantage of lexicographic sources such as as Stutchkoff (1950) and Astravuch (2008).
- In a similar vein, the entries will provide information on **equivalents in (Old) Czech**, mainly based on Newerkla's comprehensive study on German loans in Czech and Slovak (Newerkla 2011).
- Each entry features a **textual commentary** with an assessment of the history of the loanwords.

Word variants, derivatives, and word senses are each assigned "localisations" (i.e. areas of dialectal attestation) which may range from whole dialectal and subdialectal groups down to the level of individual counties (*powiaty*).

For each word sense, representative citations for select localisations will be offered. It has to be taken into consideration that in contrast to dictionaries based on written texts, the fixation of a word in a concrete dialect or subdialect, i.e. in the corresponding entry of the card index, sometimes only exists in the form of "expression E (with meaning M) in village / region V".) In such cases no citations can be offered.

6 Online Presentation

The DGlPd online entry presentation will be fully bilingual, which means that even word sense definitions, paraphrases, textual commentary, etc. will be offered in both German and Polish.

The entries will feature interactive cartographic representations that visualize the available fine-grained localisations, aggregating the data on roughly the level of a traditional map of Polish dialect space, with the traditional dialect areas differentiated in greater detail based on the time periods of German (speaking) rule. There will be at least one global map for each entry, but depending on the degree of complexity of formal variation, polysemy and/or derived forms further maps on specific words or meanings may be necessary for reasons of transparent exposition.

All entries containing data from the as yet unpublished parts of the SGP will contain hyperlinks to digital images of the relevant cards – many of them hand-written – of the card SGP index.

The DGlPd will be integrated into the *Lehnwortportal Deutsch* (LWP; cf. Meyer 2013), an online publication platform for a growing number of dictionaries on German lexical borrowings in other languages. In the LWP, the lexicographical data is internally managed as a cross-dictionary network (technically, a graph database) of relations between word forms (etyma, loanwords, the corresponding meanings and expression variants, derivatives, etc.) of all included dictionaries. In a third-party project funded by the Fritz Thyssen Foundation, the LWP, which is hosted and maintained at the IDS, is currently undergoing a complete redesign with respect to underlying technologies and user interface design. The resulting system, publicly available in 2022, will be based on a refined, manually curated and autonomous graph-based abstraction layer on top of the 'native' dictionary data that allows users to find and navigate through lexical units and their relations to each other in real time in an interactive, cross-resource fashion (Meyer & Eppinger 2018). The revamped version will start with about ten newly added resources, most of them originally print dictionaries, covering many important European recipient languages such as English, French, Dutch, Czech, Slovak, and Hungarian. Even now, the LWP hosts digitally enhanced versions of two Polish-related resources, viz. the WDLP and the WDLT, soon to be complemented by an online version of a recently published frequency dictionary on German loans in the Silesian dialect of Polish (Hentschel, Tambor & Fekete 2021). Starting 2022, the most comprehensive study on German loans in Czech and Slovak (Newerkla 2011) will be part of the system. Somewhat later, a specialised dictionary on parallel lexical borrowings from German in Polish and in the East Slavic Languages (Belarusian, Ukrainian, Russian) will be available in the LWP (cf. Meyer 2015 for an early presentation). Similar to the DGlPd, this dictionary is based on a cooperation between the Institute of Slavic Studies at the University of Oldenburg and the Leibniz Institute for the German Language.

Several of the resources named here (WDLP; WDLT; Newerkla 2011) are systematically referenced in the DGlPd entries;

together with the mostly etymon-based cross-resource links in the LWP this will embed the DGlPd data in a large network of Polish-related information concerning Literary and dialectal Polish and important contact languages.

The newly designed search options will allow users to find information based both on general, cross-dictionary criteria and on dictionary-specific criteria. For the DGlPd, the latter will include the possibility to query the fine-grained localisation data as well as the highly specialised classifications, named in section 5, for dialectal variants and word senses. Users will be able to formulate arbitrarily complex queries (including, of course, Boolean operators, regular expressions etc., but also descriptions of multiple-word borrowing constellations) through a visual query builder (Meyer 2019) integrated seamlessly into the main page.

7 Conclusion

The ongoing lexicographical endeavour outlined in this paper aims to present the linguistic results of longstanding lexical borrowing processes from German into dialectal Polish almost literally as a complex landscape – to be digitally explored by experts and interested laypersons alike. Besides a more traditional, linear textual dictionary entry format, users are offered cartographic visualizations of the dialectal distribution of loanwords and their meanings. A dense network of cross-references to other resources on German loans in Polish and other, mainly European, languages, can be leveraged in advanced search queries that can take into account the relations between lexical units available in the LWP's graph database. Users may even get a glimpse of the lexicographical practice underlying the DGlPd's main source, looking at images of index cards collected during decades and now used for a publication the first time.

The DGlPd can be seen as a further step toward a full representation of lexical traces of German cultural and linguistic contact in Eastern Europe. Through its integration in a platform of interlinked multilingual resources it will remain open towards future lexicographical, dialectal, and historical research.

8 References

Astravuch, A. (2008). *Idyš-belaruski sloǔnik.* Minsk: Medysont.

Bednarczuk, L. (1986-1988) Języki indoeuropejskie. Vol. I / II. Warszawa: PWN.

Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. In: M. Haspelmath, U. Tadmor (eds.) (2009), pp. 35-53.

Haspelmath, M., Tadmor, U. (eds.) (2009). Loanwords in the world's languages: A comparative handbook. Berlin: Mouton de Gruyter.

Hentschel, G. (2009). Intensität und Extensität deutsch-polnischer Sprachkontakte von den mittelalterlichen Anfängen bis ins 20. Jahrhundert am Beispiel deutscher Lehnwörter im Polnischen. In C. Stolz (eds.) *Unsere sprachlichen Nachbarn in Europa. Die Kontaktbeziehungen zwischen Deutsch und seinen Grenznachbarn*. Bochum: Brockmeyer, pp. 155-171.

Hentschel, G. (2010). *Zur Einführung [Introduction to the WDLP] online.* Accessed at: http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp/46704.html [15/04/2021].

Hentschel, G., Tambor, J. & Fekete, I. (2021). *Frequenzwörterbuch deutscher Lehnwörter im Schlesischen der Gegenwart. Mit Kommentaren zur Etymologie*. Oldenburg: BIS-Verlag.

Měchura, M. B. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19- 21 September 2017, Leiden, The Netherlands.* Leiden, The Netherlands. Brno, pp. 662-679.

Meyer, P. (2013). Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation. In: S. Kempgen, M. Wingender, N. Franz, M. Jakiša (eds.): *Deutsche Beiträge zum 15. Internationalen Slavistenkongress*. Minsk, München: Otto Sagner, pp. 233-242.

Meyer, P. (2015). Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography. In K. Kallas, I. Kosem, S. Krek (eds*.) Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton*. Trojina: Institute for Applied Slovene Studies, pp. 198–210.

Meyer, P., Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej et al. (eds.) *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17-21 July*, *Ljubljana.* Ljubljana: University Press, pp. 1017-1022.

Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Bilder für lexikografische Property-Graphen. In P. Sahle (eds.): *Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), Frankfurt am Main, Mainz, 25.3.2019 – 29.3.2019. Konferenzabstracts.* Frankfurt a.M., pp. 312-314.

Muysken, P. (2000). *Bilingual speech. A typology of code-mixing*. Cambridge: Cambridge University Press.

Newerkla, S. (2011). *Sprachkontakte Deutsch-Tschechisch-Slowakisch*. Frankfurt a. M., Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Poplack, S., Dion, N. (2012). Myths and facts about loanword development. In *Language Variation and Change* 24(3), pp. 279-315.

Simões, A., Salgado, A., Costa, R., & Almeida, J.J. (2019). LeXmart: A Smart Tool for Lexicographers. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Sintra, Portugal. Brno, pp. 453-466.

Stachowski, K. (2011). A note on Levenshtein distance versus human analysis. In *Studia linguistica Universitatis Iagellonicae Cracoviensis 128.* Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, pp. 155-160.

Stutchkoff, N. (1950). *Der Oytser Fun Der Yiddisher Sprakh.* Accessed at:

http://www.cs.uky.edu/~raphael/yiddish/searchOytser.cgi [15/04/2021]

Walsleben, A. (1997). *Romanische Lehnwörter in polnischen Texten des 17. Jahrhunderts*. München: Kubon & Sagner.

LWP = *Lehnwortportal Deutsch*, ed. Leibniz Institute for the German Language. Accessed at: http://lwp.ids-mannheim.de [15/04/2021].

SALDAS = *Sprachwissenschaftliche Arbeiten zu Lehnwörtern aus dem Deutschen in anderen Sprachen*, ed. Leibniz Institute for the German Language. Accessed at: http://lwp.ids-mannheim.de/saldas [15/04/2021]

SGP = *Słownik gwar polskich* (*A – hepnąć),* Sources and vol. I, ed. M. Karaś, J. Reichan; vols. II-V, ed. J. Reichan, S. Urbańczyk; vol. VI, ed. J. Okoniowa, J. Reichan; vols. VII-VIII, ed. J. Okoniowa, J. Reichan, B. Grabka, vol. IX, ed. B. Grabka, R. Kucharzyk, J. Okoniowa, J. Reichan. Issue 1, Wrocław-Warszawa-Kraków-Gdańsk; since issue 2, Łódź; vol. III, Wrocław-Warszawa-Kraków; since vol. IV – Kraków, 1977-2017.

USTAWA = *Ustawa o mniejszościach narodowych i etnicznych oraz o języku regionalnym* z dnia 6 stycznia 2005 r. Accessed at: http://mniejszosci.narodowe.mswia.gov.pl/mne/prawo/zapisy-z-konstytucji-r/6447,Ustawa-o-mniejszosciach-narodowych-i-etnicznych-oraz-o-jezyku-regionalnym.html [14/06/2021].

WDLP = *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts*, ed. A. Vincenz and G. Hentschel. Accessed at*:* *http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp* [15/04/2021].

WDLT = *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen*, ed. Th. Menzel and G. Hentschel, second, amended and enlarged electronic edition 2005. Accessed at: https://www.bkge.de/Publikationen/Online/Woerterbuecher/Deutsche_Lehnwoerter_im_Teschener_Dialekt [07/04/2021].

Acknowledgements

The project described in this paper is funded by the German Research Foundation (HE 1566/16-1; ME 4968/2-1).

EURALEX XIX
Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Historical and Scholarly Lexicography and Etymology

Drawing the line between synchrony and diachrony in historical and dialectal lexicography

Manolessou I., Katsouda G.

Academy of Athens, Greece
manolessou@academyofathens.gr
katsouda@academyofathens.gr

Abstract

The purpose of the article is to discuss the interaction between synchrony and diachrony in the domain of historical and dialectal lexicography. The discussion is organized on the basis of the various components/"information slots" of a dictionary entry, and more specifically: a) the headword or lemma form (selection of a form belonging to a specific synchrony vs. creating an artificial 'a-chronic' form), b) the formal section, where the variant forms of the word are listed (belonging or not to the 'same' synchrony, presented or not in 'chronological' order), c) the etymological section, where the origin and the morphological analysis of the word is given (by definition the locus of diachronic presentation), and d) the semantic section, where the various senses of the word are listed (again, belonging or not to the same synchrony, and presented or not in chronological order). The discussion is based principally on the *Historical Dictionary of Modern Greek* (ILNE) of the Academy of Athens, the largest on-going lexicographic project in Greece.

Keywords: historical lexicography; dialectal lexicography; synchrony; diachrony, morphology

1 Introduction

A historical dictionary is by definition the *par excellence* type of lexicographic work which is based on the notion of diachrony. Nevertheless, it is in several cases necessary to view its subject-matter in a synchronic way (or more specifically as a present synchrony being investigated regressively towards the past; on the notion of "gegenwartsbezogene historische Lexikographie" see Reichmann (2012: 19-21).[1] Conversely, dialectal dictionaries by definition deal with synchronic data (e.g., *in situ* fieldwork with native speakers of extant dialects). However, it is almost always the case that this data must also be viewed diachronically (especially in the case of obsolescent dialects with terminal speakers). The present paper aims to discuss this interaction between synchrony and diachrony, on the basis of specific historical and dialectal lexicographic works on the Greek language, and principally the *Historical Dictionary of Modern Greek* of the Academy of Athens.

The *Historical Dictionary of Modern Greek, both of the Standard Language and the dialects* (Ἱστορικὸν Λεξικὸν τῆς Νέας Ἑλληνικῆς, τῆς τε κοινῶς ὁμιλουμένης καὶ τῶν ἰδιωμάτων), or ILNE for short, is the national lexicographic enterprise of Greece, published by the Academy of Athens over a period of several decades (Manolessou & Bassea-Bezantakou 2013, Manolessou & Katsouda forthcoming b). The first volume appeared in 1933, and since then 7 volumes have been published intermittently, with a several years' gap between each volume, due to consecutive changes in the institution's research and publishing policies. The last volume (up to the entry *δόγης* [ˈðojis] 'doge') was published in 2021. Currently, work is under way for the publication of the volume 7b, which will end the treatment of the letter Δ (delta), and is due to appear in the coming year.

The *ILNE* is primarily a "classic" historical dictionary, conceived as a plan in the early 20th c., by the founder of the discipline of linguistics in Greece, Georgios Chatzidakis. As such, its content and entries include the standard components identified in the typology of historical dictionaries (Reichmann 1990: 1594): a headword, variant forms, part-of-speech and inflectional information, etymology and word history, frequency, collocations, examples from everyday language, quotations from press and literature, synonyms/antonyms and bibliography. So, in many respects it resembles "traditional" historical dictionaries such as the *Oxford English Dictionary* (*OED)*, *Wörterbuch der bairischen Mundarten in Österreich* (*WBÖ*), the *Deutsches Wörterbuch* (*DWB)*, the *Woordenboek der Nederlandsche Taal (WNT)* and the *Svenska Akademiens Ordbokhttp (SAOB)*, which are all long-term multi-volume enterprises, aiming to document the national language in all its detail and in its diachronic dimension.

Despite its categorization as a historical dictionary, the ILNE differs substantially from other national lexicographical enterprises in one crucial respect: as its title reflects, it covers not only the standard language, but all its dialects as well. This objective multiplies considerably the size of the corpus and consequently lengthens, even more considerably, the time required for the completion of the project. However, it was and remains a necessary decision, rendered inevitable by the special sociolinguistic circumstances of Greek linguistic history.

Modern Greek has a great number of dialects, all deriving from the Hellenistic Koine, which in turn descends from the

[1] In a similar vein, current approaches to etymology in historical dictionaries reject the traditional approach of "forward" presentation of a word's history, in favour of a "regressive" investigation starting from the present form (Petrequin & Andronache 2008: 1167-1168).

Attic dialect of Ancient Greek. There is no national dictionary covering all the dialects (i.e., the equivalent of the *English Dialect Dictionary* (*EDD*)). This renders historical research, both on a formal and a semantic/lexicographical level difficult, since by definition diachronic linguistics requires comparative data. But quite apart from this problem, the main reason behind the necessity of including dialectal data in the historical dictionary of Modern Greek lies in the dual nature of linguistic history of Greek. Since the first centuries AD, the oral and the written tradition of the language had begun to diverge, eventually leading to a state of diglossia: a contrast between a high-level (conservative, written, official) and a low-level (vernacular) language variety, which would last for more than 2000 years. In the relevant literature on the topic (e.g., Ferguson 1959), Greek is in fact frequently considered as a paradigm case of diglossia.

For linguistic research, again, reliance on the high-level variety is not conducive to reliable results concerning diachronic evolution (mechanisms, causes and patterns of change). But at the time of the *Historical Dictionary*'s inception, the "national language" was just such a high-register, learned, archaic, and to a certain extent "artificial", diglossic variety, something which was evident on levels of analysis: phonology, morphology, syntax and vocabulary. Therefore, for the reliable and in-depth investigation of the history of Greek, recourse to the dialects was inevitable. For example, it frequently happens that a phonetic change appears only sporadically in Standard Modern Greek, under the influence of the learned high register language, whereas the examination of dialectal data might reveal the great degree of regularity than one would normally expect from a phonetic change. After the resolution of the so-called 'language question' in the last decades of the 20th c. (Mackridge 2007), the Standard Modern Greek language which emerged was based on the "low", "vernacular" variety, but had admitted in its structure

In the case of Modern Greek, therefore, the 'admixture' of synchrony and diachrony is an element of the structure of the language itself, on all levels (phonology, morphology, syntax and especially vocabulary – see e.g., the list of 'learned' elements in Standard Modern Greek given in Anastasiadi-Symeonidi & Fliatouras 2019). Up to a point, consequently, the lexicographic treatment even in general dictionaries of Modern Greek only entails the penetration of diachrony in all slots of the dictionary entry (forms, etymology, senses). When one attempts to treat the dialects as well, the diachronic admixture increases as dialects preserve many archaic features no longer surviving in the Standard.[2]

In what follows, we shall discuss in more detail the interaction of synchrony and diachrony in the *Historical Dictionary of Modern Greek*, with examples taken from the two more recent volumes, 6 and 7a.

2 Headword/ Lemma Form

In a dictionary of the standard language, the selection of headword poses no problems: it is the 'quotation form' of the (most common variant) of the standard language in its current synchronic stage. But both in historical and in dialectal lexicography, such a selection is far from obvious, due to the great number and potentially great divergence of the variant forms than need to be subsumed under the same heading. The result is often a form which 'mixes' synchrony and diachrony. More specifically, the solutions which may be adopted in headword selection may be the following (Katsouda 2012: 124-127):

- The oldest – most conservative variant, from which all other variants may be derived through diachronic processes of phonological or morphological change. In this case, the irruption of diachrony into synchrony is obvious, since the headword belongs to an earlier stage than many of the variant forms.
- An a-chronic artificial variant, which subsumes all 'real' variants. This is for example the solution adopted for the Flemish dialect dictionary (Rys & Keymeulen 2009), as well as in several other dialectal dictionaries.
- The Standard form, without regarding the stage of its diachronic evolution

In the case of the *ILNE*, the headword is selected on the basis of both synchronic and diachronic criteria, as described in detail in its Manual of Regulations (*ILNE- MR*):

-If the word belongs only to Standard Modern Greek, no difficulty arises when it presents only a single variant. For example, in the case of relatively high-register words like *δεινοπαθῶ* [ðinopaˈθo] 'suffer' or *δεινόσαυρος* [ðiˈnosavros] 'dinosaur', the only attested form is the one given as headword.

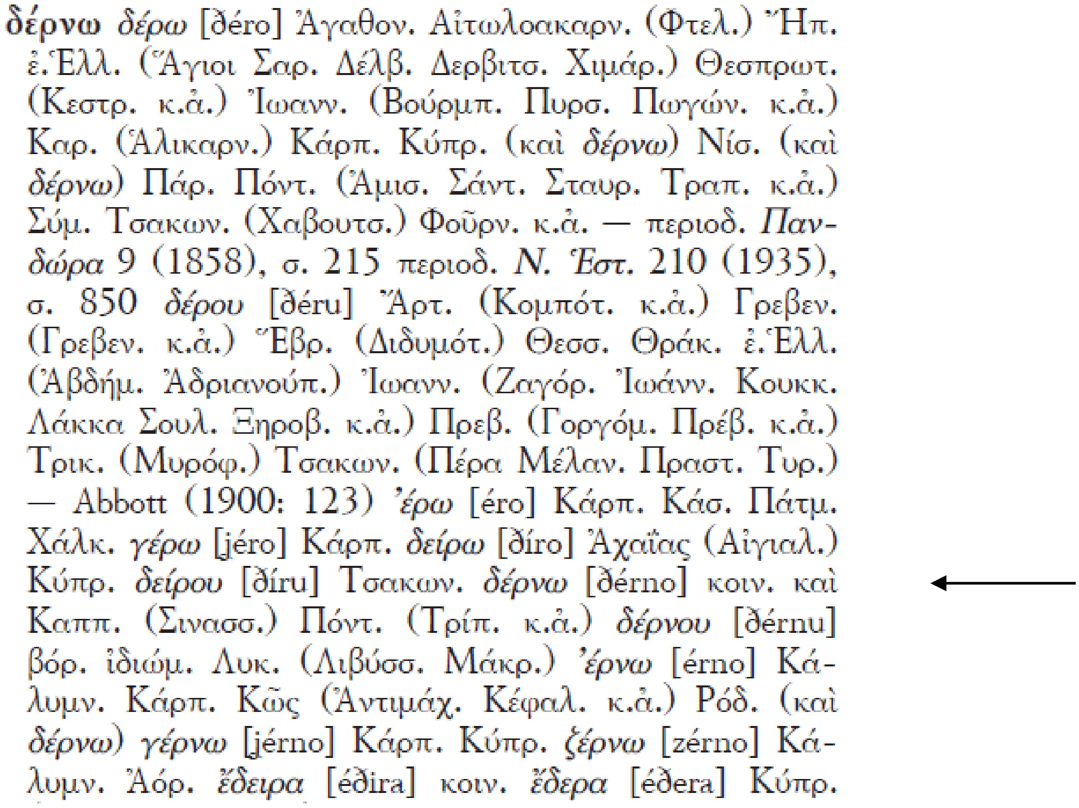
-When a Standard word, however, presents two or more, usually phonological, variants, a diachronic criterion enters into the picture: the headword takes the form which is closer to the original etymon of the word, i.e., the diachronically older form. For example, given the two variants *δεκαοκτώ* [ðekaoˈkto] and *δεκαοχτώ* [ðekaoˈxto] 'eighteen', the headword will assume the more conservative form *δεκαοκτώ* [ðekaoˈkto], which does not display the manner dissimilation of consecutive stop consonants.[3] Sometimes of course the language presents two variants which do differ phonologically, but are not distinguished graphematically, since the spelling system does not make sub-phonemic distinctions. As two typical examples one may mention *δεκαπέντε* [ðekaˈpende] / [ðekaˈpede] 'fifteen', where the spelling system cannot differentiate between a pre-nasalized and a non-presalized realization of voiced stops, or *διάφανος* [ðiˈafanos] / [ˈðjafanos] 'transparent', where the spelling system does not distinguish between a monosyllabic and a bisyllabic realization of high+low vowel sequences (with vs. without synizesis).[4]

[2] Here the term 'archaism' is employed in a specialized dialectological meaning, not as an 'an old word or phrase no longer in general spoken or written use' to be found in the Standard language (as per the definition in Crystal 2008, s.v.), but as 'phonological, morphological or lexical features of earlier linguistic phases, not surviving in the Standard language, but present in the dialects', as employed e.g., in Andriotis (1974) and Tzitzilis (2013), and defined in the *ILNE-MR*.

[3] On this phonetic change, datable to the medieval period and appearing with variable regularity both in Standard Modern Greek and the Modern Greek dialects see Newton (1972: 106-112), Holton et al. (2019: 185-193) and references therein.

[4] For this landmark change in the history of Greek, which has the value of a distinctive feature for learned/high-register vocabulary as

-When a word belongs both to the Standard and to the dialects, the criterion for headword selection is again primary synchronic, since the standard form is always preferred for reasons of findability, even though the dialects may preserve several variant forms which are closer to the original etymon. For example, for the verb ‘to beat, thrash’ *δέρνω* [ˈðerno] with an added nasal formant is the headword form, despite the fact that many dialects preserve the original Ancient Greek form *δέρω* [ˈðero] < [déro:]. Similarly, for the verb ‘to give’, *δίνω* [ˈðino] is the headword form, despite the dialectal attestation of more archaic variants like *δίδω* [ˈðiðo], closer to AG *δίδωμι* [dído:mi]. In such cases, the diachronic precedence of the the dialectal variants is expressed through their relative ordering in the formal section of the dictionary entry: As discussed below in (3), variant forms are listed in chronological order in this section of the dictionary, which entails that the conservative variant forms will appear in the list before the Standard form; see fig. 1 for an exemplification of this practice from the above-mentioned entries of the ILNE.



δέρνω *δέρω* [ðéro] Ἀγαθον. Αἰτωλοακαρν. (Φτελ.) Ἤπ. ἐ.Ἑλλ. (Ἅγιοι Σαρ. Δέλβ. Δερβιτσ. Χιμάρ.) Θεσπρωτ. (Κεστρ. κ.ἀ.) Ἰωανν. (Βούρμπ. Πυρσ. Πωγών. κ.ἀ.) Καρ. (Ἁλικαρν.) Κάρπ. Κύπρ. (καὶ *δέρνω*) Νίσ. (καὶ *δέρνω*) Πάρ. Πόντ. (Ἀμισ. Σάντ. Σταυρ. Τραπ. κ.ἀ.) Σύμ. Τσακων. (Χαβουτσ.) Φοῦρν. κ.ἀ. — περιοδ. *Πανδώρα* 9 (1858), σ. 215 περιοδ. *Ν. Ἑστ.* 210 (1935), σ. 850 *δέρου* [ðéru] Ἄρτ. (Κομπότ. κ.ἀ.) Γρεβεν. (Γρεβεν. κ.ἀ.) Ἕβρ. (Διδυμότ.) Θεσσ. Θράκ. ἐ.Ἑλλ. (Ἀβδήμ. Ἀδριανούπ.) Ἰωανν. (Ζαγόρ. Ἰωάνν. Κουκκ. Λάκκα Σουλ. Ξηροβ. κ.ἀ.) Πρεβ. (Γοργόμ. Πρέβ. κ.ἀ.) Τρικ. (Μυρόφ.) Τσακων. (Πέρα Μέλαν. Πραστ. Τυρ.) — Abbott (1900: 123) *’έρω* [éro] Κάρπ. Κάσ. Πάτμ. Χάλκ. *γέρω* [ȷéro] Κάρπ. *δείρω* [ðíro] Ἀχαΐας (Αἰγιαλ.) Κύπρ. *δείρου* [ðíru] Τσακων. *δέρνω* [ðérno] κοιν. καὶ Καππ. (Σινασσ.) Πόντ. (Τρίπ. κ.ἀ.) *δέρνου* [ðérnu] βόρ. ἰδιώμ. Λυκ. (Λιβύσσ. Μάκρ.) *’έρνω* [érno] Κάλυμν. Κάρπ. Κῶς (Ἀντιμάχ. Κέφαλ. κ.ἀ.) Ῥόδ. (καὶ *δέρνω*) *γέρνω* [ȷérno] Κάρπ. Κύπρ. *ζέρνω* [zérno] Κάλυμν. Ἀόρ. *ἔδειρα* [éðira] κοιν. *ἔδερα* [éðera] Κύπρ.

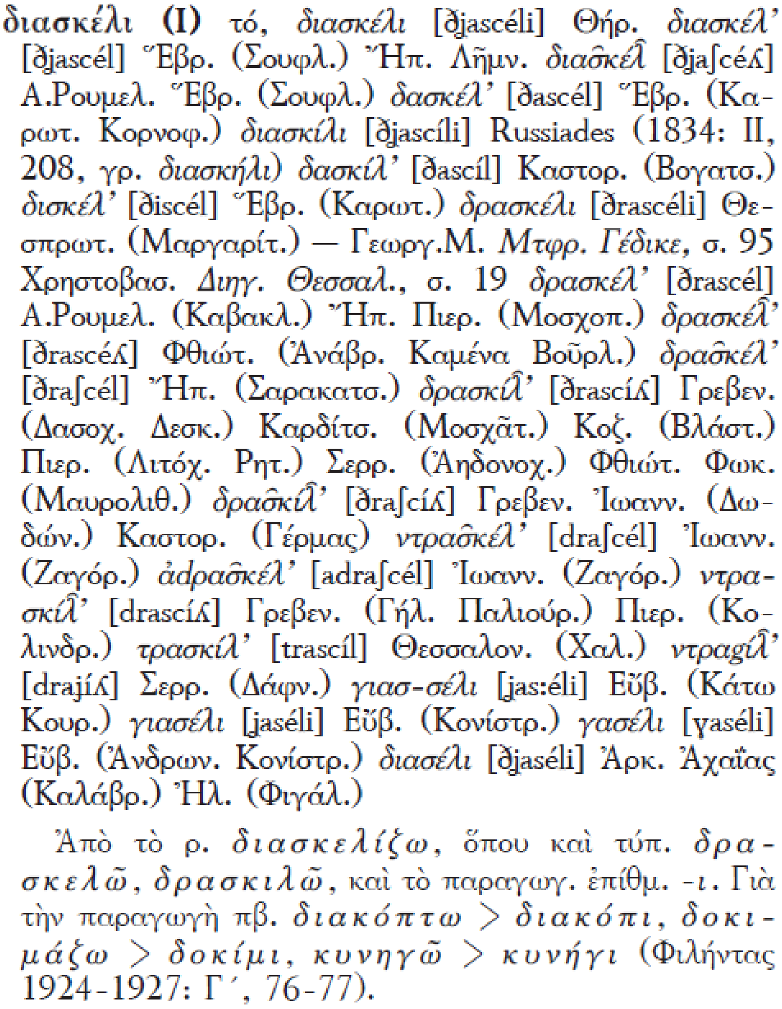
Figure 1: The formal section of the entry **δέρνω** [ˈðerno] ‘to beat up’ (ILNE, vol. 6).

-When a word is not part of the standard language, but is only to be found in dialects, the primary criterion, based on findability again, is that of frequency: the most widespread variant becomes the headword. A characteristic instance of this principle occurs in the case of loanwords whose original etymon is a word starting with [d], but which are adapted into Greek in a form with initial [ð], which becomes more widespread than the (also extant) [d]-initial variant. One may mention in this conjunction examples like *δεπουτάτος* [ðepuˈtatos] vs. *ντεπουτάτος* [depuˈtatos] ‘church official’ (< Latin *deputatus* or Italian *deputato*), *δερβίσαγας* [ðerˈvisaγas] vs. *ντερβίσαγας* [derˈvisaγas] (< Turk. *derviş ağa*) ‘Ottoman military or religious official’, *δεσένιο* [ðeˈsе́ɲo] vs. *ντεσένιο* [deˈseɲo] (< Venetian *dessegno*) ‘blueprint’. To give another example, the dialectal word *διαρμίζω* [ðjarˈmizo] ‘to tidy’ appears in this form in most island dialects (Cyclades, Dodecanese, Crete), and so this variant has been chosen as headword, despite the existence of the form *διορμίζω* [ðjorˈmizo] from the island of Kasos, which is closer to the medieval etymon *διορμῶ* [ðiorˈmo] attested in the Lexicon of Photius. In this case, the preservation of an earlier variant in a single sub-dialectal variety does not weigh sufficiently against the wide distribution of a diachronically more “altered” variant attested throughout three dialect groups.

-When no form has quantitative precedence over the others, i.e., when each geographic area presents a different variant, then diachrony enters into the picture again, and the form closest to the original etymon becomes the headword. To give a simple example, when having to choose between the two forms *διπλόφουχτα* [ðiˈplofuxta] ‘quantity which can fit into the two palms’and *διπλόχουφτα* [ðiˈploxufta] with consonant metathesis, both of which have roughly equal distribution, the first form takes precedence due to its relative closeness to the original etymon *φούχτα* [ˈfuxta] < AG *πυγμή* [pugmɛ́:]. As a more composite instance, consider the entry for the dialectal noun *διασκέλι* [ðjaˈsceli] ‘step, stride’, as depicted in fig. 2 below: although the specific form [ðjaˈsceli] is attested only in a single area (the island of Thera/Santorini), it is selected as headword over the more than 20 other variants (such as [ðraˈsceli], [ðjaˈseli], [draˈʃcel], [traˈscil], [draˈʝiʎ] etc.) some of which are more widely attested. This primarily solves the problem of having to measure quantitatively the distribution of each variant (an impossible task given the chronological spread and the uneven nature of the dictionary’s sources, as

well as for the delimitation of dialectal isoglosses see Newton (1972: 30-41), Holton et al. (2019: 98-109) and references therein.

described below in section 3) but also guarantees the findability of the entry: when appearing as headword, the more conservative variant, being closer to the original etymon (the verb *διασκελίζω* [ðiaskeˈlizo] 'to stride, to step' with back-formation), ensures that the entry will appear, in the alphabetic ordering of the dictionary, adjacently to the etymon, and together with all other entries which make up the whole word-family.



διασκέλι (I) τό, *διασκέλι* [ðjascéli] Θήρ. *διασκέλ’* [ðjascél] Ἔβρ. (Σουφλ.) Ἤπ. Λῆμν. *διασ̂κέλ̂* [ðjaʃcéʎ] Α.Ρουμελ. Ἕβρ. (Σουφλ.) *δασκέλ’* [ðascél] Ἕβρ. (Καρωτ. Κορνοφ.) *διασκίλι* [ðjascíli] Russiades (1834: II, 208, γρ. *διασκήλι*) *δασκίλ’* [ðascíl] Καστορ. (Βογατσ.) *δισκέλ’* [ðiscél] Ἕβρ. (Καρωτ.) *δρασκέλι* [ðrascéli] Θεσπρωτ. (Μαργαρίτ.) — Γεωργ.Μ. *Μτφρ. Γέδικε*, σ. 95 Χρηστοβασ. *Διηγ. Θεσσαλ.*, σ. 19 *δρασκέλ’* [ðrascél] Α.Ρουμελ. (Καβακλ.) Ἤπ. Πιερ. (Μοσχοπ.) *δρασκέλ̂’* [ðrascéʎ] Φθιώτ. (Ἀνάβρ. Καμένα Βοῦρλ.) *δρασ̂κέλ’* [ðraʃcél] Ἤπ. (Σαρακατσ.) *δρασκίλ̂’* [ðrascíʎ] Γρεβεν. (Δασοχ. Δεσκ.) Καρδίτσ. (Μοσχᾶτ.) Κοζ. (Βλάστ.) Πιερ. (Λιτόχ. Ρητ.) Σερρ. (Ἀηδονοχ.) Φθιώτ. Φωκ. (Μαυρολιθ.) *δρασ̂κίλ̂’* [ðraʃciʎ] Γρεβεν. Ἰωανν. (Δωδών.) Καστορ. (Γέρμας) *ντρασ̂κέλ’* [draʃcél] Ἰωανν. (Ζαγόρ.) *ἀδρασ̂κέλ’* [adraʃcél] Ἰωανν. (Ζαγόρ.) *ντρασκίλ̂’* [drascíʎ] Γρεβεν. (Γήλ. Παλιούρ.) Πιερ. (Κολινδρ.) *τρασκίλ’* [trascíl] Θεσσαλον. (Χαλ.) *ντραgίλ̂’* [drajíʎ] Σερρ. (Δάφν.) *γιασ-σέλι* [jas:éli] Εὔβ. (Κάτω Κουρ.) *γιασέλι* [jaséli] Εὔβ. (Κονίστρ.) *γασέλι* [ɣaséli] Εὔβ. (Ἀνδρων. Κονίστρ.) *διασέλι* [ðjaséli] Ἀρκ. Ἀχαΐας (Καλάβρ.) Ἠλ. (Φιγάλ.)

Ἀπὸ τὸ ρ. *διασκελίζω*, ὅπου καὶ τύπ. *δρασκελῶ*, *δρασκιλῶ*, καὶ τὸ παραγωγ. ἐπίθμ. *-ι*. Γιὰ τὴν παραγωγὴ πβ. *διακόπτω* > *διακόπι*, *δοκιμάζω* > *δοκίμι*, *κυνηγῶ* > *κυνήγι* (Φιλήντας 1924-1927: Γ´, 76-77).

Figure 2: The formal and etymological section of the entry **διασκέλι** [ðjaˈsceli] 'stride, step' (ILNE, vol. 7a).

3 Formal Section

It is a well-known issue in dialectal lexicography that the listing of variant forms may conceal an admixture of synchrony and diachrony, since these forms are drawn from a variety of sources, not all of which belong to the same synchrony (Katsouda 2016). In the case of Greece, this is exacerbated by the abrupt and large-scale changes in the geographical spread and demographic composition of the Greek-speaking world during the 20th c. due to major political events (wars resulting in border expansion, exchange of populations etc.). As a result, Greek dialectal lexicography functions, in any case, with the tacit assumption that the dialectal picture it describes is not truly "synchronic", but rather represents a past synchrony of the late 19th-early 20th c. (cf. Trudgill 2003: 48).

As a result, the *ILNE* in fact treats sources which cover a period of roughly 150 years as belonging to the same synchrony: the oldest written fieldwork recordings come from the mid-19th c., while a large of material also predates the war of 1922 which resulted in the massive population exchange with Turkey and in the relocation of all the Greek (dialectal) speakers of Asia Minor in mainland Greece. Furthermore, data collection through fieldwork continued uninterrupted throughout the 20th c from all Greek-speaking areas. To take a single example, for the island of Sifnos, there are 10 manuscript collections in the *ILNE* archive, the earliest dating from 1912 and the latest from 2017. This long process of consecutive documentation allows the researcher to acquire a picture of local variants decade-by-decade, and thus to observe their gradual processes of change or (usually) attrition and obsolescence. This is occasionally reflected in the formal section of *ILNE* entries, in the case when two variant forms from the same location are listed (Manolessou 2012, 28-29). As an instance one may mention the case of dialectal forms from the South-Eastern dialect area (Dodecanese, Cyprus), where, due to the phenomenon of voiced fricative deletion and subsequent hypercorrect restitution,[5] many lexical items from the same island present simultaneously three variant forms: one with the original voiced fricative intact, one with deletion of the fricative, and one with the "wrong" fricative restituted. Examples include *διπλός* [diˈplos] – *'ιπλός* [iˈplos] – *γιπλός* [ʝiˈplos] 'double' or *δικάντζω* [ðiˈkandzo] – *'ικάντζω* [iˈkandzo] – *γικάντζω* [ʝiˈkandzo] 'to judge' (Karpathos; entry δικάζω). In Fig. 1 above, one may see this variation as presented in the case of the verb δέρνω [ˈðerno]: the dictionary entry lists attestations for both [ˈðerno] and [ˈerno] from the island of Rhodes, for both [ˈerno] and [ˈʝerno] from the island of Karpathos, and for both [ˈerno] and [ˈzerno] (< *[ˈʑerno] < *[ˈʝerno] with additional palatalization) from the island of Kalymnos.

However, a diachronic explanation for synchronic variation of forms is not the only possibility; variation may be due to factors like the informant's code-switching between a standard and a dialectal form, the existence of intra-dialectal differentiation/microvariation within the same variety, or to sociolingustic causes such as sex, age or class differentiation (Manolessou, Beis & Bassea-Bezantakou 2012: 182). As has already crucially been observed in the case of the Romance languages (Banniard 2002: 782):

"Des prononciations et des réalisations distinctes pouvant cohabiter sur une même aire dialectale, il est imprudent d'étirer mécaniquement des successions de changements en leur attribuant des indices générationnels. Les phénomènes de tuilage et de chevauchement ont autant de probabilité de s'être produits en diachronie qu' en synchronie."

The admixture of synchrony and diachrony in the formal section of a historical dictionary is therefore unavoidable, since reasons of lexicographic economy do not permit the interpretation of the observed and recorded variation. The variation is always presented, for reasons of lexicographic economy, as a diachronic phenomenon, i.e. the variant forms are listed in the formal section in "chronological" and not in geographical order (depending on the number of phonetic and morphological changes each variant presents with respect to the original etymon).[6] Nevertheless, the dictionary user needs to be aware that if a certain geographic area presents variation between earlier and later forms, this may be due to more factors than the gradual nature of the spread of linguistic changes. It is left upon the dictionary user to draw their own conclusions concerning the synchronic or diachronic causes of variation of the described data, something which can only be achieved if the dictionary provides full documentation for the data presented, so that recoverability of information on the basis of the dictionary archive can be ensured (on the necessity of recoverability of dialectal data in the *ILNE* see Katsouda 2016: 155-156).

4 Etymological Section

4.1 Diachrony-in-Synchrony in Morphological Analysis

The etymological section of standard, historical and dialectal dictionaries is of course the main locus of presentation of diachronic data. However, it may also be a locus where synchronic analysis of lexical items also takes place, especially when a morphologically complex (derived/compound) word needs to be segmented into its composing parts. In such cases, it is frequently possible to assign alternative analyses, depending on whether one wishes to assume a synchronic or diachronic viewpoint, both with competing claims to "reality": the first on the actual, active, linguistic capacity of the native speaker and the second on the non-falsifiable, passive, record of the written text (Manolessou 2012). Nevertheless, an absolute distinction is often difficult to draw, as the relevant data are frequently lacking (e.g. the dating of the creation of an innovative suffix, the productivity of an affix during a certain chronological period or in a certain dialect, or, similarly, the productivity of a morphological mechanism such as backformation or conversion in a certain period or dialect).[7] Furthermore, the segmentation decision may depend on the overall morphological system one assumes for a specific time period or a specific dialectal variety, although it is hardly possible for any lexicographical enterprise to have developed a fully-fledged morphological model for the linguistic varieties it treats.

This difficulty in the synchronic vs. diachronic etymological treatment of morphologically complex dictionary entries is in fact a quite well-known issue in theoretical morphological discussions, and usually takes the form of problems in the synchronic analysis of a series of words which were derived via affixation with no longer extant derivational affixes or via no longer productive inflectional or derivational processes. Typical English examples include:

- The suffix *-ful* normally attaches to nominal bases in order to form adjectives, e.g.: *care → careful, beauty → beautiful, shame → shameful, thought → thoughtful*, etc. However, there are a number of cases where the suffix *-ful* synchronically seems to attach to verbal bases, as an exception to the rule: *forget → forgetful, resent → resentful, mourn → mournful*. Viewed from a diachronic viewpoint, the problem disappears, since at the time of formation of the relevant lexical items the English vocabulary did contain corresponding nominal forms which served as the derivational basis (example from Ruszkiewicz 1997: 96-100).
- Determinative compounds denoting actions are formed with the suffix *-er* on the basis of apparent "verb + object" combinations, e.g., *pay tax: taxpayer*, *own land: landowner*, *maw lawn: lawnmower*. However, there are instances where no such combination exists, such **say sooth*: *soothsayer*. The problem arises "as productive derivation is confused with historical derivation", since the noun *sooth* 'truth' did exist at the time of formation of the compound (example from Pilch 1985: 409-410).

[5] For a description of this phenomenon see Newton (1972, 60-73) and Holton et al. (2019, 153-154).

[6] On this methodological issue, and on the difference between assumed chronological precedence based on comparative reconstruction vs. 'real' chronological precedence based on historical attestation/documentation see Manolessou (2012: 54-58).

[7] For such processes in the diachrony of Greek see Manolessou & Ralli (2015).

4.2 Greek Lexicographic Examples

It is indeed the case that the Modern Greek vocabulary contains a number of morphologically complex lexical items, the result of non-apparent morphological processes or derived on the basis of no longer extant or productive affixes; these problematic cases receive variable lexicographical treatment. As exemplification, may mention the following instances: The word *φοβητσιάρης* [foviˈtsçaris] “fearful, coward” has as its base some form of the Standard word *φόβ-ος* [ˈfov-os] ‘fear’ or the verb *φοβ-άμαι* [fov-ame] “to fear”, while the adjective-forming suffix is the equally Standard and common *-άρης/ -ιάρης* [ˈaris]/[jaris] (fom Lat. < *arius*). However, this leaves unexplained the sequence [ts] in the middle of the word. There are no forms of the nominal or the verbal stem ending in [t] or [ts], and there is no suffix [ˈtsiaris] in Greek. Furthermore, in Standard Modern Greek there are no other adjectives ending in [ˈtsjaris], except those where [ts] is part of the stem, such as (1):

(1a) *γλίτσα* [ˈɣlitsa] ‘slime’ → *γλιτσ-ιάρης* [ɣliˈtsçaris] ‘slimy’*,*
(1b) *γκριμάτσα* [griˈmatsa] ‘grimace’ → *γκριματσ-ιάρης* [grimaˈtsçaris] ‘habitual mugger’
(1c) *κλωτσιά* [ˈklotsça] → *κλωτσ-ιάρης* [kloˈtsçaris] ‘habitual kicker’.

In order to interpret the form, the Modern Greek etymological dictionary of Andriotis (1983), followed by all Standard Modern Greek dictionaries, provides an interpretation which brings the diachrony-in-synchorny problem to the fore: It suggests that *φοβητσιάρης* [foviˈtsçaris] is a Medieval adjective, made up by the Ancient adjective *φοβητός* [phobɛːˈtos] > [foviˈtos] and the Modern suffix *-ιάρης* [ˈjaris]. Such an analysis is quite problematic both from a theoretical and a lexicographic viewpoint: *φοβητός* [foviˈtos] or even *φοβητικός* [fovitiˈkos] does not exist in Medieval or Modern Greek, and the two parts of the word never co-existed synchronically. To compound the problem, the authoritative general Standard dictionary *LKN* adds an extra step in the derivation: an assumed “strengthening of the articulation” [t] > [ts] before the semivowel /j/ and its allophones [ɟ]/[ç], i.e., *φοβητός* [foviˈtos] > **φοβητιάρης* [foviˈtçaris] > *φοβιτσιάρης* [foviˈtsçaris]. However, the articulatory fronting (affrication)[8] [t] > [ts] is not a phonetic rule of Standard Modern Greek; among other things, it would have resulted, in the case of similar derivatives, in unattested forms like *μεροκάματο* [meroˈkamato] ‘daily wage’ → *μεροκαματιάρης* [merokamaˈtçaris] ‘day labourer’ > **μεροκαματσιάρης* [merokamaˈtsçaris], *έρωτας* [ˈerotas] ‘love →*ερωτιάρης* [eroˈtçaris] ‘amorous’ > **ερωτσιάρης* [eroˈtsçaris].

As already pointed out above, methodological works on historical lexicography, and the practice of major historical dictionaries internationally, are clear on this point: an etymology, or a morphological analysis provided in a dictionary should consider the word in the specific synchrony when it was created, and analyse it only on the basis of elements and processes available at the time (as discussed at length in e.g., Chauveau 2005). This is also the principle followed by the *ILNE*, as stated in its *Manual of Regulations* (*ILNE-MR*).

For the interpretation of the lexical item *φοβητσιάρης* [foviˈtsçaris], an alternative analysis must therefore be sought. Of course, in the case of the *ILNE*, the solution to the problem is not a primary responsibility, since the word is already attested before the Modern period (in fact, its first attestation occurs in the *Erotokritos*, a Cretan Renaissance literary work dated around 1600), and a historical dictionary of a specific period should have as its task to fully analyse only the words which fall within its period of examination. So, in the case of Modern Greek and the *ILNE*, only words attested after 1800, the start of the Modern period, are provided with a full morphological analysis, which takes the place of an etymology. In other words, for “modern”, post 1800- words, diachronic etymology and synchronic morphological analysis coincide, something which cannot be true for past phases of the language (any language).

Nevertheless, even if the specific Standard word *φοβητσιάρης* [foviˈtsçaris] need not be treated etymologically in the *ILNE*, the process responsible for its creation does have to be addressed, since there are several other lexical items, belonging to various Modern Greek dialects, that present the problematic suffix [ˈtsjaris]. In the already published volumes of the *ILNE* one may find the forms *γέλιο* [ˈɟeʎo] ‘laughter’: *γελατσιάρης* [ɟelaˈtsçaris] ‘mirthful’ and *δειλία* [ðiˈlia] ‘cowardice’: *δουλιατσάρης* [ðuʎaˈtsaris] ‘cowardly’. One may also add a couple of Standard words which are not attested before 1800, such as *θυμός* [θiˈmos] ‘anger’: *θυμωτσάρης* [θimoˈtsaris] ‘irritable’ (listed in Anastasiadi-Symeonidi 2003).

It is in fact the dialectal words treated by the *ILNE* that point the way towards the solution to this etymological problem: *γελατσιάρης* [ɟelaˈtsçaris] and *δουλιατσάρης* [ðuʎatsaris] are not treated as dictionary entries *per se*, but as mere dialectal variant forms of the more widely attested words *γελασιάρης* [ɟelaˈsçaris] and *δειλιασιάρης* [ðiʎaˈsçaris]. This reveals the derivational path leading to their creation: the forms are derived from verbal stems, and specifically perfective (aoristic) stems augmented by the aoristic formative [-s-] denoting perfective aspect, in a pattern which is productive both in the Standard and in the dialects. Characteristic examples are common words like (2):

(2a) *ξεχνώ* [kseˈxno] ‘to forget’: *ξέχασ-α* [ˈksexas-a] ‘I forgot (past perfective)’ → *ξεχασιάρης* [ksexaˈsçaris] ‘forgetful’
(2b) *(ε)παινώ* [epeˈno] ‘to praise’: *(ε)παίνεσ-α* [eˈpenesa] ‘I praised (past perfective) → *παινεσιάρης* [peneˈsçaris] ‘braggart’
(2c) *αγαπώ* [aɣaˈpo] ‘to love’: *αγάπησα* [aˈɣapisa] ‘I loved (past perfective) → *αγαπησιάρης* [aɣapiˈsçaris] ‘sentimental’

The pattern appears also with deponent (medio-passive) verbs which do not present an active past perfective, but where the perfective formative [s] can be seen in derived nouns, e.g. (3):

(3a) *σιχαίνομαι* [siˈçenome] ‘to be disgusted’: *σιχασ-ιά* [sixaˈsça] ‘disgust’ → *σιχασιάρης* [sixaˈsçaris] ‘squeamish’
(3b) *καυχιέμαι* [kaˈfçeme] ‘I brag’: *καυχησ-ιά* [kafçiˈsça] ‘bragging’ → καυχησιάρης [kafçiˈsçaris] ‘braggart’.

It becomes obvious, therefore, that the apparently isolated and problematic derivational process leading to the formation of the word *φοβητσιάρης* [foviˈtsçaris] is in fact part of a well-established and productive pattern. Furthermore, the phonetic change responsible for the creation of the problematic affricate [ts], has so far been misunderstood: it is indeed a fronting process triggered by the semivowel [j] (the initial sound of the suffix /jaris/), but the effect caused by it is not the

[8] On the phenomenon see Holton et al. (2019: 122-123) and references therein.

"strengthening of [t]", but the palatalisation/affrication of [s].[9]

The comprehensive 'retrograde-regressive' investigation of current Standard vocabulary, only available in a large-scale historical dictionary like the *ILNE*, would also eventually have led to the solution of the problem, and avoided the perpetuation of the error in Modern Greek lexicography. As already described in section 1.2 above, the *ILNE* is called upon to provide a dating for the first appearance of its entries and all their variant forms, and does so through primary research both in textual corpora and in the earlier lexicographic tradition, thus serving as a sort of 'linguistic registry office' for Modern Greek.[10] Such an investigation, in the case of *φοβητσιάρης* [foviˈtsçaris] would have revealed that in the early 19th c. dictionaries it is still possible to find the earlier variant *φοβησιάρης* [foviˈsçaris] minus the affrication process; see e.g. the relevant entries in the Dictionaries of Gazis (1835) and Skarlatos Byzantios (1835).

So, the previous discussion aimed to demonstrate the following: On a first and obvious level, the "correct" etymological analysis of a word, even a well-known common word, cannot be achieved without a comprehensive historical investigation, which, in the case of Modern Greek, should also include the Modern Greek dialects. But more importantly, and generally, that it should be and is a principle of both historical lexicography and theoretical morphology that segmentation into morphemes only takes into account elements extant in the same synchrony.

Of course, it is not possible to oust diachrony completely from any lexicographical treatment of synchrony. In reality each lexicographic "synchrony" takes up several decades, often centuries. In the case of the *ILNE*, as discussed above, it takes up 200 years (from 1800 onwards); in the case of the *Medieval Greek Dictionary* (Kriaras 1968-) it takes up 600 years (1100-1669), and Ancient Greek dictionaries, ranging from Homer until the Hellenistic period take up even longer, covering about a thousand years. Within these periods, many phonetic and morphological changes operate, and these will appear, in the dictionaries, as "synchronic variation" whereas in fact they are the result of diachronic variation. Nevertheless, it is necessary to cut up time in larger chunks, otherwise we would need a different dictionary for every 50 years or so; and it is quite possible, in fact it is always being done both in Greek and in international lexicographic practice, to allow for a more general notion of "synchrony", as more extended time periods, delimited by basic major linguistic changes.

To turn now to a different example, the mingling of synchrony and diachrony also becomes evident in cases where the etymology and the dating of a modern dictionary entry is achieved, due to the lack of direct historical attestations of the simplex word, with the assistance of derivatives or compounds which do happen to be attested in earlier phases of the language, but which are not included in the *ILNE* as they are no longer extant.

One may mention as a typical instance the word *διάτανος* [ˈðʝatanos] 'devil', which is not attested before the 19th c. The lack of attestations is to be expected, given the 'vulgar' nature of the lexical item, normally employed as a swear-word (it is in fact a taboo deformation, a blend of the Koine words *διάβολος* [ˈðʝavolos] 'devil' and *σατανάς* [sataˈnas] 'Satan'). Although the oldest attestations of the word that research has uncovered come from theatrical plays of the early 19th c., it should be assumed that the word was formed sometime during the Dedieval period, given that there does exist a *hapax* medieval attestation of the derived noun *διατανοσύνη* [ðʝatanoˈsini] 'devilry, evil' dated to the late 15th c. (see Kriaras 1968-, s.v.). A similar example is the dialectal adjective *διαρμιστής* [ðʝarmiˈstis], feminine *διαρμίστρα* [ðʝarˈmistra] 'tidy, orderly, neat', both derived from the verb *διαρμίζω* [ðʝarˈmizo] 'to tidy up, clean, do housework'. Neither form of the derived adjective is attested before the modern period; however, 17th c. literature again contains a *hapax* attestation of a compound form of the feminine, *κακοδιαρμίστρα* [kakoðʝarˈmistra] 'bad housewife', which leads one to conclude that the word was in existence since the Early Modern period despite the lack of attestations.

As a third instance of diachrony-in-synchrony, one should also consider the issue of morphological segmentation. It has already been claimed that a basic lexicographic principle is to provide a morphological analysis involving only the elements extant in the same synchrony, and in the form they had in that specific synchrony. But this can be quite difficult, because it is not always possible to be certain about the actual form that a formative element had in a specific period. In other words, especially when one is dealing with derivational affixes, which diachronically undergo reanalysis and grow larger (through accretion) or smaller (through truncation), it is not always possible, in historical lexicography, to know where to set the morpheme boundary.[11] The "same" word, i.e. a lexical item that retains the same phonological shape in different periods, may have a different morphological analysis in each period, without one being able to ascertain when the boundary "moved". To give a concrete example:

A very common and productive Ancient Greek adjective-forming suffix is *-ρός* [ˈros], with a multiplicity of vowels that may precede it[12] (4):

(4a) *πόνος* [pónos] 'toil, trouble, pain' → *πονηρός* [ponɛ:rós] 'painful > base, cowardly'
(4b) *τόλμη* [tólmɛ:] 'courage, boldness' → *τολμηρός* [tolmɛ:rós] 'bold, daring'
(4c) *κράτος* [krátos] 'might' → *κρατερός* [kraterós] 'mighty'
(4d) *ἰσχύς* [isk^{h}ýs] 'power' → *ἰσχυρός* [iskhyrós] 'powerful'

At some point in Medieval Greek, the suffix was reanalyzed as *-ερός* [eˈros], so although the above words survive in Modern Greek as [poniˈros], [tolmiˈros], [krateˈros], [isxiˈros], innovative derivatives created through the suffix only present the vowel /e/ (5):

(5a) *λάδι* [ˈlaði] 'oil' → *λαδερός* [laðeˈros] 'oily'

[9] For this phenomenon and its geographical distribution see Holton et al. (2019: 122) and references therein.

[10] On the issue see Manolessou & Katsouda (forthcoming a). In fact, the comprehensive diachronic investigation of Standard vocabulary, not provided in any other major general dictionary of Modern Greek, leads to the revision both of the etymology and of the dating of dozens of words.

[11] For the diachronic changes affecting derivational suffixes, with various examples also taken from Greek, see Haspelmath (1995).

[12] On the suffix *-ρός* [ros] and its productivity and variants in Ancient Greek see Probert (2006: ch. 6).

(5b) *λίγδα* [ˈliɣda] ‘grime, gunk’ → *λιγδερός* [liɣðeˈros] ‘grimy, full of gunk’
(5c) *βαμβάκι* [vamˈvaci] ‘cotton’ → *βαμβακερός* [vamvaceˈros] ‘made of cotton’
(5d) *σούβλα* [ˈsuvla] ‘roasting spit’ → *σουβλερός* [suvleˈros] ‘sharp’
(5e) *σίχαμα* [ˈsixama] ‘disgust’ → *σιχαμερός* [sixameˈros] ‘disgusting’
As a next evolutionary step, which leads to a modern lexicographic problem of synchronic vs. diachronic analysis, at some point the suffix underwent accretion and acquired a variant form -*τερός* [teˈros], on the analogy of stem forms ending in [t] such as (6):
(6a) *μύτη* [ˈmiti] ‘nose, tip’ → *μυτερός* [miteˈros] ‘pointed, sharp’
(6b) *αστράφτω* [aˈstrafto] ‘to sparkle’ → *αστραφτερός* [astrafteˈros] ‘sparkling.
Consequently, a number of innovative derivatives show a suffix [teˈros], without the presence of final [t] in the stem. Examples include (7):
(7a) *γυαλίζω* [ʝaˈlizo] ‘to shine’: *γυάλισ-α* [ˈʝalisa] ‘past perfective → *γυαλισ-τερός* [ʝalisteˈros] ‘shiny’
(7b) *διαβάζω* [ðʝaˈvazo] ‘to read’: *διάβασ-α* [ˈðʝavasa] ‘past perfective → *διαβασ-τερός* [ðʝavasteˈros] ‘book-worm’
(7c) *γαμώ* [ɣaˈmo] ‘to fuck’: *γάμησ-α* [ˈɣamisa] ‘past perfective’ → *γαμησ-τερός* [ɣamisteˈros] ‘fucking good’
But the dating of the morpheme boundary shift is difficult to determine. The Medieval Greek Dictionary (Kriaras 1068-) includes very few such forms, The earliest seems to be λυπώ [liˈpo] ‘sadden’ → λυπητερός [lipiteˈros] ‘saddening’ .
However, in the case of verbs which also might form verbal adjectives in -τός [tos], it is difficult to decide which variant of the derivational suffix is involved, [eˈros] or [teˈros]. For example (8):
(8a) βράζω [ˈvrazo] ‘to boil’ → βραστός [vrasˈtos] ‘boiling’ → βραστ-ερός [vrasteˈros] ‘easily boiled’ or
(8b) βράζω [ˈvrazo] ‘to boil’ → έβρασ-α [ˈevrasa] ‘past perfective’ → βρασ-τερός [vrasteˈros] ‘easily boiled’
As a result, in cases like these the irruption of diachrony, in the guide of "older form of a suffix" in the synchronic morphological analysis, cannot be avoided.

5 Semantic Section

The mingling of synchrony and diachrony in historical and dialectal lexicography is also to be met with in the semantic section. This is to be expected, to a certain extent, in that historical dictionaries need to provide a dating for each listed sense, and the examples and quotations are often presented in chronological order. Cf. the similar concerns about the presence of both synchronic and diachronic descriptions in the semantics sections of the *SAOB* expressed by Stille (2001: 228-229).

In the case of the ILNE, the connection to earlier phases of linguistic history is not achieved only through the fact that all senses are assigned a dating as to the overall period of the language they first appear (i.e., ancient, postclassical, medieval, early modern, or modern), but also frequently through quotations. These are deemed indispensable, as they serve to corroborate or justify the dating of a sense, in cases when the ILNE provides a different dating than that usually assumed in the major general dictionaries of the Standard language.

To give an example among many, for the entry *δικέφαλος* [ðiˈcefalos] ‘two-headed’ the ILNE needs to document that the attestation of this adjective with specific reference to muscles is already medieval, with a textual excerpt from a medical work of the 7th c. Similarly, for the entry *δευτερόλεπτο* [ðefteˈrolepto] ‘second’ the ILNE again needs to document that the attestation of this adjective with reference to a measure of time is medieval, with a textual excerpt from a Byzantine astronomer. In both cases, general dictionaries of Greek consider these senses to be recent translation loans from French *biceps* and *seconde* respectively. The issue is discussed in more details and further examples in Manolessou (2016) and Manolessou & Katsouda (forthcoming a).

Given the Dictionary’s time-frame, 1800-today, the diachronic dimension also makes its presence strongly felt in the case of senses which were current in the 19th c., but are no longer to be met with (except, of course, in historical accounts or narrations). A typical example is constituted by the names of various types of coinage, see e.g., the entries *δεκάρα* [ðeˈkara] ‘coin of ten cents of the drachma’, *δεκάρικο* [ðeˈkariko] ‘coin of ten drachmas’, *δεκαχίλιαρο* [ðekaˈciʎaro] ‘banknote of ten thousand drachmas’, *δίφραγκο* [ˈðifrago] ‘coin of two francs’*, δίλεπτο* [ˈðilepto] ‘coin of two cents’*, δίλιρο* [ˈðiliro] ‘ coin of two pounds or liras’*, δίγροσο* [ˈðiɣroso] ‘coin of two piasters’*, διόβολο* [ðiˈovolo] ‘coin of two alms’ in the latest volumes of the *ILNE*. In certain cases, the whole entry does not belong to the modern synchrony, as the coin in question is no longer in use, but is included in the dictionary as it was current in the standard language during the 19th and (part of the) 20th c. In other cases, the coin in question is no longer in use, but a semantic change has taken place and its name has been retained in order to refer to a new coin (e.g., *δίλεπτο* [ˈðilepto] ‘two cents of a drachma’ > ‘two cents of a euro’). Of course, it is also the case that in various dialectal sources names of outdated coins are frequently to be found, both because the dialectal material may date as early as the mid-19th c. and because it is retained through oral tradition in popular songs, games, sayings, nursery rhymes etc. Indeed, the preservation of older coin names in set phrases and proverbs (e.g., *δε δίνω δεκάρα* [ðe ˈðino ðeˈkara] ‘I don’t give a damn’*,* or *τέρμα τα δίφραγκα* [ˈterma ta ˈðifraga] ‘end of story’, ‘game over’) necessitates the inclusion of such entries even in the general dictionaries of Standard Modern Greek.

6 Conclusions

On the basis of the above discussion, the mingling of synchrony and diachrony is an inevitable facet of historical and dialectal lexicography. It is deemed necessary in order to present and interpret the evolution of words, forms and senses, so long as the reader is forewarned and made fully aware of this principle. This can be ensured in the lexicographical work’s introduction or manual of regulations, where the methodology adopted is set out, and the entry slots where it occurs are noted. The systematic application of a parallel synchronic and diachronic examination of each entry emerges, then, not as an accidental byproduct, but rather as a conscious methodological tool.

7 References

Anastasiadi-Symeonidi, A. (2003). *Αντίστροφο Λεξικό της Νέας Ελληνικής* [Reverse Dictionary of Modern Greek]. Thessaloniki: Institouto Neoellinikon Spoudon, Manolis Triantafyllidis Foundation.

Anastasiadi-Symeonidi, A. & Fliatouras, A. (2019). Το λόγιο επίπεδο της σύγχρονης Νέας Ελληνικής: Συγχρονικές και διαχρονικές τάσεις [The learned register in Standard Modern Greek: synchronic and diachronic tendencies]. In A. Fliatouras & A. Anastasiadi-Symeonidi (eds), *Το λόγιο επίπεδο στη Νέα Ελληνική*. Athens: Patakis, pp. 15-56.

Andriotis, N. P. (1974). *Lexikon der Archaismen in neugriechischen Dialekten.* Wien: Österreichische Akademie der Wissenschaften.

Andriotis, N.P. (1983). *Ἐτυμολογικὸ λεξικὸ τῆς κοινῆς Νεοελληνικῆς* [Etymological dictionary of Standard Modern Greek] (3rd edition). Thessaloniki: Institute of Modern Greek Studies [Manolis Triantafyllidis Foundation].

Banniard, M. (2002). Sur la notion de fluctuation langagière en diachronie longue (IIIe+VIIIe s.) à la lumière des enquêtes dialectologiques contemporaines. In *Revue belge de Philologie et d'Histoire, Année* 80(3), pp. 779-788.

Chauveau, J.-P. (2005). Remarques sur la dérivation dans les notices historiques et étymologiques du Trésor de la langue française. In E. Buchi (éd.), *Actes du Séminaire de méthodologie en étymologie et histoire du lexique (Nancy/ATILF, année universitaire 2005/2006).* Accessed at: www.atilf.fr/wp-content/uploads/manifestations/seminaires/atilf_seminaire_melh_Chauveau_2005-11-16.pdf [10.4.2021].

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. Blackwell.

DWB: Grimm, J. & W. Grimm 1854-1960. *Deutsches Wörterbuch,* Vol. I-XVI. Leipsig.

EDD: Wright, J. (1898-1905). *English Dialect Dictionary*, vol. 6. Oxford: Henry Frowde.

Ferguson, C. A. (1959). Diglossia. In *Word,* 15(2), pp. 325-340.

Gazis, A. (1835). *Λεξικὸν τῆς ἑλληνικῆς γλώσσης τρίτομον [...] νῦν δὲ τὸ δεύτερον ἐπεξεργασθέν, διασκευασθὲν [...] καὶ εἰκοσιτέσσαρσι περίπου χιλιάσι λέξεων προσεπαυξηθὲν* […]. [Three-volume dictionary of the Greek language, in second revision and with the addition of about 24.000 new words], Wien.

Haspelmath, M. (1995). The Growth of Affixes in Morphological Reanalysis. In Booij, G. & J. van Maarle (eds.) *Yearbook of Morphology* 1994, pp. 1-29.

Holton, D., Horrocks, G., Janssen, M., Lendari, T., Manolessou, I. & Toufexis, N. (2019). *The Cambridge Grammar of Medieval and Early Modern Greek.* Cambridge: Cambridge University Press.

ILNE: *Ἱστορικὸν Λεξικὸν τῆς Νέας Ἑλληνικῆς, τῆς τε κοινῶς ὁμιλουμένης καὶ τῶν ἰδιωμάτων*. [Historical Dictionary of Modern Greek, both of the Standard and the dialects] Vol. 1-7, α-δόγης. Athens: Academy of Athens, 1933.

ILNE-MR: *Manual of Regulations of ILNE*: *Κανονισμός Συντάξεως του Ιστορικού Λεξικού της Νέας Ελληνικής.* Λεξικογραφικόν Δελτίον Παράρτημα 6. Athens: Academy of Athens, 2012.

Katsouda, G. (2012). Διαλεκτική Λεξικογραφία: Επισκόπηση και Ζητήματα [Dialectal lexicography: Overview and issues]. In *Λεξικογραφικόν Δελτίον*, 26, pp. 77-159.

Katsouda, G. (2016). Ο νέος τόμος του Ιστορικού Λεξικού της Νέας Ελληνικής της Ακαδημίας Αθηνών: συγχρονικές προοπτικές [The new volume of the Historical Dictionary of Modern Greek of the Academy of Athens: synchronic perspectives]. In *Studies in Greek Linguistics*, 36, pp. 151-160.

Kriaras, E. (ed.) (1968-). *Λεξικό της Μεσαιωνικής Ελληνικής Δημώδους Γραμματείας 1100-1600* [Dictionary of Vernacular Medieval Greek Literature, 1100- 1600). Thessaloniki: Centre for the Greek Language.

LKN = Ινστιτούτο Νεοελληνικών Σπουδών (Ίδρυμα Μανόλη Τριανταφυλλίδη). 2007.[7] *Λεξικό της Κοινής Νεοελληνικής* [Dictionary of Standard Modern Greek]. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

Mackridge, P. (2007). *Language and National Identity in Greece 1766–1976*. Oxford: OUP.

Manolessou, I. (2012). Το Ιστορικόν Λεξικόν της Νέας Ελληνικής ως ιστορικό λεξικό. [The Historical Dictionary of Modern Greek as a historical dictionary]. In *Λεξικογραφικόν Δελτίον*, 26, pp. 9-75.

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]. In *Studies in Greek Linguistics,* 36, pp. 239-349.

Manolessou, I. & Bassea-Bezantakou, Ch. (2013). The Historical Dictionary of Modern Greek: dialectological issues. In *Dialectologia*, Special issue IV, pp. 25-48.

Manolessou, I., Beis, St. & Bassea-Bezantakou, Chr. (2012). Η φωνητική απόδοση των νεοελληνικών διαλέκτων και ιδιωμάτων [The phonetic transcription of Modern Greek dialects]. In *Λεξικογραφικόν Δελτίον* 26, pp. 161–222

Manolessou, I & Katsouda, G. (forthcoming a). Η ετυμολογία στο Ιστορικό Λεξικό της Νέας Ελληνικής [Etymology in the Historical Dictionary of Modern Greek]. In *Proceedings of the 2nd International Conference in Greek Etymology (2-3 November 2018),* Thessaloniki.

Manolessou, I. & Katsouda, G. (forthcoming b). The making of the Historical Dictionary of Modern Greek: Problems and solutions in the domain of historical and dialectal lexicography. In *Proceedings of the 10th International Conference on Historical Lexicography and Lexicology* (12-14 June 2019), Fryske Akademy, Leeuwarden.

Manolessou, I. & Ralli, A. (2015). From Ancient Greek to Modern Greek. In P. O. Müller, I. Ohnheiser, S. Olsen, Fr. Rainer (eds.), *Word-Formation. An International Handbook of the Languages of Europe*. Berlin: De Gruyter, pp. 2041-2061.

Newton, B. (1972). *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology.* Cambridge: CUP.

OED: *Oxford English Dictionary*. Accessed at: http://public.oed.com/how-to-use-the-oed/glossary/ [06/04/2020]

Petrequin, G. & Andronache, M. (2008). Le programme TLF-Étym: apports récents de l'étymologie comparée-reconstruction. In E. Bernal and J. DeCesaris (eds.), *Proceedings of the 13th EURALEX International Congress*. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 1165-1173.

Pilch, H. (1985). The synchrony-diachrony division in word-formation. In J. Fisiak (ed.) *Historical Semantics - Historical Word-Formation*. Berlin/New York: Mouton De Gryuter, pp. 407-434.

Probert, Ph. (2006). *Ancient Greek Accentuation: Synchronic Patterns, Frequency Effects, and Prehistory*. Oxford: OUP.

Reichmann, O. (1990). Formen und Probleme der Datenerhebung I: Synchronische und diachronische historische Wörterbücher. In Hausmann et al. (1989–1990), vol 2, pp. 1588–611.

Reichmann, O. (2012). *Historische Lexikographie. Ideen, Verwirklichungen, Reflexionen an Beispielen des Deutschen, Niederländischen und Englischen*. Berlin/Boston: De Gruyter.

Ruszkiewicz, P. (1997) On the diachrony-in-synchrony analysis in morphology. *Folia linguistica historica* 31. In *Historica,* 18 (1-2), pp. 81-118.

Rys, K. & Van Keymeulen, J. (2009). Intersystemic Correspondence Rules and Headwords in Dutch Dialect Lexicography. In *International Journal of Lexicography* 22, pp. 129-150.

SAOB: *Svenska Akademiens Ordbokhttp*. Accessed at: http://g3.spraakdata.gu.se/saob/ [29/6/2021]

Skarlatos Byzantios, D. (1835). *Λεξικὸν τῆς καθ' ἡμᾶς ἑλληνικῆς διαλέκτου μεθηρμηνευμένης εἰς τὸ ἀρχαῖον ἑλληνικὸν καὶ τὸ γαλλικόν*. Athens.

Stille, P. (2001). Working on a Historical Dictionary: The Swedish Academy Dictionary Project. In *Lexikos* 11, pp. 222-230.

Trudgill, P. (2003). Modern Greek dialects: a preliminary classification. In *Journal of Greek Linguistics,* 4, pp. 45-64.

Tzitzilis, Ch. (2013). Archaisms in Modern Dialects. In G. Giannakis (ed.) *Encyclopedia of Ancient Greek Language and Linguistics*. Leiden/Boston: Brill, pp. 158-171.

WBÖ: Kranzmayer, E. (ed.) (1963-). *Wörterbuch der bairischen Mundarten in Österreich* vol. 5. Wien: Österreichische Akademie der Wissenschaften.

WNT = Woordenboek der Nederlandsche Taal (1864-1998). Rotterdam: AND Publishers b.v. Online version 1999, Accessed at: https://ivdnt.org/onderzoek-a-onderwijs/lexicologie-a-lexicografie/wnt [29/6/2021]

New words in old sources: Additions to the lemma list of a historical scholarly dictionary

Johannsson E.T., Battista S.

Department of Nordic Studies and Linguistics, University of Copenhagen
ellert@hum.ku.dk, sb@hum.ku.dk

Abstract

This paper accounts for recent additions to the lemma list of *A Dictionary of Old Norse Prose (ONP)*, which is a historical dictionary describing the medieval language of Iceland and Norway. The dictionary was established in 1939 and has throughout the years built up a large database containing about 800.000 example citations illustrating the vocabulary of all prose genres. The lemma list consists of about 65000 words with accompanying citations, but is continuously being revised. After giving a brief account of the history of this project we give an overview of the editorial principles, the criteria used for defining a lemma and discuss different types of lemmas found in the dictionary. We describe the characteristics of entries in ONP and mention different types of entries found in the online version. We then focus on the period from 2010-2019 and present a study into new additions to the lemma list during those years. We analyze these more recent words, divide them into eight groups and give some examples that illustrate the processes involved when new headwords are established. The results of the study show that most of the later additions to the lemma list come about in relation to editorial work on other words. A significant proportion of new words are established when new compounds are identified while editing uncompounded, simplex words, but other factors are in play as well.

Keywords: historical lexicography; morphology; lexicology

1 Introduction

A Dictionary of Old Norse Prose (ONP) is a dictionary project hosted at the University of Copenhagen and part of the Arnamagnæan Institute of Old Norse Manuscript Studies. This dictionary accounts for the vocabulary of the language of medieval Iceland and Norway, from around 1150 to 1370 (Norway) and to 1540 (Iceland). The corpus consists of texts preserved in manuscripts, with all the implications of text transmission, which make every version of a text unique. The lexical material is the result of extensive excerption work mostly from scholarly editions of manuscript texts and in some cases directly from manuscripts. ONP has since 2010 been available as an online resource at onp.ku.dk, which provides access to the material from the published volumes of the dictionary (1995-2004) as well as more recently edited dictionary entries and unedited dictionary material. The work on the dictionary continues with new entries published online, as well as addition of new features to the online version.

The paper is organized as follows: After giving a brief account of the background and history of the project we discuss some of the editorial principles, the criteria used for defining a lemma and account for different types of lemmas in ONP. Next, we describe the characteristics of entries in the dictionary and give examples of the different types of entries found in the online version. We then account for a study into new additions to the lemma list of ONP during the period from 2010 to 2019 and present the results of our findings.

2 Background

The ONP dictionary project was established in 1939. The focus of the lexicographic work has always been on the language of medieval prose texts from Norway and Iceland as the poetic language had already been described in great detail with the publication of a revised dictionary of the poetic language a few years earlier (Jónsson 1931). The dictionary was originally meant to be a supplementary to the *Ordbog over det gamle norske Sprog* (1886-1896), by Johan Fritzner. In the meantime, it had become clear that Fritzner's work had some limitations with its normalization practices and citing of many text editions that had become obsolete. It was decided that a new lexical description of Old Norse prose was needed, with the aim of giving an exhaustive representation of the vocabulary of Old Norse prose excerpted from all known texts in a scholarly edition or directly from manuscripts. A new dictionary of this kind would also fit well into the Danish lexicographic tradition and would strengthen Copenhagen as an important research center for the Nordic cultural heritage, with many important manuscripts being preserved there.

In the first decades, the dictionary staff was mostly concerned with gathering material for an eventual print publication. This meant selectively excerpting all known medieval texts, representing the vocabulary of different Old Norse prose genres, by collecting examples of word use. Selected citations were written onto slips, which were then filed under a particular headword in alphabetical order (cf. figure 1). The citation collection was intended to be very detailed and illustrate the range in meaning of every word.

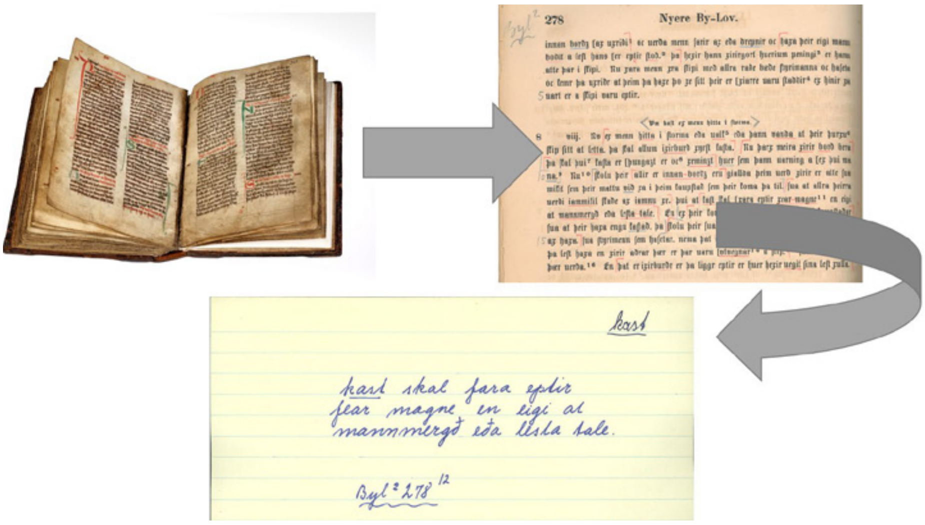


Figure 1: The medieval manuscript is edited in a scholarly edition, which in turn is excerpted by underlining relevant citations and writing them down on a paper slip.

A few key works – representing different genres, from particularly relevant manuscripts – were exhaustively excerpted, i.e., each and every word in those texts was copied onto a slip along with its syntactic context and filed in the dictionary archive. This provided additional examples of many commonly used words that were underrepresented in the citation collection. The exhaustively excerpted texts represent different genres and among these we find a section of *Snorra Edda* (mythological tales) from the manuscript GKS 2367 4º, a fragment of *Egils saga Skallagrímssonar* (an Icelandic family saga) from AM 162 A θ fol, and *Íslendingabók* (an historical account of the settlement of Iceland) from AM 113 fol^x.[1] As the decades passed and more scholarly text editions were published, the citation collection grew. This eventually resulted in an archive consisting of around 750.000 handwritten slips, organized under 65.000 lemmas (cf. Johannsson & Battista 2014).

Once the examples had been collected, plans were made for the publication of a multi-volume print dictionary. The first volume was an index volume published in 1989 followed by three volumes of dictionary entries covering the alphabet from *a-em*. In 2005, after the third volume had come out, the print publication was put on hold and preparation began for the transition to a digital online dictionary (cf. Johannsson & Battista 2016). ONP Online (onp.ku.dk) was launched in 2010 and combined material from the already published volumes, newly edited entries and unpublished material in the form of scanned citation slips (cf. Johannsson 2019). The online version has more recently been redesigned and enhanced in different ways with linking to other digital resources for Old Norse (cf. Wills & Johannsson 2019).

The editing work is ongoing and focuses on word groups rather than alphabetical order of lemmas. The headwords are divided into twelve groups for the purpose of editing: simplex nouns (with fewer than ten citations) simplex nouns (with ten or more citations), compound nouns, verbs, simplex adjectives, compound adjectives, simplex adverbs, compound adverbs, pronouns, numerals, conjunctions and prepositions (cf. Johannsson and Battista 2016). The grouping is mostly based on part of speech, but also morphological features and frequency. The first group to be edited following this new procedure was simplex nouns with more than ten citations, followed by simplex nouns with fewer citations. Since this new editing procedure was put in place all the simplex nouns have been edited along with simplex adjectives and adverbs. Pronouns, numerals, conjunctions and prepositions are also close to being finished. The editing of verbs is done in two rounds with an initial round of editing focusing on argument structure and formal categories rather than meaning. This initial editing is now completed. The second round, which involves the semantic editing of verbs, is currently underway. The largest groups that remain completely unedited are compound nouns and compound adjectives.

3 Editorial Principles and Lemma Criteria

The ONP dictionary has from the beginning followed certain clearly defined editorial principles. An important feature, which distinguishes ONP from its predecessors, involves adhering to the original orthography of the source texts and maintaining rigorous philological standards. This entails that the citations are as far as possible taken from diplomatic scholarly editions or even unpublished manuscripts, which means that the orthography of the citation examples is highly irregular (cf. Johannsson & Battista 2016: 118-119).

Even though the example citations are not normalized, the lemma list of ONP is normalized according to a normalization standard developed by ONP. This normalization is similar to the classic Old Norse spelling often used in text editions, which reflects the phonological state of Icelandic around year 1200 (a detailed overview of normalization practices and the principles of Old Norse normalization is found in Bernarðsson et al. 2019). ONP's orthography differs in some significant ways as it tries to take into account both Norwegian and Icelandic language development. It does not reflect some special Icelandic sound changes, such as vowel lengthening before certain consonant clusters, e.g., standard Old

[1] See, e.g., handrit.is for more information about these manuscripts and shelf marks.

Norse *úlfr* 'wolf' with a long u-vowel is normalized as *ulfr* by ONP as this lengthening rule did not take hold in Norway. Another important difference is the consistent use of the acute accent to mark all long vowels by also using the lesser-known characters *ǽ* and *ǿ,* which are not part of the traditional Old Norse standard orthography. This approach has some pedagogical advantages and constitutes a good compromise when it comes to developing standards that can be used for editions of both Norwegian and Icelandic texts. ONP's orthography, however, has only a limited tradition in text editions. It has also been revised a few times, most recently in 2003, which could give the impression that it is not as well established as other orthographic standards (cf. discussion in Johannsson and Battista 2020).

When choosing the form of lemma, ONP follows the criteria already established by its predecessors. Nouns are listed in their nominative singular form, verbs are given in the infinitive, active voice and adjectives in the masculine, nominative singular form. (cf. ONP Nøgle/Keys 2004)

The main components of a typical dictionary entry is the headword, along with grammatical classification, details of inflection and the example citations found in the dictionary archives. The citations are either typed in or shown as scanned paper slips (see section 4 below for detailed look at different types of entries). Unlike most dictionaries, ONP lists all the examples it has registered for each lemma, and, in many cases, these are all the attested examples of word use.

The ONP dictionary features several different types of entries and this is reflected in the lemma list. There is a distinction between so-called standard headwords and secondary headwords. Standard headwords contain citations arranged chronologically according to senses (if the entry has already been edited), and a concluding section with supplementary information. The secondary headwords are in principle "registrations of a word's existence, with no semantic explanation, but with references to other dictionaries and glossaries (Gloss.), and occasionally to secondary literature" (cf. ONP Nøgle/Keys 2004). These secondary entries can be further divided into several subtypes:

- *Poetical words*. These are words, which, in spite of appearing only in poetry context, are recorded in the dictionary for their lexicographic value. These entries only rarely have example citations (from poetic use in prose texts). They contain reference to a relevant glossary over the poetic language and are labelled (*poet*.). An example would be *ǫglir,* a word for hawk or falcon with multiple occurrences in poetry but never in prose.
- *Non-assimilated foreign words*. These are foreign words that appear in an Old Norse context, but are not adapted to the language, either phonologically or morphologically. Such words are labelled (*alien*.) or (*foreign*) and spelled according to the language of origin. Some examples would be *cherub*, *schismus* and *synecdoche*. Integrated loanwords however are treated in the same way as other Old Norse words.
- *Starred words*. These words are of various sorts. They can be words that appear in other Old Norse dictionaries, but fall outside ONP's defined scope. They can also be so-called "ghost words" which are the result of an erroneous reconstruction in a text edition or a misreading of a manuscript, which has found its way into a published text, or are based on an interpretation with which ONP does not agree. Many such words stem from earlier dictionaries, based on material from old and obsolete editions that have since been replaced by more precise scholarly editions. A good example is the hapax *duma*, which is a result of a misreading of the quite common verb *dvína* in an early text edition. It is often difficult to see the difference between -*in*- and -*m*- in manuscripts so such errors can arise. A more careful reading in a later edition has revealed the mistake. The word *duma* is still recorded but receives a star, as it does not have any philological basis.
- *Questionable words*. Some words of uncertain status, which are attested in the actual medieval material, but are most likely a result of an error, are preceded by a question mark and are usually followed by a reference to a likely "correct" form. An example would be the adjective ?*gozkr* for *girzkr* 'Greek'
- *Prefixes and suffixes*. These are items listed in other dictionaries/glossaries. An example is the suffix -*geðjaðr*, which only exists in derived words such as *lausgeðjaðr* 'indecisive'.
- *References*. There are references from so-called alternative forms, which are variant forms that share most characteristics with the main form, i.e., high frequency and straight forward normalization, as well as references to so-called special forms, which are isolated occurrences that are not suited to normalization but cannot be a result of an error. An example would be *skemmtan* which is a common alternative form to *skemmtun* 'entertainment', which is linked to a single entry *skemmtun, skemmtan sb. f.*
- *Final element of a compound*. Such elements are not attested as independent words and thus have no definition or citations, but only a reference to the attested compounds where the element occurs, as well as relevant references to older dictionaries and secondary literature. An example would be -*saltaðr* 'salted', which is only attested in the adjectival compound *ósaltaðr* 'unsalted'.
- *Out of scope words*. These are words that have been registered by ONP but do not fall within the defined scope of the dictionary. These words can be for example words that are only attested in younger text sources but are deemed to have relevance to the description of the medieval vocabulary, such as *allraheilagramannamessuaftann* 'the eve of all saints mass' only attested in a Norwegian text from around 1380 but shows an example of a very long nominal compound. Another type would be place names and personal names, such as *England* or *Egill*. The words that fall under this heading are clearly marked as such.

The standard headwords follow certain principles. As stated in the User's Guide for the dictionary the oldest and etymologically most original form is usually chosen as the main form.[2] If a variation between forms is purely

[2] The User's Guide is a helpful aid originally intended for the users of the print dictionary. It explains the editorial principles of ONP in some detail as well as the structure of the entries. The guide was published in a booklet that accompanied each printed volume, and contained corrections along with a list of abbreviations and symbols. The latest one is ONP Nøgle/Keys 2004. The User's Guide is now

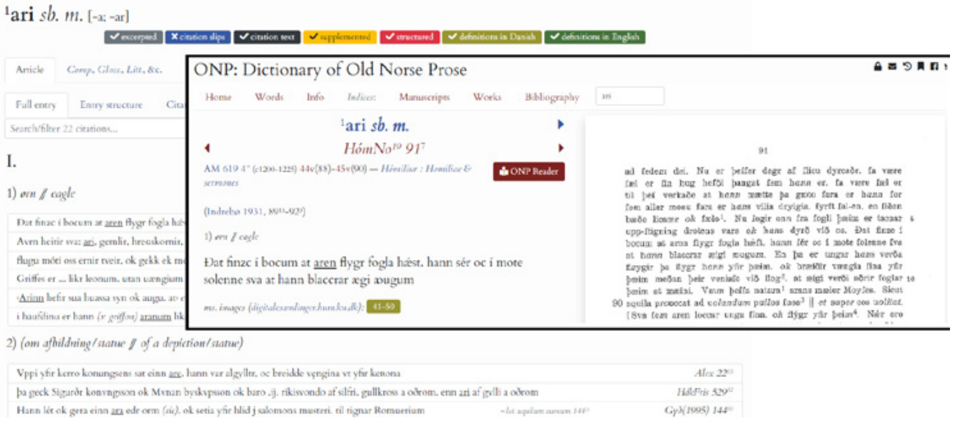
orthographical or a result of a clearly understood phonetic variation, i.e., dialectal, the variants are treated under the same headword. If the difference between variant words is the result of inflectional discrepancies, the words in question will each be listed in their own entry with their own set of citations. This is often the case with verbs that are attested with forms belonging to different conjugational patterns or nouns that display forms that belong to different declensional classes. This is also the case with compound words where the members of the compound are the same but joined together by different morphological elements, e.g., *barns-faðir*, *barna-faðir* and *barn-faðir*, which all have the same basic meaning as 'a father of child/ren'. There are two exceptions, where variant forms are grouped together under the same lemma. One is words where the second element starts with an s and variant forms with both *-s-* and *-ss-* are recorded the lemma is normalized as with one *s* in brackets, e.g., *dóm(s)sæti* 'court seat'. Another rare exception is when the first member ends in *-ar* and alternatively in *-a* in some Norwegian sources, e.g., *atfara(r)þing* 'assembly called to obtain an order of distraint' (cf. ONP Nøgle/Keys 2004).

4 Types of Entries in ONP Online

In its current form, the online dictionary displays one of four types of entries for each standard headword. The reason for these different entry types can be explained by the complicated history of the project as a partial print publication and later on as an online work in progress. In the sections below we will give a brief description and examples of each type.

4.1 Edited Entries from the Print Edition

The edited entries that also appear in the printed volumes are published online without any major changes. The structure contains all the same detailed information found in print as well as the keyed-in citation examples. Unlike the print edition all the citations that accompany each headword are listed and most citations are displayed along with a scanned page from an edition of the text (cf. figure 2). When publishing in print the number of citations often had to be reduced so the editors would pick the best representative examples for publication. In the online version of ONP all the citations found in the dictionary archive have been made available and the unpublished citations have been fully integrated into the entry structure and placed under the appropriate sense.



I.

1) *ørn* // *eagle*

Þat finzc i bocum at aren flygr fogla hæ

Avrn heitir sva: ari, gemlir, hreskornir,

flugu móti oss ernir tveir. ok gekk ek m

Griffes er ... likr leonum. utan uængium

‹Arinn hefir sua huassa syn ok augu. at

i haufdinu er hann (*r griffon*) aranum lik

2) (*om afbildning/statue* // *of a depiction/statue*)

Vppi yfir kerro konungsens sat einn are. hann var algylltr. oc breidde vęngina vt yfir kenona		*Alex* 22¹³
þa geck Sigurðr konvngsson ok Mvnan byskvpsson ok baro .ij. rikisvondo af silfri. gullkross a oðrom. enn ari af gylli a oðrom		*HákFris* 529¹²
Hann lét ok gera einn ara edr orm (*sic*). ok setia yfir hlid j salomons musteri. til tignar Romuerium	~*lat. aquilam aureum* 144³	*GyðÞ(1995)* 144¹⁰

¹ari *sb. m.*

*HómNo*¹⁹ *91*⁷

AM 619 4° (c1200-1225) 44v(88)-45v(90) — *Hómilíur : Homiliae & sermones*

(Indrebø 1931, 89¹³-92⁹)

1) *ørn* // *eagle*

Þat finzc i bocum at aren flygr fogla hæst. hann sér oc i mote solenne sva at hann blaccrar ægi augum

ms. images (digitalesamlinger.hum.ku.dk): 41-50

91

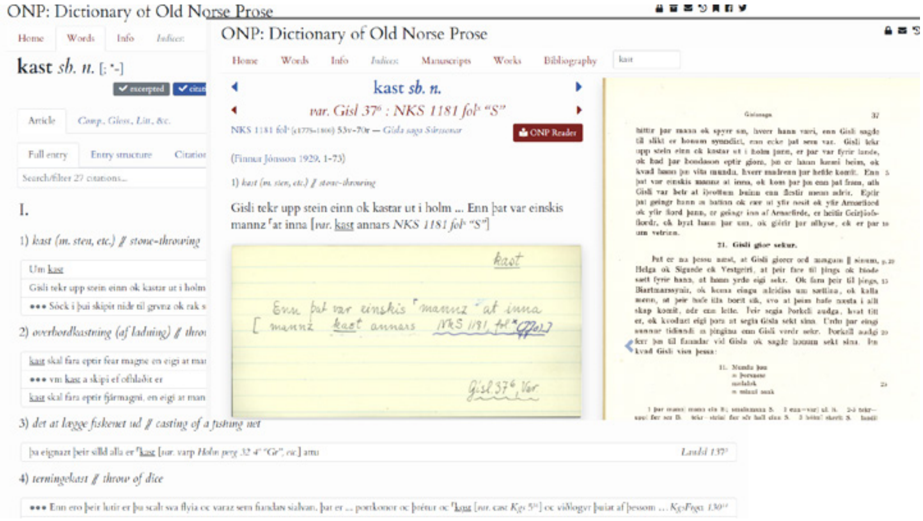
ad ſedem dei. Nu er þeſſer dagr af ilicu dyrcaðr. ſa være ſæl er ſin hug hefði þangat ſem hann er. ſa være ſæl er til þeſ verkaðe at hann mætte þa goðo fara er hann for ſem aller menn fara er hans vilia dryigia. fyrſt ſal-en. en ſíðan baðe licame ok ſale¹. Nu ſegir oss fra fogli þeim er tacnar upp-ſtigning drotens vars ok hans dyrð við os. Þat finze i bocum at arna flygr fogla hæſt. hann ſér oc i mote ſolenne ſva at hann blaccrar ægi augum. En þa er ungar hans verða fleygir þa flygr hann yfir þeim. ok breiðir vængia ſina yfir þeim meðan þeir veniaſc við flog². at eigi verði aðrir foglar þeim at meini. Venn þeſſa natura¹ arans maler Moyſes. Sicut aquila provocat ad volandum pullos ſuos³ || et super eos uolitat. [Sva ſem aren loccar unga ſina. ok flygr yfir þeim⁴. Nær ero

Figure 2: Screenshots showing an entry that was previously published in print. Additional citations have been added to the entry structure and for each citation a scanned page from the relevant edition is made available (second screenshot superimposed).

4.2 Entries that Have Been Edited since the Print Edition Was Put on Hold

The entries that have been edited in the period since the print edition are very similar to the ones described above. All the available citations have been organized and placed in the entry structure. For many headwords selected citations have been marked with three bullets (●●●). These show citations that the editor of the relevant entry has highlighted to demonstrate a particular representative usage of the word (cf. figure 3). In addition to the keyed-in citation examples, the scanned slips are also visible and the extra information they may contain is accessible in this way.

accessible on the ONP website but is no longer updated.



ONP: Dictionary of Old Norse Prose

kast *sb. n.* [: ·-]

Search/filter 27 citations...

I.

1) *kast (m. sten, etc.)* // *stone-throwing*

Um kast

Gísli tekr upp stein einn ok kastar ut i holm

••• Sóck i þai skipit niðr til grvnz ok rak s

2) *overbordkastning (af ladning)* // *throw*

kast skal fara eptir fear magne en eigi at mar

••• vm kast a skipi ef ofhlaðit er

kast skal fara eptir fjármagni, en eigi at man

3) *det at lægge fiskenet ud* // *casting of a fishing net*

þa eignazt þeir silld alla er ⸢kast [*var.* varp *Holm perg 32 4° "Gr", etc.*] amu — *Landl 137*

4) *terningekast* // *throw of dice*

••• Enn ero þeir lutir er þu scalt sva flyia oc varaz sem fiandan sialvan. þat er ... portkonor oc þrétur oc ⸢kast [*var.* cast *Kgs 5*] oc viðlogvr þuiat af þessom ... *KgsFragm 130*

ONP: Dictionary of Old Norse Prose

kast *sb. n.*

var. Gísl 37⁶ : NKS 1181 fol^x "S"

NKS 1181 fol^x (c1775-1800) 53v-70r — *Gísla saga Súrssonar*

(Finnur Jónsson 1929, 1-73)

1) *kast (m. sten, etc.)* // *stone-throwing*

Gísli tekr upp stein einn ok kastar ut i holm ... Enn þat var einskis mannz ⸢at inna [*var.* kast annars *NKS 1181 fol^x "S"*]

kast

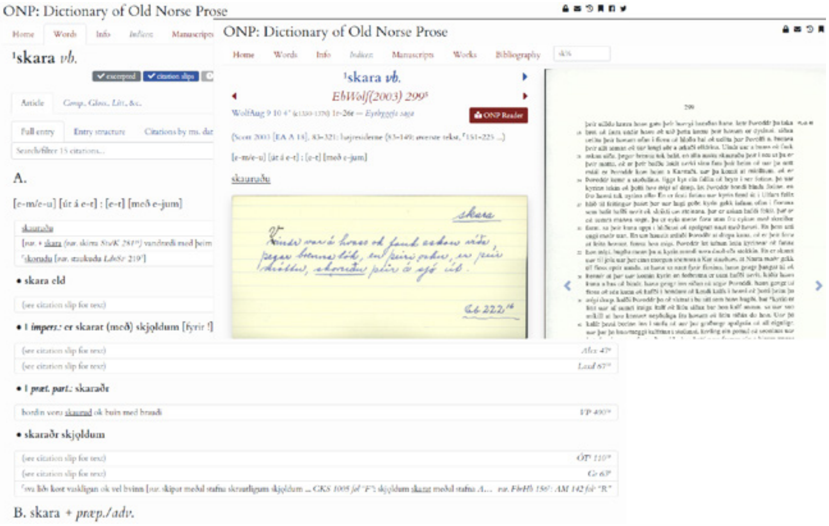
Enn þat var einskis ⸢mannz ⸣at inna
[mannz kast annars NKS 1181, fol^x "S"]

Gísl 37⁶, Var

Figure 3: Screenshots showing an entry that was edited for ONP Online. Three bullets highlight selected citations and each citation is provided with a scanned citation slip as well as a scanned page from the relevant edition (second screenshot superimposed).

4.3 Structurally Edited Verb Entries

After the print edition was put on hold, it was decided that the editing of verbs should be done in two stages. The first stage would include structuring the verbs, i.e., grouping the citations according to structural criteria, e.g., participles, active and middle voice forms as well as verbal clitics. The entry consists of detailed information about the verbal argument structure and the accompanying citations are displayed as scanned slips. However, there is no information about meaning and the citations have not been keyed in (although sometimes the actual form has been entered (cf. figure 4). All the verbs have been structurally edited in such a manner. The second stage of the editing of the verbs has recently commenced and includes semantic structuring of the verbs in line with what is found for verbs in the print edition, but such fully edited verbs have not yet been published online.



ONP: Dictionary of Old Norse Prose

¹skara *vb.*

Search/filter 15 citations...

A.

[e-m/e-u] [úr á e-t] : [e-t] [með e-jum]

skaurodu

[*var.* + skara (*var.* skirra *StvK 281ⁿ*) vandræði með þeim

⸢skoradu [*var.* staukudu *LbSv 219*]

• skara eld

(see citation slip for text)

• I *impers.*: er skarat (með) skjǫldum [fyrir !]

(see citation slip for text) — *Alex 47*

(see citation slip for text) — *Laxd 67*

• I *præt. part.*: skaraðr

borðin voru skaurud ok búin með brauði — *ÍP 400*

• skaraðr skjǫldum

(see citation slip for text) — *ÓT 110*

(see citation slip for text) — *Gr 67*

⸢sva lìðr kost vaskligan ok vel bvinn [*var.* skipat meðal stafna skrautligum skjǫldum ... *GKS 1005 fol "F"*: skjǫldum skarat meðal stafna *A*... *var. FlatIIb 156*: *AM 142 fol "R"*

B. skara + *præp./adv.*

ONP: Dictionary of Old Norse Prose

¹skara *vb.*

EbWolf(2003) 299⁵

WolfAug 9 10 4° (c1350-1370) 1r-26r — *Eyrbyggja saga*

(Scott 2003 [EA A 18], 83-321: hajresiderne (83-149: øverste tekst, ⸢151-225 ...)

[e-m/e-u] [úr á e-t] : [e-t] [með e-jum]

skaurudu

skara

Kindr voru á hvasi ok fundr eaðow víða, þegar birtna tók, en þeira raðu, er þeir drottdu, skorudu þeir á sjó út.

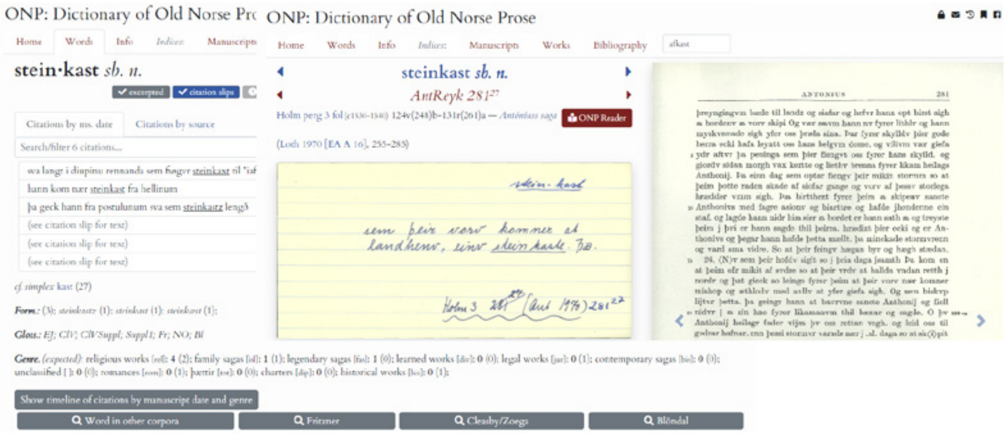
Eb 222¹⁶

299

Figure 4: Screenshots showing a verb entry that has been partially edited to illustrate structural features. There are no definitions and most citations have not been keyed in. Each citation is displayed as a scanned citation slip accompanied by a scanned page from the relevant edition (second screenshot superimposed).

4.4 Unedited Entries

In the first version of the online dictionary, the majority of entries was of this kind, consisting only of headword along with the accompanying scanned slips with no entry structure and only a list of the citations in chronological order (cf. figure 5). As the editing work has progressed this entry type is most common for compound nouns and adjectives, which are the largest groups of words that have not yet been edited. Eventually, as the editing work progresses, these types of entries will gradually be replaced by fully edited and structured entries.



ONP: Dictionary of Old Norse Prc

Home Words Info *Indices:* Manuscripts

stein·kast *sb. n.*

✓ excerpted ✓ citation slips

Citations by ms. date Citations by source

Search/filter 6 citations...

sva langt i diupinu rennandi sem fingvr steinkast til "iaf
hann kom nær steinkast fra hellinum
þa geck hann fra postulunum sva sem steinkastz lengð
(see citation slip for text)
(see citation slip for text)
(see citation slip for text)

cf. simplex kast (27)

Form.: (3): *steinkastz* (1): *steinkast* (1): *steinkost* (1):

Gloss.: *EJ; CIV; CIVSuppl; Suppl I; Fr; NO; Bl*

Genre. *(expected):* religious works [rel]: 4 (2): family sagas [isl]: 1 (1): legendary sagas [fas]: 1 (0): learned works [dic]: 0 (0): legal works [jur]: 0 (1): contemporary sagas [bis]: 0 (0): unclassified [ ]: 0 (0): romances [rom]: 0 (1): þættir [tot]: 0 (0): charters [dip]: 0 (0): historical works [his]: 0 (1):

Show timeline of citations by manuscript date and genre

Word in other corpora | Fritzner | Cleasby/Zoëga | Blöndal

ONP: Dictionary of Old Norse Prose

Home Words Info *Indices:* Manuscripts Works Bibliography aflkast

steinkast *sb. n.*

AntReyk 281²⁷

Holm perg 3 fol (c1530-1540) 124v(248)b-131r(261)a — *Antóníuss saga* ONP Reader

(Loth 1970 [EA A 16], 255-283)

stein-kast

sem þeir vorv kommer at
landhenv, eitt steinkaste. Þa.

Holm 3 281²⁷ (Loth 1970) 281²⁷

ANTONIUS 281

þreyngingvm baedo til landz og siafar og hefvr hann opt birst sigh a bordenv a vorv skipi Og var ssavm hann nv fyrer lithlr og hann myskvnnado sigh yfer oss þraela sina. Þar fyrer skylldv þier gode herra ecki hafa leyatt oss hans helgvm dome. og vilivm var giefa 5 ydr aftvr þa peninga sem þier fengvt oss fyrer hans skyld. og giordv sidan morgh vax kertte og liethv brenna fyrer likam heilags Anthonij. Þa einn dag sem optar fengv þeir mikit stormm so at þeim þotte raden skade af siofar gange og vorv af þesser storlega hradder vmm sigh. Þa birtthest fyrer þeim a skipenv sancte 10 Anthonivs med fagre asionv og biartire og hafde jhonderne ein staf. og lagde hann nidr hia sier a bordet er hann sath a og treyste þeim j þvi er hann sagde thil þeirra. hrædizt þier ecki eg er Anthonivs og þegar hann hafde þetta mallt. þa minckade stormvrenn og vard sma vidre. So at þeir feingv haegan byr og hagh stædan. 15 24. (N)v sem þeir hofdv sigIt so j þeia daga jsamth Þa kom en at þeim ofr mikit af ordre so at þeir vrdv at hallda vndan retth j nordr og þat gieck so leingo fyrer þeim at þeir vorv nær kommer mishop og sthlodv med avllv at yfer giefa sigh. Og sem biskvp lijtvr þetta. þa geingr hann at barrvne sancte Anthonij og fiell 20 nidvr j a sin hne fyrer likamanvm thil hennar og mgle. O þv sæla Anthonij heilage fader vijtu þv oss rettan vegh. og leid oss til godrar hafnar. enn þessi stormvr varade nær j .xl. daga so at sia(þ)pit

Figure 5: Screenshots showing an entry that has not been edited. There is no entry structure and most citations have not been keyed in. The citations appear in chronological order. Each citation is displayed as a scanned citation slip accompanied by a scanned page from the relevant edition (second screenshot superimposed).

4.5 The Ideal Entry

Only the first two types of entries discussed above (4.1 and 4.2) are representative of the ideal form of all entries once the editing work has been completed. An ideal entry in the ONP should include the following features:

- The headword in a cardinal form in normalized orthography.
- Morphological information (inflectional pattern and verb conjugation) based on texts (mainly the actual examples found in the dictionary database).
- Semantic tree.
- Two target languages: Danish and English.[3]
- References to foreign parallel texts (esp. Latin).
- Keyed-in citations with non-normalized orthography, i.e., the orthography of the relevant manuscript or scholarly edition is rendered as closely as possible, with frequent use of special characters.
- Detailed system of sigla indicating not only reference to an edition but also the actual manuscript for each section of the text (in some cases different manuscripts are used within the same edition).
- Syntactic information (especially verb complements and prepositional use).
- Phrases and collocations.
- References to glossaries and, where relevant, references to secondary literature.
- Scanned editions and/or links to images of manuscripts.

Currently, about half of the entries in ONP Online have all or most of the features listed above. For the other half, the editing work is ongoing and only several of the features are already available. In addition to the features listed here ONP Online provides the users with various innovative ways to access the data from the dictionary (cf. e.g., Wills & Johannsson 2019) as well as links to various secondary sources, such as earlier dictionaries, electronic text editions and scanned manuscript images.

5 The Current Study

Since the textual material is known and clearly defined in space and time, we would expect the list of lemmas to be stable as there is in principle no new material being added to the text corpus. However, it turns out that the lemma list of the dictionary continues to evolve with some words being removed as well as new words being added. In its current form the lemma list consists of about 90.000 items and the database contains about 800.000 example citations. About 65.000 of these items are associated with citations, the other items on the lemma list are various types of secondary headwords, which do not contain any examples of usage.

In order to better understand the dynamics of the lemma list and how new items are added to it we conducted a small study. We decided to limit the study to a ten-year period from 2010-2019. As all the lemmas in ONP are organized in a database and each item in the lemma table is assigned a number, it was relatively easy to figure out the additions for the defined time period. After we had filtered out obvious mistakes and entries that were old but had received a new number, we were left with 3789 new items. Most of those items turned out to be secondary headwords of the type that do not contain any citations. Of those there were 2855 references to alternative forms or side forms as well as 194 items, which were suffixes, affixes and second members of compounds. Additional 153 items were different types of secondary entries without citations.

[3] The entries that also appeared in the print edition have two target languages. The entries that only have been published online are still mostly monolingual (either in Danish or English), but work is ongoing to add the other target language where it is missing.

We were then left with 586 new words that had some citations associated with them. Not all of these could be classified as standard headwords as some secondary headwords also have citations. We found that out of the 586 words with citations 42 were outside the scope of the dictionary. Further 100 were labeled questionable and we decided not to account for them further, as these headwords are inherently problematic. The remaining 444 words are standard headwords and can be divided into eight groups:

Compounds added from examples of simplex words (208 examples)
This is the largest group of words, as could be expected, since in Old Norse, as in all Germanic languages, compounds are a very productive word category. Besides the morphological variation in the manuscript material variable factors are whether a compound has been written as one or two words and whether phonological changes have occurred at the juncture of the two elements of a compound. These variables can reflect individual scribal traditions but also the degree of lexicalization of a given compound (cf. Bakken 1995: 170 ff.), or in other words the extent to which it was considered a semantic unit. An example of newly added compound is the hapax *þurfandahjǫlp* 'help for the needy', a nominal compound the first element of which is a noun in the gen. pl. This word has been added to the lemma list while editing the simplex *hjǫlp* 'help', as it has been considered a neologism created on the basis of the foreign Latin model *Auxilium Egentium*. There are also examples of adjectival compounds such as *nýgerðr* 'newly done', and *nýklyppðr* 'newly shaved', which are formed by the adjective *nýr* 'new' + participle. We also find examples of compound verbs, which are often prepositional verbs with occurrences where the conjugated form has a preposition as the first element, for instance *upplíða* vs. *líða upp*, *upptelja* vs. *telja upp*, or *viðkennask* vs. *kennask við*.

Homograph reorganization (43 examples)
Another significant group of new words is the result of homograph reorganization. All in all, we found 42 examples of a new word that had been added to the lemma list where an identically spelled counterpart already existed. In most cases, this is a result of an editing process of a particular headword where either morphological or semantic evidence has suggested that the examples should be divided up between different homonyms that fulfill the criteria for an independent headword. There are many examples here of verb forms that show the characteristics of a different conjugational pattern than a more established type. An example would be the verb *lúka* 'close' which usually is a strong verb with a vowel change in the root, present form *lýkr* 'closes', but ONP has recorded one example with a present form *lúkar*, which indicates a weak verb conjugation and has therefore given rise to a "new" weak verb *lúka*^{2}. Slightly different are cases like the noun *slím* which most commonly means 'slime' but is also found in the meaning 'hindrance', which seems to be of different origin and has given rise to a new homographic headword.

Words from related words (70 examples)
Another group of words has been labeled as originating from related words. This group is similar to the homograph group and most of the examples are explained in a similar manner as resulting from editing an already existing word. The related words are usually a different kind of word formation, either with a different suffix or different inflectional pattern. We can take as an example the neuter noun *tagl* 'tail of a horse' which has given rise to a new word *tǫgl*, a feminine noun that shows a different vowel development, but means the exact same thing. The only example of this new word was erroneously filed under the neuter noun, but after all the examples were categorized as part of the editing process the example was found to represent a different noun class that could only be accounted for with a new headword. This word was subsequently added to the lemma list.

Words added from similar words (21 examples)
Yet another group of words has been labeled as arising from graphically similar words, that are probably not related. The inception of these kind of words is very similar to the category above, except that the words are not related, e.g., the word *firnska* is established in relation to the editing of a similar word *fíflska* 'foolishness' and seems to mean the same thing.

Words added from dissimilar words (5 examples)
In this small group we find a few words that have been added to the lemma list in relation to work on completely different words, where the dictionary editor has come across them. The only way to determine this for sure is when the editor in question has written a note about how this came about. For example, we find the word *vábein* of unclear meaning, where a note has been added that this word is established in relation to work on the word *kvikvendi* 'living thing'.

Misplaced words (9 examples)
This is a small group of nine words that were added to the database after a batch of old citation slips was found in a drawer that had been used in conjunction with the editorial work on the first print volume. Most of these words are compounds where the word *altari* 'alter' is a member, e.g., *formessualtari*, *guðsmóðuraltari*. These slips were discovered by coincidence and seem to be a result of a filing error.

Newly excerpted words (46 examples)
This group mostly contains examples of words that have been discovered in new scholarly editions of lesser-known manuscripts. Even though most Old Norse prose texts have been published, there still remains a large body of manuscripts of various textual significance that has not yet been thoroughly accounted for. The words in this group are found mostly in small fragment texts that have recently been published, such as *hógværisandi* 'spirit of modesty' taken from a 2018 edition of a prayer text. In some cases, a new edition gives rise to a reclassification of a known example,

which requires the establishment of new headword, e.g., *huggøði* > *hugøði* in a new edition of a bishop's saga from 2018.

Unclear (42 examples)
Sometimes it is impossible to figure out why a particular word was added to the list of lemmas. It is likely that in most of such cases the dictionary editors have simply come across an interesting word in relation to their work on the rest of the vocabulary and subsequently have decided to add it to the database. It is therefore probable that most of these have a similar history as the group of words added in connection with work on unrelated words and should perhaps be counted with them.

6 Results

The results show that an overwhelming majority of the additions to the lemma list is a consequence of editorial work on other related or unrelated words where (re)evaluation of textual evidence has brought to light new independent headwords. In most cases, the additional lemmas are compounds that the editors came across when editing the simplex form of one of the members of the compound. Another significant contributive factor is the reconsidering of morphological forms and homographic variation. Only a small portion of the words in question are completely new words, which have been overlooked in previous lexicographic descriptions of the language or have been found in newly reevaluated text material.

7 References

Bakken, K. (1995). *Leksikalisering av sammensetninger. En studie av leksikaliseringsprosessen belyst ved et gammelnorsk diplommateriale fra 1300-tallet*. Avhandling for graden doctor artium. Oslo: Universitetet i Oslo.

Bernharðsson, H., Haugen, O.E. & Berg, I. (2019) "Normalisation." Ch. 10 of *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources*. Gen. ed. Odd Einar Haugen. Version 3.0. Bergen: Medieval Nordic Text Archive. Accessed at: http://www.menota.org/handbook.xml [15/06/2021].

Fritzner, J. (1886, 1891, 1896): *Ordbog over Det gamle norske Sprog 1–3, rev. udg*. Kristiania: Den norske forlagsforening.

handrit.is. An online catalogue over Icelandic and Nordic manuscripts. Accessed at: https://handrit.is [15/06/2021].

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*. Copenhagen: University of Copenhagen, Faculty of Humanities, pp. 250-258.

Johannsson, E. & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition. In *Abel, A. et al. (eds.) Proceedings of the XVI EURALEX International Congress: The User in Focus, 15-19 July 2014*. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 169-179.

Johannsson, E. & Battista, S. (2016). Editing and Presenting Complex Source Material in an Online Dictionary: The Case of ONP. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 6-10 September 2016*, Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 117-128.

Johannsson, E. & Battista, S. (2020). Brugere og forskellig anvendelse af ONP. In C. Sandström et al. (eds.) *Nordiske Studier i Leksikografi 15:* Rapport fra 15. Konference om Leksikografi i Norden, Helsinki 4.-7, pp. 152-161.

Jónsson, F. (1931). *Lexicon poeticum antiquæ linguæ Septentrionalis / Ordbog over det norsk-islandske skjaldesprog*. Original edition by Sveinbjörn Egilsson. 2nd ed. Copenhagen: Kongelige nordiske oldskriftselskab.

ONP = Degnbol, H., Jacobsen, B.C., Knirk, J.E., Rode, E., Sanders, C. & Helgadóttir, Þ. (eds.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose*. ONP Registre (1989). ONP 1: a-bam (1994). ONP 2: ban-da (2000). ONP 3: de-em (2004). ONP Nøgle/Keys (2004). Copenhagen: Den Arnamagnæanske Kommission.

ONP Online. Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose. Accessed at: http://onp.ku.dk [20/03/2021].

Wills, T. & Johannsson, E. (2019). Reengineering an Online Historical Dictionary for Readers of Specific Texts. In Kosem, I. et al. (eds.) *Electronic lexicography in the 21st century: Smart lexicography: Proceedings of the eLex 2019 conference*. Brno: Lexical Computing CZ, pp.116-129.

Stereotypes and Taboo Words in Dictionaries from a Diachronic and a Synchronic Perspective – The Case Study of Croatian and Croatian Church Slavonic

Lazić D., Mihaljević A.

Institute of Croatian Language and Linguistics, Old Church Slavonic Institute
dlazic@ihjj.hr, amihaljevi@gmail.com

Abstract

The paper deals with the lexicographic treatment of derogatory and sensitive vocabulary, in particular vocabulary related to social groups, in historical and contemporary Croatian (and Croatian Church Slavonic) dictionaries. The analysis of the dictionary data, motivated by the insights into the relation between dictionaries and society, is conducted to show how dictionaries reflect the worldview of the time, explain the diachronic development of the lexicographic approach to sensitive content, and propose improvements to contemporary descriptions based on social awareness. For that purpose, the treatment of selected lexical items from the following domains is presented: male and female, sexuality and taboo, ethnicity. It is shown that there is a clear distinction in worldview and lexicographic approach between historical and contemporary dictionaries, which is facilitated by the fact that contemporary dictionaries have to balance between political correctness and the corpus. However, the examples given in this paper show that there is still room for improvement.

Keywords: social stereotypes; offensive language; critical lexicography; historical lexicography; Croatian

1 Introduction

Description of derogatory and sensitive vocabulary – for instance, swear words, vulgar expressions, taboo words, etc. – has always presented a challenge for lexicographers. A type of vocabulary that can be perceived as offensive is the one related to social groups, such as ethnic, religious, gender, age, etc. The offensiveness of such vocabulary may stem directly from its meaning and the intention of the speaker to say something negative about someone or something, or indirectly from stereotypes and prejudices about a group they are grounded in (Schutz 2002: 638).

In recent years, the following factors have to a greater extent shifted such vocabulary into the focus of lexicographers' interest:

1) Political correctness has become an important topic in society and consequently in linguistics and lexicography (cf. Atkins & Rundell 2006: 422-430; Cloete 2014). Social awareness regarding gender equality, marginal social groups, different nationalities and religions, people with disabilities, etc. has changed considerably (cf. Allan & Burridge 2006; Mills 2008; Wodak & Benke 2000; Talbot 2005). This has sometimes led to the process of euphemisation and disphemisation and a constant shift of attitude towards certain words or expressions (a well-known and often-quoted example are words like *Negro*, *Black*, *Afro-American, African American*).
2) Modern dictionaries are usually based on large computer corpora. With higher availability of such corpora, the corpus approach (cf. Tognini-Bonelli 2001) has become the norm in lexicography as an objective approach to language description. However, since actual language usage as attested in corpora is not always polite and politically correct, the question has arisen how sensitive vocabulary, especially vocabulary related to social groups, can be described in a dictionary that aims at being both descriptive and socially responsible.

In addition to being potentially offensive, the vocabulary itself, its usage, as well as its lexicographic description often reflect stereotypical views and values which are culture- and time-specific, and in that sense, the dictionary material can testify to the worldview of a certain society and time.

In this paper, the lexicographic treatment of several sensitive groups of lexical items in historical and contemporary Croatian Church Slavonic and Croatian dictionaries is presented in order to:

1) show how dictionaries reflect the worldview of the time;
2) explain the diachronic development of the lexicographic approach to such content;
3) propose improvements and strategies that could be applied in a modern, socially responsible dictionary.

2 Key Concepts and Previous Research

Sensitive lexical items can be defined as lexical items that have "strong connotative values and derogatory implications" (Cloete 2014: 482). The notion of sensitive vocabulary encompasses a diverse group of lexical items. Harteveld and van Niekerk (1996: 382, 385, 387, 389, 391) proposed the following categorization: 1) racist lexical items, 2) sexist terms and sensitive lexical items which indicate stigmatized sexual phenomena, practices, and preferences, 3) sensitive lexical items which indicate stigmatized physical or mental conditions and phenomena, 4) sensitive lexical items within a social, political, and religious structure, and 5) obscene and vulgar lexical items, abusive language, and swear words. Since these categories differ in the degree and nature of their sensitivity and offensiveness, a slightly different lexicographic treatment for each category was suggested.

Offensiveness is a concept closely related to sensitivity. Jay (1992: 160-161) defines offensiveness as denoting "the degree to which a certain word or concept possesses negative or aversive properties", i.e. the degree of negative content. Very offensive words have the potential of becoming taboo words, words we refrain from using. In that sense, sensitivity and offensiveness can be regarded as quasi-synonyms, although offensiveness can imply a higher degree of negativity, thus denoting a somewhat narrower concept. Janschewitz (2008: 1067) relates offensiveness to the reaction of a person who hears (or reads) a word, the extent to which they perceive it as "personally offensive or upsetting". Tabooness, on the other hand, is the extent to which a word is "offensive and upsetting" in society in general. Sometimes the term *offensive* is differentiated from the term *derogatory*. The former refers to the reaction of the listener or reader to a negative content, and the latter the intentions of the speaker or writer to express a negative attitude towards the referent (Norri 2000: 77). Since offensiveness can be caused unintentionally and not necessarily by a derogatory way of speaking, in this paper, we will mainly use the term *offensive* and the broader term *sensitive*, unless we have the intention of the speaker in mind.

Dictionaries have often been criticized for including offensive vocabulary as well as portraying certain social groups in a stereotypical and/or negative manner, thus codifying and strengthening prejudices that might exist in the society.[1] Having that in mind, a lexicographer can feel urged to omit sensitive content or change linguistic facts, an approach that can be criticized as falsifying reality. However, there is an overall agreement that a dictionary should reflect the real language usage of a certain period (Cloete 2014: 482). Rather than omitting sensitive items that meet the inclusion criteria (e.g. frequency in the corpora), they should be labelled and described properly. In other words, linguistic and lexicographic facts should be distinguished (Bratanić 2005: 39). While language usage does not have to be correct and should be described as such, its lexicographic treatment should be guided by social awareness and should not contribute to strengthening and maintaining the inequality which might be reflected in the language.

Among the elements of lexicographic description which have been discussed with regard to vocabulary sensitivity are the following (Harteveld & van Niekerk 1996; Cloete 2014): 1) choice of headwords, 2) usage labels, 3) metalanguage, 4) references to semantically related items (synonyms, antonyms), where referring to offensive items can be a problem, 5) expressions containing sensitive items, and 6) the choice of illustrative material, such as collocations, editorial usage examples, and citations. Moreover, the lexical treatment of such items will depend on (Cloete 2014: 482): 1) the type of dictionary (e.g. the approach is likely to be different in a school or learners' dictionary and a dictionary aiming at adult native users), 2) the category of a sensitive item, 3) attitudes within a certain community.

As mentioned above, sensitive items can vary in their degree and nature of sensitivity and/or offensiveness (cf. Coffey 2010: 1278-1279). Moreover, offensiveness can be context-dependent as an item can be offensive in all or only in some of its senses (cf. Harteveld & van Niekerk 1996: 383). Schutz (2002: 638) pointed out two aspects of offensiveness, direct offensiveness originating from the speakers' intention to say something negative about someone or something, and indirect offensiveness caused by the stereotype a lexical item is grounded in. Thus, he discriminates between directly offensive items (e.g. *nigger*), indirectly offensive items (e.g. *Dutch treat*), and both directly and indirectly offensive items (e.g. *Jew*, *unmanly*). All these aspects should be kept in mind when describing the usage of an item in a dictionary.

Until now, numerous studies have been conducted in the field of lexicographic treatment of sensitive vocabulary and the social aspect of lexicographic work. They have analysed different grounds of discrimination, such as gender (Fournier & Russel 1992; Russel 2012; Moon 2014, etc.), ethnicity (e.g. Moon 2014), age (e.g. Moon 2014), or illness and disability (e.g. Norri 2019). Some studies regarded certain elements of lexicographic description, such as usage labels (e.g. Norri 2000), or examples (e.g. Fjeld 2015, for Nordic dictionaries). Challenges and lexicographic choices regarding social sensitivity in ongoing dictionary projects (Danish and Swedish) are explained in Jensen et al. (2018) and Petersson & Sköldberg (2020). Additionally, several studies have been conducted on Slovenian lexicography (Gorjanc 2004, 2005; Trojar & Žagar Karer 2013). Few studies have analysed the lexical items connected to sexuality in dictionaries compiled before the 20th century (Dykstra 2006; Schweickard 1997; Lebsanft 1997; Radtke 1986). In the Croatian context, the research has merely focused on the presence of gender inequality and gender stereotypes in dictionaries (Bratanić 2005; Dakić 2017; Pišković 2017), while no studies have concerned users' reactions and expectations regarding the socially sensitive content in dictionaries. Historical Croatian dictionaries have up to now not been studied from the social perspective.

3 Corpus and Methodology

In this paper, the analysis of the lexicographic presentation of the sensitive content is conducted from the point of view of historical and contemporary lexicography. Similar entries and entries from the same domains are analysed in historical and contemporary Croatian dictionaries to determine changes in the lexicographic approach to such content. The corpus for our analysis consists of the following dictionaries:

1) historical dictionaries: Vrančić (1595), *Dictionarium quinque nobilissimarum Europae linguarum*; Kašić (around 1600), *Hrvatsko-talijanski rječnik;* Mikalja (1649), *Thesaurus linguae Illyricae* – the first modern-type Croatian dictionary; Habdelić (1670), *Dictionar ili reči slovenske;* Jambrešić (1742) *Lexicon Latinum interpretation Illyrica, Germanica et Hungarica locuples;* Stulli (1801), *Lexicon latino-italico-illyricum;* (1805),

[1] One of the famous examples has been the definition of Bangkok as "a place often mentioned where there are a lot of prostitutes" (*The Herald* 1993). In Nordic context, the inclusion of the entry *grønlænderstiv* ('drunk as a Greenlander') in a Danish dictionary is an example that has gained a lot of publicity (Farø & Jensen 2018: 219). A Croatian example that has been criticized is the definition of woman as a being opposite of a man (*Libela* 2013). However, in Croatia, the criticism of that kind is not very common and has almost exclusively come from the activists and minority groups themselves, and rarely from the general public.

Rjecsosloxje; (1810), *Vocabolario italiano-illirico-latino;* Šulek (1860), *Njemačko-hrvatski rječnik*; the dictionary of the Croatian Academy of Sciences (= ARj) that was compiled from 1880 to 1976 and was based on the corpus from the 12th century to the contemporary period; *Slovník jazyka staroslověnského* (= *Slovník,* 1966-1997), a Church Slavonic Dictionary that is being used as the basis for the RCJHR; *Rječnik crkvenoslavenskoga jezika hrvatske redakcije* (= RCJHR, *Dictionary of the Croatian Redaction of Church Slavonic*), an ongoing project of compiling a dictionary of Croatian Church Slavonic based on the corpus from the 11/12th century to the 16th century;

2) contemporary dictionaries: the printed dictionary *Rječnik hrvatskoga jezika* compiled by Šonje (= RHJ; 2000), *Hrvatski jezični portal* (= HJP; http://hjp.znanje.hr) – currently the only Croatian comprehensive monolingual open-source web dictionary and thus presumably the one most frequently used by adult native speakers, and *Veliki rječnik hrvatskoga jezika* (= VRH; 2015) – the most recent comprehensive monolingual dictionary, available in a printed and a digitized version. Only general dictionaries for adult native speakers have been analysed; special-purpose dictionaries, such as school dictionaries, are beyond the scope of this study.

The lexicographic treatment of sensitive lexical items will be illustrated by selected entries from the following domains: male/female (sex distinction), sexuality and taboo words, ethnicity. The analysis will focus on the selection of headwords, definitions, usage labels, choice of examples, collocations, idioms, and pragmatic notes.

4 The Analysis

4.1 Male – Female

4.1.1 Historical Dictionaries

If not stated otherwise, the examples and definitions are from ARj, which has the greatest number of examples and the most elaborate definitions. Older Croatian dictionaries tend not to be socially sensitive towards the stereotypes regarding female identity. In ARj, women are defined as persons whose organism is designed to give birth – “osoba kojoj je organizam udešen za rađanje djece”. In some definitions, it is stated that the word *woman* can be used to denote a man acting like a woman, i.e. someone acting cowardly – “oznaka ili obiļežje čovjeka koji se vlada kao žena, koji je plašļiv, kukavica”. The adjective *female* is defined as the opposite of male in strength, courage, and boldness. It is stated that the adjective denotes someone weak, soft, limp, timid, fearful, and cowardly – “suprotan muškom u hrabrosti, smjelosti i snazi, tj. slab, mek, mlitav, bojažļiv, strašļiv, kukavički”. In ARj, it is stated that the words derived from the word *woman* can sometimes be used to denote a man if he is fearful as a girl, soft, sensitive, spoiled, weak, and timid, e.g. “u prenesenom smislu o muškarcu, koji je bojažljiv kao djevojka”; “muškarac, koji je sličan ženi, koji radi i vlada se kao žensko, koji je mek, osjetļiv kao žensko, kome žena zapovijeda, koji voli ženske i trči za ńima”; “kao žena, ženski, t. j. razmažen, slab, bojažļiv činiti da tko bude mek i osjetļiv kao ženska”; “mek, slab kao žena”. Some older dictionaries list the most common collocations with the words denoting a woman, such as: *ill-tempered*, *evil*, *wild*, *quarrelsome*, *disgraced*, *dishonorable*, *dishonest*, *insatiable*, *lewd*; *harlot*, *sinner*, *adulteress* – “zloćudna, zlobna, zlopametna, divlja i grda, karljiva, osramotjena, nečastna, nepoštena, razpuštenica, zla, bludnica, zlica, nečastnica, grešnica, hotnica, priljubovca, mrska, pogana, nezasitna, nepoštena, nikad sita…”. In some dictionaries (e.g. Mikalja), a lot of collocates are connected to the meaning of being married or pregnant. Older dictionaries often give examples, which are also full of stereotypes: a good woman gives birth only if she is married; a new-born is fortunately male; women do not have the brain of a man, but of a child; a man is the head, and a woman is the grass; women have long hair, but a short brain; a woman should be quiet when a man speaks; dogs should bark, and women should be quiet; you should not trust a woman because she changes like the Moon; a man should be a hero, and not act like a woman – “Da bude rodila ne budući za mužem, ne bi bila držana za ženu dobru”; “da j' dite na su sreću muško”; “Mi žene … neimamo pamet mušku na djetsku”; “čovjek je glava, a žena trava”; “Žena je dugokosa, a kratkoumna”; “Žena jezik za zube kad muž govori”; “Kučka nek laje, a žena nek muči”; “Ne vjeruj ženi, er se kako mjesec mijeni”; “Ta nemoj me ženski udarati, već me udri, čim s' junaci biju”. In the dictionaries, it is often stated that there are jobs and duties fit for a woman and those fit for a man – “poslovi su odijeļeńi: čovjek u poļu, a žena u kući”; “ti žeńkari lepo šiju … i druge ženske poslove rade … samo da ne traže muškaraca”; “ženskadija pravi večeru”; “To u nas radi ženskadija”; “o muškarcu, koji se ponaša kao ženska glava i zna obavļati ženske poslove”. In ARj, domestic violence is depicted as normal: he who does not hit his wife is not a man; you should hit a horse and a woman for them to be obedient – “Ko ženu ne bije, on čovjek nije”; “Ženu i konja udri, ako želiš, da su ti pokorni”. In Šulek’s dictionary, it is stated that the woman is more cunning than the devil – “žena je lukavija od vraga”, and in Stulli’s, that she is the worst beast – “žena je na svijetu najgora zvir”.

Some gender stereotypes are also visible in (Croatian) Church Slavonic Dictionaries and are mostly influenced by the fact that their corpora mostly consist of Biblical and other religious texts. In *Slovník* and RCJHR, women are depicted as weak, they are often shameless, and they brought destruction to Adam, Joseph, and David. It is stated that men should not listen to women and that women often present themselves as much prettier than they are. It is also stated that the devil often appears in the form of a woman – “muži že takožde sь svoimi živuĉe ženami êko nemoĉnêisei veĉi žen'scêi vzdajuĉe čьst'”; “ženom' bo adamь iz raê spuen' bi ženom bo pr(a)vdni osip' zatv(o)ren' bê v' tamnici ženomь d(a)vidь uriju stv(o)ri ubiti”; “zač ti posluša glasa ženi tvoee ku ti dah' pod' tvoju oblast' i na tvoju volju”; “o gorko i čermernotim' ženom' ke lice svoê pomazuju i lipše se čine nere ih' e bog' učinil'”; “mnogo bo krat' dêvl' prihoêše k nemu va obraze žen'sceem'”.

4.1.2 Contemporary Dictionaries

Even though the content of the entries related to women and men in contemporary dictionaries clearly shows how the worldview has changed over time – for example, women are no longer depicted as having the primary role of giving birth[2] nor is domestic violence promoted – some stereotypes persist, and they regard, in the first place, what are thought to be typical male/female characteristics, and to some extent the attitude to sexuality of the respective sexes.

The stereotypical view and inequality can be observed in the first place in definitions, collocations, and examples. The following examples from HJP speak for themselves: *male* (*muški*, adj.) – one of the senses: *worthy of a man, a real man* – “dostojan muža, pravog muškarca”; *“male hand”* (*muška ruka*): *a man who in a household does the typical work which is not suitable for a woman* – “muškarac koji u kući obavlja tipične poslove koji nisu za ženu”; *manly/masculine* (*muževan*, adj.): *having all qualities of an adult or an honourable man [manly appearance]* – “koji ima sve odlike odrasla muškarca ili časna muža [*muževna pojava*]”; *“male old wife”* (*muška baba*): *a man with some female characteristics (e.g. talks a lot and the like)* – “muškarac s nekim ženskim osobinama (mnogo priča i sl.)”. In the examples, the positive attitude towards typical masculine traits is emphasized and they are not questioned. Similar sexist definitions can also be found in RHJ, e.g. *“female head”* (*ženska glava*): *a woman with her peculiar way of thinking* – “žena sa svojim osebujnim mišljenjem”.

If the definitions are compared with those in VRH, it is obvious that the latter are more neutral and less emotionally coloured, but the social construct of a typical masculine nature is taken as given: *manly* (*muški*, adv.) – *like a man, in a manner of a man [to act / say / drink / hit someone like a man]* – “kao muško, na način muškoga [*muški postupiti* / *reći* / *piti* / *udariti koga*]”; *manly/masculine* (*muževan*, adj.): *showing characteristics of the male sex [masculine look; masculine attitude; masculine appearance]* – “koji pokazuje odlike muškoga spola [*muževan izgled* / *nastup*; *muževna pojava*]”; expressions *“male hand”* and *“male old wife”* are not listed in the dictionary.

Social stereotypes of typical male and female characteristics have been lexicalized in the adjectives *ženskast* and *muškobanjast*, which denote femininity in men (*womanish*) and masculinity in women (*mannish*) respectively, and are normally used disparagingly or mockingly. They can thus be regarded as both directly offensive to the person or the group they are used for, and indirectly offensive to women and men because of the stereotype they are grounded in. However, none of this is indicated in Croatian dictionaries – the words are not even labelled as derogatory/offensive. Moreover, the examples given in VRH reveal further stereotypical view on certain professions which are socially not regarded as masculine/feminine: *a masculine policewoman* – “muškobanjasta policajka”; *For his feminine and somewhat unnatural movements, he could be a ballet dancer* – “Po svojim ženskastim i pomalo namještenim kretnjama mogao bi biti baletni plesač.” It can be argued that such stereotyping is not necessary for describing the meaning of the headwords, neither is it the typical context of their use[3], and should thus be avoided.

Collocates given for words from the domain men and women often differ in dictionaries, those for women often being related to beauty and emotions and those for men expressing physical strength. For example, in VRH a woman is *energetic*, *young*, *pretty*, *unhappy*, while a man is *brave*, *unknown*, *real*, *threatening*, *average*, *middle-aged*, *loved*. On the other hand, some of the examples in VRH show that an attempt has been made to make it more inclusive and up to date, e.g. in the entry *women* (*žena*), the following examples have been included: *fighting for women's rights* (“borba za prava žena”); *a man trapped in a woman's body* (“muškarac zarobljen u tijelu žene”).

4.2 Sexuality

4.2.1 Historical Dictionaries

Words related to sexuality[4] are rarely included in older dictionaries. Sexual relations outside marriage and those with the same sex are described as unnatural and sinful – “objašńava to starim vjerovańem po kome se smatralo, da su tjelesni odnosi između muža i žene i začeće religijski nečisti i griješni”. Sodomy is defined in ARj as unnatural sexual intercourse and in the usage example, it is stated that a man or a woman who has committed sodomy should be burned – “sodomija **–** nenaravno spolno općeńe – sodomija jest, kada muški polь ima čińenje s muškim spolom”; “kadano čovjek sgriješi z' ženom naopako, to jest učini sodomiju”; “sodomski – sedmi grih je proti naturi aliti sodomski”; “ako bi se tko naša u grihu nepodobnu, ča se zove grih sodomski, ali bi bila muška glava ali ženska … ima se sažgati”. Prostitutes and mistresses are described as *unclean, sinful, wicked*, and it is stated that they will not go to paradise – “Nisi čista, da bludnica”; “sagriješiti s ženom bludnicom jest blud preprost”; “bludnici i bludnice … biti će polivani gorućim paklom”; “bludnici ne će ulisti u raj”. Words denoting homosexual men are not attested in older dictionaries and vocabulary related to sexuality and relationships is always described traditionally, having the heteronormative relationships in mind.

4.2.2 Contemporary Dictionaries

Due to social changes, today we are more familiar with various forms of sexual orientation and preferences, so it is not surprising that the associated vocabulary is more present in contemporary dictionaries than in historical ones. No signs of reluctance to include such vocabulary have been observed in analysed dictionaries, but several other problems can be

[2] An exception is the definition of the word *woman* (*žena*) in RHJ, which has often been quoted in Croatian publications on sexism in language and dictionaries (cf. Bratanić 2005; Dakić 2017): *a human being of opposite sex than a man, who can give birth to children and take the main care of the upbringing and education of children* [“ljudsko biće po spolu suprotno muškarcu, koje može rađati djecu i preuzeti glavnu brigu za uzgoj i odgoj djece”].

[3] The nouns *policajka* and *baletan* do not appear as common collocates of the adjectives *muškobanjast* and *ženskast* in *Croatian Web Corpus – hrWaC* (accessed 26/02/2021).

[4] “Central aspect of being human throughout life and encompasses sex, gender identities and roles, sexual orientation, eroticism, pleasure, intimacy and reproduction.” (European Institute for Gender Equality 2021).

discussed: usage of labels, cross-references, and definitions of headwords related to sexuality.

In contemporary dictionaries, there are several words denoting a homosexual man, varying from neutral through colloquial to derogatory and vulgar. However, sometimes labels are missing or do not correspond to actual usage (usually they are too mild). For example, the word *peder* is labelled only as colloquial, while it can be argued that it is often used disparagingly and can be perceived as offensive. Moreover, unlike vulgar words (discussed in the next section), which are usually not listed as synonyms / cross-referenced within neutral entries, the neutral entry *homoseksualac*, for example, contains several synonyms in HJP and VRH ranging from the colloquial *homić* to the very offensive *dajguz* (literally a "butt giver"), but the stylistic value of the synonyms is not indicated. It can be discussed whether and how the offensive synonyms should be listed within neutral entries. While it can be useful for the user to get a list of similar words to choose from in text production, an uncritical listing of offensive expressions contributes to the negative view of social groups and could cause public disapproval. A possible solution could be to include appropriate labels both when the expressions appear as headwords and when they are listed within other entries.

A traditional, heteronormative view often occurs in descriptions of vocabulary related to sexuality and relationships in general. Even though some improvement can be noticed, e.g. *"male virgin"* (*djevac*) is defined as *the one who is living as a virgin, who has renounced or is deprived of the touch of a woman* – "onaj koji živi u djevičanstvu, koji se odrekao ili je lišen dodira žene" in HJP, but as *a man without sexual experience* – "muškarac bez spolnoga iskustva" in VRH, in many cases, such as in definitions and examples provided for words *girlfriend* (*cura*), *boyfriend* (*dečko*), *lover* (*ljubavnik*/*ljubavnica*), etc. the meaning and usage are described having heterosexual relationships in mind.

A dissimilar social attitude towards female and male sexuality is also reflected in dictionaries, and it can be observed in the first place in the inventory of expressions included in a dictionary and their definitions, which is also more prominent in HJP than in VRH. In HJP, expressions like the following can be found: *secondary virginity* (*drugo djevičanstvo*), defined as *the condition of a married woman who, according to social and economic reasons and customs, has an older husband and is left without an erotic life in her vital years* – "stanje udate žene koja prema društvenim i ekonomskim razlozima i običajima ima starijeg muža i u vitalnim godinama ostaje bez erotskog života"; *"mental female prostitute"* (*duševna prostitutka*), defined in one of the senses as *a woman who takes advantage of a man and keeps him hoping he will be successful; one who promises or gives hope of an intimate relationship she does not intend to get into* – "ona koja iskorištava muškarca i drži ga u nadi da će postići uspjeh; ona koja obećava ili daje nade u intiman odnos u koji ne misli ući". In the examples, female sexuality, age, etc. are portrayed as a means of taking advantage of men, gaining social and financial security, etc. There are no male counterparts in the dictionary and definitions feel outdated and one-sided since there is no comment on the social context they have arisen from or their usage today. Moreover, there is hardly any evidence of their usage in available Croatian corpora (e.g. *Croatian web corpus – hrWaC*; accessed 26/02/2021), so it is not unexpected that the expressions do not appear in the newer VRH. However, some vulgar expressions referring to female sexuality and character are to be found in both dictionaries, for example, the very vulgar *"cold cunt"* (*mrzla pizda*), meaning "a frigid woman" and *"wolf with a pussy"* (*vuk s pičkom*), defined as *a very determined, enterprising, strict woman* – "Vrlo poduzetna, odlučna, oštra ženska osoba". These expressions are also very scarcely attested in contemporary corpora, so it is questionable whether they should be included in a dictionary.

4.3 Taboo Words

4.3.1 Historical Dictionaries

As examples of taboo words[5], we have chosen three word-formation clusters – *kurac* and its derivatives – an offensive word for penis, *pička* and its derivatives – an offensive word for vagina, and *jebanje* and its derivatives – an offensive word for sexual intercourse. Older Croatian dictionaries (e.g. Vrančić and Kašić) generally do not list those words. In Jambrešić's bilingual Latin-Croatian dictionary, the words *mentula* and *penis* are listed, but without their Croatian equivalents.[6] In the entry *penis*, it is stated that the translation can be found in the entry *mentula*, but the definition found there is also in Latin – *membrum pudendum viri*, "shameful male body part". In ARj, although it is a monolingual dictionary, and the entries are in Croatian, some taboo headwords are defined only by their Latin equivalents or have a Latin definition: *kurac – mentula*; *kurat – mentulatus*; *kurcati se – penis vocabulo abuti*; *kurcoglavac – senecio vulgaris L.*; *kurčev – mentulae*; *kurčevit – ut mentula*; *kurčiti se – penem imitor*. Some derivatives have a Croatian definition, but a part of the taboo word is censored: *kura – hyp. …ac, kurcovina – augm. od …cov, kurčekanja – augm. …ac, kurčenje – djelo kojijem se ko …či*. It is interesting to note that the headword is not censored, but the word derived from the same root in the definition is. Sometimes there is a combination of the censored Croatian definition followed by a Latin definition, e.g. *kurcanje – djelo kojijem se ko …ca*, *penis creber usus in loquendo*. In most of these entries, there are no examples of usage, although this dictionary usually gives many examples. In one of the entries where the example is given, the taboo word is also censored: *kurcov – ti si već …cov.* One word is defined in German: *kurcokret – 'ein komisches wort fur celer'*. Words denoting vagina are found only in ARj and are similarly defined. Only Latin translations are given when defining

[5] On the history, definitions, and features of taboo words see Allan & Burridge 2006; Jay 1977, 1992, 2000, 2009; Jay, Caldwell-Harris & King 2008; Janschewitz 2008. The term *taboo words* describes "the lexicon of offensive emotional language. A taboo is a 'ban or inhibition resulting from social custom or aversion' (The American Heritage Dictionary of the English Language, 2000). Taboo words are sanctioned or restricted on both institutional and individual levels under the assumption that some harm will occur if a taboo word is spoken." (Jay 2009: 154). "Taboo words represent a class of emotionally arousing references with respect to body products, body parts, sexual acts, ethnic or racial insults, profanity, vulgarity, slang, and scatology (Jay, 1992, 2000)." (Jay, Caldwell-Harris & King 2008: 83).

[6] On the difference between the two and on Latin sexuality vocabulary in general see Adams 1982.

the following words: *pica – cunnus*; *pičkar – fututor, amans cunni*; *pičkaroš – fututor, amans cunni.*[7] The word *pička* is defined by both German and Latin equivalents: *pička – Scham, cunnus.* The word derived from the same root *pican* is defined in Croatian as the boy who likes to play and be friends with girls and is spoiled – "dječak koji je razmažen i rado se s djevojčicama igra i druguje".

In older dictionaries, male genitalia can also be found as entries *ud* and *udo* ('extremity/limb') with adjectives *dishonest*, *shameful*, *ashamed*, *secretive*, *childbearing*, and genitalia of both men and women as entries deriving from the words denoting shame and disgrace: *sram* and *stid*, usually defined as the thing that women and men hide.

Words derived from the root *jeb-* are found only in ARj and Stulli. In ARj, they are often defined by Latin equivalents, e.g. *jebač – futuens.* Two words have both German and Latin equivalents: *jebac – der hurer, fututor validus*; *jebaonica – das bordell, lupanar*. Some words have Croatian definitions with taboo words censored in the definition: e.g. *jebane – djelo kojijem se ...e*. The word *jebičina* is defined as *augm. od jebica,* without the taboo word being censored. Only one example is given also in its censored form – "Taman laže, vsi mu je.li majku!". The equivalents in other Slavic languages are given in uncensored forms although they are very similar to the Croatian word, while the uncanonical forms of these Slavic words are censored – "usporedi novoslov. jebati… češ. jebati, praes. …am i ...u, polj. jebać, …ę".

4.3.2 Contemporary Dictionaries

The words and their derivatives discussed above appear as headwords in two contemporary Croatian dictionaries – HJP and VRH – while reluctance to include them can be noticed in RHJ. In the latter, only neutral words such as *vagina* (as well as its Croatian synonym *rodnica*), *vulva* (as well as *stidnica*) and *penis* (as well as *udo*) are found, while their vulgar synonyms are not mentioned either within the entry of a neutral headword or as a separate headword. The headword *jebati* exists in the dictionary and is labelled as vulgar, but its description is deficient since only the following definition is provided: *to have sexual intercourse with a woman*. However, the word can denote any type of sexual intercourse, and that has been taken into account in the other two analysed dictionaries, where the gender of those involved is not specified. Thus, it is evident that the definition in RHJ reflects a traditional view by mentioning only heterosexual intercourse and associating men with an active role. Furthermore, other senses of the word *jebati* ('to bother someone; to ignore someone/something'), its usage as a swearword as well as numerous expressions and derivatives it appears in have been omitted in RHJ, so it can be concluded that the lexicographic description is not up to date and that some aspects of language usage – especially colloquial, have been disregarded.

In the remaining two dictionaries, where vulgar words for genitalia and sexual intercourse are listed as well as their various derivatives, their stylistic value is indicated by means of:

1) a stylistic label *vulg.* ('vulgar'), which is found in all entries discussed here;
2) (rarely, non-systematically) an explanation which is:
    a. a part of the definition, e.g. one of the senses of the word *kurac* is defined in the following manner (HJP, VRH): *the word which as a filler often fills a pause in a sentence and which is orthographically expressed by a dash (hyphen) or a comma placed according to the intonation, in texts usually abbreviated to k...* – "riječ kao poštapalica u rečenici često popunja stanku koja se pravopisno izražava crtom (povlakom) ili zarezom postavljenim po intonaciji, u tekstovima obično kraćena k...";
    b. added separately, as additional information, e.g. within parentheses in the entry *pizda* (HJP), after the senses have been listed: *the word is very rude and inappropriate in polite communication, in texts it is usually abbreviated to p...* – "riječ je vrlo nepristojna i neprikladna za iole pristojan način izražavanja, nalazi se u tekstovima obično kraćena p…";
3) cross-references: there is a tendency to include references to neutral entries in the vulgar ones, but not vice versa; however, some exceptions exist, e.g. the neutral entry *penis* includes references to both neutral/scientific *falus*, colloquial *pimpek,* and vulgar *kurac* in VRH. In HJP, only neutral words (*spolovilo*, *udo*) are listed as synonyms (note the difference in the synonyms given in the two dictionaries); similarly, no vulgar entries denoting sexual intercourse (*jebati*, *ševiti*, *fukati*…) are referenced to within either vulgar or neutral entries. It can be discussed whether colloquial and vulgar synonyms should be listed at least within entries of a similar style, if not all of them; it could be a useful information for language production.

4.4 Ethnicity

4.4.1 Historical Dictionaries

In older dictionaries, ethnonyms[8] are often defined depending on their social and historical background. The ethnic groups more closely related to Croatia are described in more detail and with more examples and stereotypes, e.g. in ARj, the Turks are described as warriors and are sometimes described negatively as enemies, non-believers, and liars – "Turci nas su oplinili i požgali"; "u Turčina nigda vire nije"; "laže ka Turčin"; they are also described as people who smoke a lot and drink coffee and wine – "Turci vino piju"; "pije kavu ka Turčin"; "puši ka Turčin". Most stereotypes are attested in ARj in entries derived from the words denoting Gypsies. Gypsies are often described as people who *deceive*, *cheat*, *tell fortune* and *wonder*, *steal* and *lie*, and are *lazy* – "cigančiti – cigančiti je osobito prositi ili iskati navaļujući, ne odstupajući, kao što čine Ciganke"; "ciganiti = varati, prosjačiti"; "svit Cigane vrlo kori od svi ļudi da su gori, jer su lini od kolina i lupeži od starina"; "još ciganski i dlane gledate"; "prijateļa ki ukani, još je gorši neg Cigani". It is stated that calling

[7] On the etymology and the lexemes denoting genitals see Reinhart 1994.
[8] I.e., the names of ethnic groups. For more see Koopman 2016.

someone a Gypsy is an insult – "poruga čovjeku koji laže i vara.", but it is not noted that the definitions of Gypsies as fraudulent are based on a stereotype.

In addition to negative stereotyping, some positive stereotyping can also be found, mostly in the entries connected to Bosnia and Croatia. Bosnian women are *pretty* – "ļepše djeve Bosankińe; oženio sam se Bošńakińom lijepom divojkom", Croatians are *good*, *famous*, *proud*, etc.

4.4.2 Contemporary Dictionaries

In contemporary dictionaries, ethnonyms are approached more cautiously and are either not included at all as headwords or, when they are, they are defined neutrally, e.g. with regard to their geographical origin, and their usage is not exemplified. On the other hand, prejudices and stereotypes tend to be revealed in related words and expressions, such as derivatives.

Prejudices about social groups often give impetus to the development of secondary meanings as well as the formation of derivatives and expressions, which are usually more or less derogatory given the fact that they are often based on characteristics that are considered socially unacceptable or undesirable. They are even indirectly offensive to the social group which the underlying stereotype regards. The example of that is the colloquial/offensive word *Gypsy* (*Ciganin*) and its derivatives like *ciganski*, *cigančiti*, *ciganija*, *ciganluk*, etc., which exist in abundance both in the language itself and in dictionaries and denote something deceitful, messy, or dishonourable either in their secondary or even primary and only sense. However, such words are not always labelled as derogatory, offensiveness to the group is never indicated, and the stereotype is rarely commented on.[9]

In addition to the negative, some positive stereotyping can also be found, as in the expression *Slavic soul*, defined in the following way: *according to the established positive prejudice, a peaceful human nature, a magnanimous person* – "po uvriježenoj pozitivnoj predrasudi nesebična, miroljubiva narav čovjeka, široka duša" (HJP, VRH). Negative stereotyping can both regard the group one belongs to (e.g. expressions *Croatian envy* (*hrvatski jal*) and *Croatian silence* (*hrvatska šutnja*) in HJP and VRH) and other groups, such as neighboring nations (e.g. RHJ, HJP, VRH: *"Bosnian pot"* (*bosanski lonac*) – regarding complicated political circumstances; *srbovati* – defined as *to express Serbian national feeling intrusively* – "nametljivo izražavati srpske nacionalne osjećaje").

In the entries related to ethnic groups, stereotypes are commented on more frequently than in other semantic groups (probably because they are perceived as especially sensitive), even though not very often, for example in the definition of *"Bosnian pot"*: *intricate political circumstances typical for Bosnia (according to prejudices outside of Bosnia)* – "zamršene političke prilike tipične za Bosnu (prema predrasudama izvan Bosne)" or *balkanština*: *primitivism and dishonest actions in public, cultural and political life, which are according to preconceptions in Western Europe considered typical of the Balkans* – "primitivizam i nečasni postupci u javnom, kulturnom i političkom životu što se po preduvjerenjima u Zapadnoj Europi smatra tipičnim za Balkan" (HJP). Sometimes, expressions like *allegedly* are used, as in the example *Croatian silence*: *allegedly the conformism common for Croatian public and politicians* – "navodno uobičajeni konformizam hrvatske javnosti i političara".

5 Conclusion

In historical dictionaries, many stereotypes have been attested, and it is obvious that dictionary compilers were not aware of the potential offensiveness of some entries. Many entries also reflect the worldview from a certain period and prevailing stereotypes. In defining the words denoting women and men, characteristics that are stereotypically perceived are given without hesitation, and many stereotypes can be found in both definitions, collocations, and examples. Being a woman is thus described mostly negatively: women are inferior to a men, they should be beaten, keep quiet, obey their husbands, and give birth to children. The headword inventory connected to sexuality is limited. Words denoting sexual relations are usually omitted, especially in dictionaries from the oldest period (e.g. Vrančić and Kašić). Sexual intercourse is defined as something sinful, and, if it is between two men, unnatural. The intercourse between two women is not commented on. The headword inventory of taboo and vulgar words is also limited. If included, the headwords are usually defined in Latin (and sometimes in German). Although the headwords are not censored, the vulgar words in the rest of the entry usually are. In the entries with vulgar headwords, usually, no usage examples are given. The ethnic groups geographically or historically closer to Croatia are described in more detail. The entries reflect historical relations (e.g. Turks are described as warriors and enemies) or prevailing stereotypes (e.g. Gypsies are described as lying, lazy, thieves, etc.).

In contemporary dictionaries, the stereotypes are more subtle than in historical dictionaries – which implies a change in the worldview over time, but also a change of the lexicographic approach towards socially sensitive content. Moreover, the dictionary material has shown that social awareness is not equally present in the treatment of vocabulary in all domains, which can be due to the fact that some of them (e.g. ethnicity), are perceived as more sensitive than others. An overview of the results of our analysis is presented in Table 1.

Although a progress can be noticed when the contemporary Croatian dictionaries are compared with historical dictionaries, and even within the group of contemporary dictionaries – the newest one being more socially sensitive than the older ones – the examples presented in this paper show that there is still room for improvement. The analysis has revealed the following elements that should be revised: 1) definitions should be checked for subjectivity (*worthy of a real man*), outdated perceptions (*typical work not suitable for woman*), sexism (*woman with her peculiar way of thinking*), etc.;

[9] An exception is the entry *ciganluk* in HJP, defined as *an ugly act of a kind that is according to the prejudice attributed to Gypsies* ["ružan postupak kakav se prema predrasudama pripisuje Ciganima"].

a comment of a stereotype lexicalized in an item can be considered (as it is sometimes done in entries concerning ethnic groups, but rarely other, e.g. *womanish*, *mannish*); 2) definitions and examples should in some cases be more inclusive, e.g. words denoting relationships and sexuality are often approached from a heteronormative perspective; 3) unnecessary stereotyping should be avoided in examples and collocations (*masculine policewoman*), 4) lists of collocations and expressions should be revised and updated, 5) labels should reflect the actual usage (they are sometimes missing or are too mild); both direct and indirect offensiveness could be considered and indicated by labels or other means, 6) synonyms / cross-references should be reviewed – for offensive items, neutral items should be given and offensive items, if listed within other entries, should be labelled. Modern e-dictionaries, often published online, have innovative features which can be useful for discovering and describing socially sensitive content. For example, the absence of space limitation allows for a more detailed description, explanatory notes, more examples to illustrate different contexts, etc. The Internet as a medium makes it easier to edit the dictionary data and enables communication with users, who can provide useful information on potentially sensitive content.

	Historical dictionaries		Contemporary dictionaries	
	characteristics	examples	characteristics	examples
male/female	– typical male/female characteristics mostly taken as given – stereotypes found in definitions, collocations, and examples – being a woman often perceived as negative or inferior to being a man – potentially offensive items usually not labelled as such	– woman – a person whose organism is designed to give birth; someone acting cowardly	– typical male/female characteristics sometimes taken as given – stereotypes found in definitions, collocations, and examples – potentially offensive items sometimes not labelled as such – progress has been made in the newest dictionary	– “male hand”: *a man who in a household does the typical work which is not suitable for a woman* – masculine policewoman – pretty woman vs. brave man – fighting for woman’s rights
sexuality	– the headword inventory is very limited – sexual relations (especially of the same sex and outside marriage) described as unnatural, shameful, and sinful	– sodomy – unnatural sexual intercourse	– the headword inventory is more inclusive – offensive headwords and cross-references are not properly labelled – different view on male and female sexuality – definitions of words denoting relationships and sexuality should be more inclusive	– “male virgin”: *the one who is living as a virgin, who has renounced or is deprived of the touch of a woman* – “mental prostitute”: *a woman who takes advantage of a man and keeps him hoping he will be successful*
taboo words	– the headword inventory is very limited – no Croatian definitions are given – equivalents in Latin (and German) – vulgar word censored	– penis – Latin translation: *mentula* – fucking – the act of f…ing	– listed without hesitation in 2 out of 3 dictionaries – vulgar expressions labelled as such	– pizda (‘cunt’), vulg.: *the word is very rude and inappropriate in polite communication, in texts it is usually abbreviated to p...*
ethnicity	– larger entries with more examples for the ethnic groups historically closely related – stereotypes found in definitions, collocations, and examples	– Turks – warriors, enemies, non-believers…	– treated more cautiously than other domains – stereotypes are sometimes commented on	– “Bosnian pot”: *intricate political circumstances typical for Bosnia (according to prejudices outside of Bosnia)*

Table 1: The treatment of vocabulary from the domains male/female, sexuality and taboo words, and ethnicity in historical and contemporary dictionaries – main features.

6 Sources

ARj = *Rječnik hrvatskoga ili srpskoga jezika* (1880-1976). Zagreb: Jugoslavenska akademija znanosti i umjetnosti.

Habdelić, J. (1670). *Dictionar ili reči slovenske*. Accessed at: http://crodip.ffzg.hr/default_hr.aspx [15/03/2021].

HJP = *Hrvatski jezični portal*. Accessed at: http://hjp.znanje.hr [15/03/2021].

Jambrešić, A. (1992). *Lexicon Latinum interpretation Illyrica, Germanica et Hungarica locuples.* Zagreb: Hrvatski filološki institut.

Kašić, B. (1999). *Hrvatsko-talijanski rječnik s Konverzacijskim priručnikom.* Zagreb: Kršćanska sadašnjost.

Mikalja, J. / Gabrić-Bagarić, D., Horvat, M., Lovrić Jović, I. & Perić Gavrančić, S. (2011). *Blago jezika slovinskoga*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

RCJHR = *Rječnik crkvenoslavenskoga jezika hrvatske redakcije* (2000-2018). Zagreb: Staroslavenski institut.

RHJ = Šonje, J. (2000). *Rječnik hrvatskoga jezika*. Zagreb: Leksikografski zavod Miroslav Krleža – Školska knjiga.

Slovník = *Slovník jazyka staroslověnského* (1966-1997). Praha: Akademia.

Stulli, J. (1801). *Lexicon latino-italico-illyricum.* Budapest: Regia Universitas Pestana.

7 References

Adams, J.N. (1982). *The Latin Sexual Vocabulary*. London: Duckworth.

Allan, K., Burridge, K. (2006). *Forbidden words: taboo and the censoring of language.* Cambridge: Cambridge University Press.

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bratanić, M. (2005). Mjesto žene u rječniku. In D. Stolac, N. Ivanetić & B. Pritchard (eds.) Jezik u društvenoj interakciji. Zagreb – Rijeka: Hrvatsko društvo za primijenjenu lingvistiku, pp. 37-46.

Cloete, A.E. (2014). The treatment of sensitive items in dictionaries. In R.H. Gouws et al. (eds.) Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. De Gruyter Mouton, pp. 482-486.

Coffey, S. (2010). Offensive items, and less offensive alternatives, in English monolingual learners' dictionaries. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress.* Leeuwarden, pp. 1270-1281.

Dakić, M. (2017). Kada ćemo postati žene u rječnicima hrvatskoga jezika? In *Jat: časopis studenata kroatistike*, 1(3), pp. 106-123.

Dykstra, A. (2006). J.H. Halbertsma: 'Sexual Language and the Lexicon Frisicum (1872)'. In *Dictionaries: Journal of the Dictionary Societyof North America*, 27, pp. 21-35.

European Institute for Gender Equality (2021). Sexuality. In *Glossary & Thesaurus.* Accessed at: https://eige.europa.eu/thesaurus/terms/1379 [11/06/2021].

Farø, K.J., Jensen, L.V. (2018). Den Danske Ordbog på nettet: en undersøgelse af version 3.0. In *LexicoNordica*, 25, pp. 215-232.

Fjeld, R.V. (2015). Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordbøker. In C. Sandström et al. (eds.) Perspektiv på lexikografi, grammatik och språkpolitik i Norden. Helsinki: Institutet för de inhemska språken, pp. 35-65.

Fournier, H.S., Russell, D.W. (1992). A study of sex-role stereotyping in the Oxford English Dictionary 2E. In *Comput Hum*, 26, pp. 13-20.

Gorjanc, V. (2004), Politična korektnost in slovarski opisi slovenščine – zgolj modna muha? In Stabej, M. (ed.) Moderno v slovenskem jeziku, literaturi in kulturi: zbornik predavanj / 40. seminar slovenskega jezika, literature in kulture, 28. 6.–16. 7. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, pp. 153-161.

Gorjanc, V. (2005). Neposredno in posredno žaljiv govor v jezikovnih priročnikih: diskurz slovarjev slovenskega jezika. In *Družboslovne razprave*, 21(48), pp. 197-209.

Harteveld, P., van Niekerk, A.E. (1996). Policy for the Tretment of Insulting and sensitive Lexical Items in the Woordeboek van die Afrikaanse Taal. In M. Gellerstam et al. (eds.) *Proceedings of the 7th EURALEX International Congress*. Göteborg, pp. 381-393.

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. In *Behavior Research Methods*, 40(4), pp. 1065-1074.

Jay, T. (1977). Doing research with dirty words. In *Maledicta: The International Journal of Verbal Aggression*, 1(2), pp. 234-256.

Jay, T. (1992). *Cursing in America.* Philadelphia: John Benjamins.

Jay, T. (2000). *Why we curse: A neuropsychosocial theory of speech.* Philadelphia: John Benjamins.

Jay, T. (2009). The Utility and Ubiquity of Taboo Words. In *Perspectives on Psychological Science*, 4(2), pp. 153-161.

Jay, T., Caldwell-Harris, C. & King, K. (2008). Recalling Taboo and Nontaboo Words. In *The American Journal of Psychology*, 121(1), pp. 83-103.

Jensen, J., Lorentzen, H., Nimb, S., Svendsen, M.-M.M. & Trap-Jensen, L. (2018). Thaipiger, muskelhunde og fulde svenskere: nedsættende ord, stereotyper og ligestilling i Den Danske Ordbog. In *Nordiske Studier i Leksikografi*, 14, pp. 141-151.

Koopman, A. (2016). Ethnonyms. In C. Hough (ed.) The Oxford Handbook of Names and Naming, pp. 251-262.

Libela (2013). Jezik pokazuje rodnu neravnopravnost. Published: 09/10/2013. Accessed at: https://www.libela.org/sa-stavom/4326-jezik-pokazuje-rodnu-neravnopravnost/ [15/06/2021].

Lebsanft, F. (1997). Sprachtabu und Euphemismus in der französischen Sprachgeschichte. In G. Holthus, J. Kramer & W.

Schweickard (eds.) Italica et Romanica, Festschrift für Max Pfister zum 65. Geburtstag. Band 3. Tübingen: Max Niemeyer Verlag, pp. 111-131.

Ljubešić, N., Klubička, F. (2016). *Croatian web corpus – hrWaC 2.1.* Accessed at: http://hdl.handle.net/11356/1064 [26/02/2021].

Mills, S. (2008). *Language and sexism.* Cambridge – New York: Cambridge University Press.

Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* Bolzano, pp. 85-105.

Norri, J. (2000). Labelling of derogatory words in some British and American dictionaries. In *International Journal of Lexicography*, 13(2), pp. 71-106.

Norri, J. (2019). Treatment of words for illness and disability in monolingual English dictionaries. In *International Journal of Lexicography*, 33(3), pp. 227-250.

Petersson, S., Sköldberg, E. (2020). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Democritus University of Thrace, pp. 381-386.

Pišković, T. (2017). Perpetuiranje rodnih stereotipa u hrvatskim rječnicima. In *Romanoslavica*, 52(2), pp. 343-363.

Radtke, E. (1986). Konstanz und Wandel in der Beurteilung von Sexualia in der Geschichte der Lexikographie. In *Osnabrücker Beiträge zur Sprachtheorie*, 35, pp. 107-117.

Reinhart, J. (1994). Slav. Pyjь "membrum virile". In P. Vavroušek (ed.) Iranian and Indo-European Studies. Memorial Volume of Otakar Klíma, pp. 219-223.

Russel, L.R. (2012). This is What a Dictionary Looks Like: The Lexicographical Contributions of Feminist Dictionaries. In *International Journal of Lexicography*, 25(1), pp. 1-29.

Schutz, R. (2002). Indirect Offensive Language in Dictionaries. In A. Braasch, C. Povlsen (eds.) *Proceedings of the 10th EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, pp. 637-641.

Schweickard, W. (1997). Tabu und Euphemismus in der italienischen Lexikographie. In M. Lieber, W. Hirdt (eds.) Kunst und Kommunikation. Betrachtungen zum Medium Sprache in der Romania, Festschrift zum 60. Geburtstag von Richard Baum. Tübingen, pp. 303-310

Talbot, M. (2005). Gender Stereotypes: Reproduction and Challenge. In J. Holmes, M. Meyerhoff (eds.) The Handbook of Language and Gender. Oxford: Blackwell Publishing, pp. 468-486.

The Herald (1993). Dictionary withdrawn in war of words over 'Bangkok prostitutes' entry. Published: 06/07/1993. Accessed at: https://www.heraldscotland.com/news/12730658.dictionary-withdrawn-in-war-of-words-over-bangkok-prostitutes-entre/ [15/06/2021].

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam – Philadelphia: John Benjamins Publishing Company.

Trojar, M., Žagar Karer, M. (2013). Družbena občutljivost v terminoloških slovarjih. In A. Žele et al. (eds.) Družbena funkcijskost jezika (vidiki, merila, opredelitve). Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, pp. 457-463.

Wodak, R., Benke, G. (2000). Gender as a Sociolinguistic Variable. In F. Coulmas (ed.) The Handbook of Sociolinguistics. Oxford – Malden: Blackwell Publishers, pp. 127-151.

Acknowledgements

This paper was written within the research projects *Croatian Web Dictionary – MREŽNIK* (IP-2016-06-2141) and *Retro-digitization and Interpretation of Croatian Grammar Books before Illyrism* (IP-2018-01-3585), financed by the Croatian Science Foundation.

Revised entries in the multi-volume edition and TEI encoding: a case of the historical dictionary of Russian

Lyashevskaya O.[1,2], **Penkova Y.**[2,1]

[1] *National Research University Higher School of Economics*
[2] *V. V. Vinogradov Russian Language Institute RAS*
olesar@yandex.ru, amoena@inbox.ru

Abstract

The Dictionary of Russian Language of the 11th – 17th centuries (DRL11–17), which covers both Old and Middle Russian periods, is an ongoing project of the Russian Academy of Sciences, with volumes 1–31 published in hardcopy in 1975–2019). Up to now, only volumes 28-30 were converted into the database and published free online (http://web-corpora.net/wsgi/oldrus.wsgi/). The online edition allows one to search for entries that contain particular grammatical properties, phraseological units, sources of etymology, texts and sources attested in the entry, historical periods they represent, etc. (Aksyonov et al. 2015, Vechkaeva 2016). This paper presents a new initiative aimed at the digitization of earlier volumes, which includes OCR, encoding the dictionary according to a TEI-compatible XML scheme, improving the integrity of entries, and additional data mining and enrichment using external resources. We focus on the issue of how to represent the revised entries, namely, those that were added, deleted, and corrected in subsequent volumes and in a supplementary volume.

The changes in the entries are usually powered by new sources taken into consideration, by new interpretations of the source documents, or by changes in editorial policy. The typology of revisions made by the authors and editors of later volumes includes: adding or deleting entries; adding or deleting certain parts of the entry (senses, examples, etymology, etc.); correcting one or several fields of the entry (definition, example, grammatical properties, bibliographic description of citations, etc.). More complex changes are decomposed into the components listed above.

The TEI-based scheme of the dictionary addresses two ways of presenting the content: (i) an online searchable version and (ii) a retro-digitized version that preserves the layout of the published volumes. In the first case, the revised entry is represented as one merged entry (Target) that incorporates data from Source (entry published in an earlier volume) and Revision (entry published as addendum in later or supplementary volumes). As neither Source no Revision presents the correct content of the entry in full, the TEI-based representation of the Target should be generated. Besides that, advanced users may have access to the history of changes made by editors and to deleted entries. We use the critical apparatus module of TEI to track the history of changes, in which the lemma contains a “preferred”, corrected reading and another reading corresponds to the content provided in earlier volumes. Taking the perspective of the retro-digitized version, Source and Revision are two separate entries with different metadata. Nevertheless, these two entries are linked to each other using the reference tags. Taken as a whole, the proposed schema outlines the principles for documenting the genetic relationships between different versions of edited lexicographic material.

Keywords: historical lexicography; TEI encoding; retro-digitizing; dictionary content revision; Old Russian; Dictionary of Russian Language of the 11th–17th centuries

1. The Dictionary of Russian Language of the 11th – 17th centuries and its revision history

The Dictionary of Russian Language of the 11th–17th centuries (DRL11-17) presents different periods of the Old Russian language including Middle Russian (15–17th centuries). The compilation of the DRL11–17 was initially aimed at creating a wide-audience popular-science manual for reading Old Russian texts (cf. the verso of the title page and the preface to (DRL11–17, 1, 5-16)), provided with a minimal number of citations from the Old Russian sources (usually, the earliest and the latest attestations were given). The concept of the DRL11–17 had been changed several times even before the release of the first volume (the editors first planned to publish the so-called “Small Old Russian Dictionary of the 11th – 17th centuries”, see (Krysko 2007, 108)). During the publication process, the concept of the Dictionary underwent significant changes. From a popular-science manual, it gradually turned into an academic dictionary, cf. (Chernysheva 2013).

Changes in the concept caused the revision of many lexicographic principles: the number of source citations highly increased; recently published Old Russian texts are constantly added to the list of the dictionary sources, outdated editions being excluded; unknown originals for Old Russian translations are regularly identified.

One of the consequences of the conceptual change was the need for a revision of the previous volumes, taking into account the recent editions of sources, the emergence of Old Russian corpora, and newly identified originals. Most productively, this work was being carried out when V.B. Krysko became the editor-in-chief of the DRL11–17, i.e., in the 27th–29th volumes, cf. (Krysko 2007).

In 2006, additions and corrections to the first volume of the DRL11–17 were published, both in the appendix to volume 27, and as a separate volume (DRL11–17, vol. 27; Chernysheva 2006). However, the editorial board decided to postpone the systematic publication of additions to the earlier volumes until all the DRL11–17 volumes were completed. The editors decided to provide each current volume only with corrections and additions to the later volumes.

Thus, volume 27 is supplemented with additions for the alphabetic segment “C” and corrections to the previous 26th volume. The 28th volume (DRL11–17, vol. 28) includes only corrections to the 2nd volume (the alphabetic segment

"B"), the additions to the segment "B" being not yet published.

The 29th volume (DRL11–17, vol. 29) contains additions for the alphabetic segment "C", omitted in previous volumes (including those omitted in additions published in DRL11–17, vol. 27: 214-216). The "Supplements" also indicate words that were mistakenly included in the 28th volume and should be removed. The 30th volume (DRL11–17, vol. 30: 316–318) is supplemented by the additions to the alphabetical segments "C–T" and the corrections to the previous 29th volume. Volume 31 (DRL11–17, vol. 31) also contains some additions to the segments "C–T"; however, it does not include any corrections planned to be published in the upcoming volumes. So, various additions and corrections are scattered across the latest five volumes (27th–31st).

2. Variety of revision strategies

The lexicographic principles for submitting corrections and additions were not completely unified in the latest volumes of the DRL11–17. In the 29th volume, references to entries included in the appendix are marked with an asterisk, cf.:

(1) САМООБѢЩАННО... – Ср. самоотвѣщанно* (DRL11–17, vol. 29: 424)
'SAMOOBĚŠČANNO… – Cf. samootvěščanno*'

In the supplementary volume (Chernysheva 2006), this symbol is used wider: it marks all newly discovered lemmas that are missing in the 1st volume. In the 31st volume, the asterisk is used to refer to the unpublished additions that will be included only in the next, 32nd volume.

However, the differences in structuring additions and corrections are not limited to ways of using an asterisk. The crucial difference between the lexicographic strategies used in different volumes is the following. In (Chernysheva 2006), additions and corrections are provided with a special commentary, such as "previous interpretation is removed"; "lemma is corrected"; "previous interpretation is specified"; "misprint in the lemma is corrected". The previous variant, as a rule, is not indicated. In some cases, it is not easy to understand what changes are being made without consulting the original article from Volume 1, cf.:

(2) БРЕЩИ. знач.4: цитаты с отрицанием переносятся в статью НЕБРЕЩИ.
Брещи себе. Исправлены опечатки. (Chernysheva 2006: 60)
'BREŠČI. sense 4: citations with negation transferred in the entry NEBREŠČI'
'brešči sebe. Misprints corrected.'

In presenting corrections to the 2nd volume, published in the appendix to volume 28 (DRL11–17, vol. 28: 291–302), another strategy is adopted. The corrections are structured in the form of a table. The left column contains a lemma of the dictionary entry being amended. The lemma is given in the form as presented in the 2nd volume. In the right column, the information that requires correction is placed to the left of the "arrow" symbol (⇒). The Arabic numerals indicate the sense number, and a certain type of font indicates the corresponding field of the dictionary entry: bold capital letters mark lemmas, bold letters – references, italics indicate interpretation, wide font – idiomatic combinations. The correct variant stands to the right of the arrow symbol; a semicolon marks boundaries of the corrections. The initials of the editors who provided the editor-in-chief with the corresponding amendment are put in square brackets. If the correction concerns only the lemma, the arrow symbol is not used, and the corrected lemma variant is put in the right column (see table 1).

ВДАНИЕ VDANIE	2. Действие по глаг. вдатися (в знач. 2) ⇒ Сдача (города) без боя, капитуляция [К.М.]; овоихъ ⇒ ово ихъ (2 раза) '2. Action related to the verb vdatisya (sense 2) ⇒ Surrender (of the town) without a fight, capitulation [K.M.]; ovoih'' ⇒ ovo ih'' (2 times)' 3. Вещь, отданная на сбережение; залог ⇒ Вещь или сумма денег, отданная в залог серьезности намерений заключить брак [К.М.]. '3. A savings item; a pledge ⇒ A thing or amount of money pledged for the seriousness of intent to marry [K.M.].'
ВДАНИЕ VDANIE	2. Действие по глаг. вдатися (в знач. 2) ⇒ Сдача (города) без боя, капитуляция [К.М.]; овоихъ ⇒ ово ихъ (2 раза) '2. Action related to the verb vdatisya (sense 2) ⇒ Surrender (of the town) without a fight, capitulation [K.M.]; ovoih'' ⇒ ovo ih'' (2 times)' 3. Вещь, отданная на сбережение; залог ⇒ Вещь или сумма денег, отданная в залог серьезности намерений заключить брак [К.М.]. '3. A savings item; a pledge ⇒ A thing or amount of money pledged for the seriousness of intent to marry [K.M.].'

Table 1: A sample of corrections presented in (DRL11–17, vol. 28: 292).

If the changes are related to removing a segment or transferring it to another dictionary entry, the arrow symbol is not used, and the correction is provided only with a comment: "citation is transferred to article N," or "entry is removed," etc.

The types of the changes approve the strategy used in (Chernysheva 2006): the supplementary volume deals mostly with adding new entries. It determined the principles for presenting additions and corrections (see above). However, the

strategy used in the 28th volume seems to be more convenient for the user of the dictionary, on the one hand, and it is easier to integrate the changes into the dictionary database, on the other hand.
The inconsistency with which additions and corrections are presented in different volumes of the DRL11–17 introduces difficulties for integrating the corresponding ones into the dictionary database. However, before discussing particular technical solutions, it is necessary to establish all the types of additions and corrections which we are dealing with.

3. Typology of the content revisions

Various reasons cause the changes that have been introduced by the editors. There are many different classifications of lexicographic errors depending on the particular purposes. One can propose a lexicographic error classification based on the reasons behind these errors, cf. (Shapoval 2016). There are various types of such errors:

- errors as a result of an incorrect reading of sources (for example, incorrect word division in a manuscript);
- transmission errors, i.e., errors brought by copying text from a source to dictionary entry;
- errors which go back to the publishers of Old Russian sources cited and uncritically reproduced in the dictionary;
- errors resulting from incorrect interpretation or reconstruction of grammatical forms, etc.

The changes in the entries can also be powered by new interpretations of the source documents, by new sources taken into consideration, or by changes in editorial policy. However, for our purposes, we need a technical classification of changes. The typology of revisions made by the authors and editors of later volumes includes adding missing entries, or parts of the entries, and revision.

3.1. Additions

Types of *additions* include:

- adding a missing dictionary entry;
- adding a missing reference entry;
- adding a missing part in an existing entry (i.e., adding Greek glosses to the citation taken from Old Russian translations, identifying a new sense, etc.);
- expanding already existing fields (i.e., citation extension).

3.2. Simple and complex revisions

We distinguish between simple and complex revisions. The former ones suggest that the correction can be made in one step. Simple corrections include deleting an existing entry or a certain part of it, replacing the contents of one of the fields, moving a part of the entry to another entry, as well as simple corrections in one of the fields (i.e., correction of misprints).
Simple replacements may refer to:

- lemma spelling;
- grammatical information;
- interpretation provided, no changes in the citations are required;
- source citation, unless it causes changes in the chronological order of the citations.

Simple transfers include:

- moving a source citation from one sense to another within the same dictionary entry, provided the latter preserves its structure;
- moving citations to another dictionary entry, provided both entries preserve their structure.

Simple deletions include:

- deleting the dictionary entry, unless it requires deletion of the reference included in another dictionary entry;
- deleting a part of the entry along with all the citations.

Complex revisions are a combination of two or more corrections, i.e., such changes, which, from a technical point of view, should be introduced in several steps. Complex corrections include the following:

- deleting a dictionary entry, along with transferring citations from the deleted entry to other dictionary entries;
- transferring citations along with creating a new dictionary entry;
- transferring citations to another dictionary entry, along with adding to the latter, a new sense field;
- replacing an existing dictionary entry with a reference entry, along with transferring the citations to other dictionary entries.

Table 2 shows an example of a complex change, a replacement of the dictionary entry with the reference entry, while the citation is transferred to another dictionary entry.

ВОЖАВСТВО VOŽAVSTVO	ВОЖАВСТВО см. вожевство; цитата переносится в статью ВОЖЕВСТВО ‘VOŽAVSTVO see voževstvo; citation transferred to the entry VOŽEVSTVO’

Table 1: A sample entry with a complex change.

3.3. Implicit corrections

We discussed the classification of revisions to the earlier volumes of the DRL11–17, proposed by the editors of the latter five volumes. However, there is also another type of correction, which can be called implicit.

Firstly, implicit corrections arise in the case when the source, used for decades, acquires a new dating, due to a separate archaeographic research. Secondly, implicit corrections include introduction of new source abbreviations. As a result, different volumes of the dictionary can refer to the same source, using different abbreviations. Thirdly, implicit corrections are powered by the termination of the use of a particular source recognized as unsuitable (i.e., if a manuscript is recognized as not featuring the Russian redaction of Church Slavonic). Fourthly, spelling and interpretation of Greek glosses (and, less commonly, glosses in other languages), is also the subject of editorial revisions in subsequent volumes. Other corrections can also occur.

4. The DRL11–17 electronic edition

Starting from 1975, the dictionary was originally published in hardcopy (the 31st volume was released in 2019). The their first attempt to create an online edition, Aksyonov et al. (2015) and Vechkaeva (2016) converted the content of volumes 28-30 into the lexicographic database and made it possible to search for head words, grammatical labels, phraseological units, sources of etymology, texts and sources attested in the entry and historical periods they represent, etc. online (http://webcorpora.net/wsgi/oldrus.wsgi/). The new objective is to expand the coverage of the database, so that it eventually includes all printed volumes, and to make access to the content more flexible and user-friendly.

The data is automatically recognized, using the ABBYY FineReader OCR system, manually checked and presented in three different ways:

- TEI-compatible XML scheme;
- SQlite database format;
- screenshots of the printed dictionary pages and individual entries (pdf view).

Figures 1 and 2 illustrate the TEI-based representation and the browser view, based on the SQlite representation of the same entry. By clicking on the “book” pictogram, the user can switch to the pdf view of the printed edition, see figure 3.

To improve the integrity of entries, some implicit information, not present in the printed version, was retrieved and added to the TEI-based representation. For that purpose, we mostly use tags, embedded within the definition and example fields. For example, the definition template elements *Том, кто…* ‘The one who (does smth.)’ is marked with the tag <defTemplate class="agent">. The corresponding nominalization, denoting the agent’s action (*спасение* ‘salvation’), has a cross-reference to the entry *спасение*. In examples, the mentioned headwords were identified and labeled with <oRef>, eg. <oRef type="pl">*сп҃сьници*</oRef> for the plural abbreviated form of the headword *спасникъ*. Glosses were classified by language and scope, see the tag <gloss lang="gr" class="ex"> for the Greek gloss that corresponds to the Greek source of the cited example. Besides that, the editors’ notes, with regard to definitions and examples, were marked and classified. Missing data tags were added, using information from the source titles, see the tag <date class="hidden" when="1073"> which has the attribute ‘hidden’. In enhanced TEI-based representation, lemmas for all Russian and Old Russian words are provided so that the user could look for particular words in definitions, examples, grammatical data, and other fields throughout the entries.

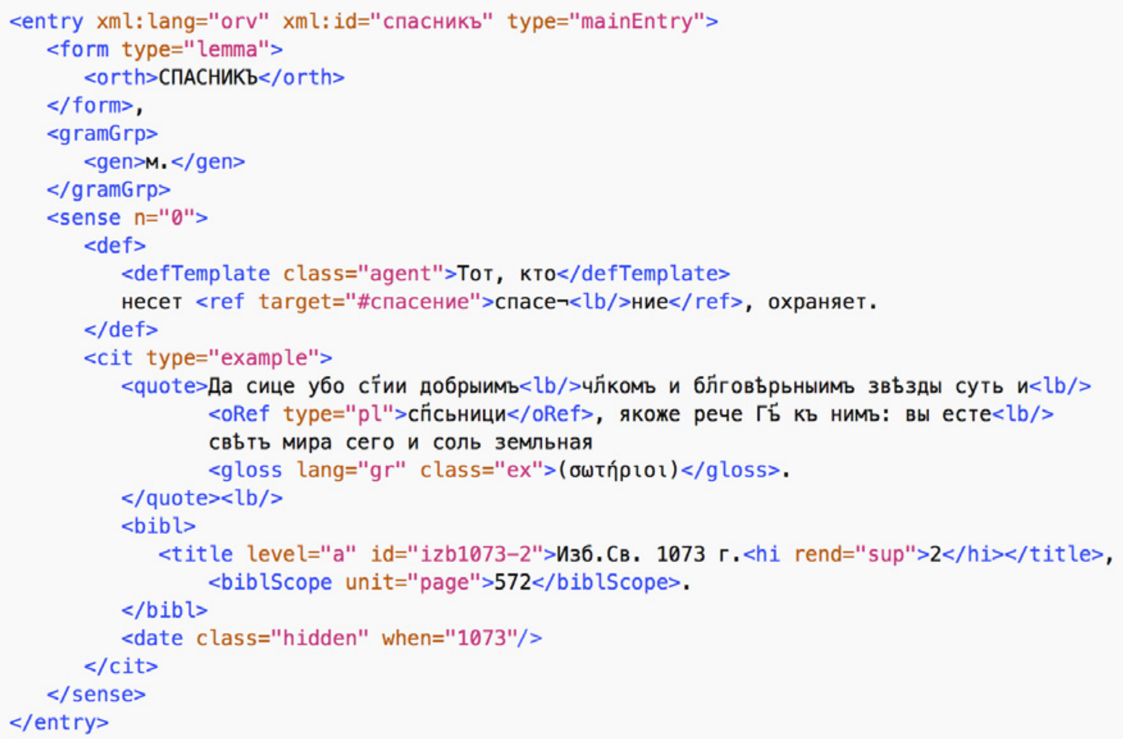
The TEI scheme of the dictionary is designed to support two ways of presenting the content:

(i) an online searchable version and

(ii) a retro-digitized version that preserves the layout of the published volumes.

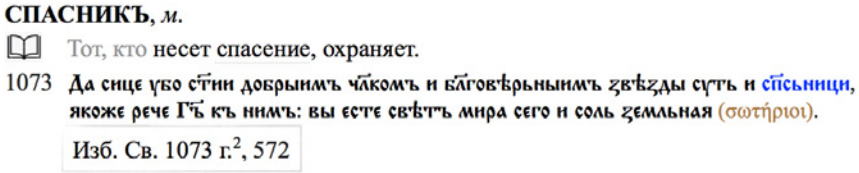
With such functionality, line break tags (<lb/>) are preserved in the TEI-based representation, but used only in (ii). Analogically, information tagged with the attribute ‘hidden’ is used in (i) and not shown in (ii).

The following section addresses the representation of the revised entries, namely, those that were added, deleted, and corrected in the TEI scheme, taking into account (i) and (ii).



```
<entry xml:lang="orv" xml:id="спасникъ" type="mainEntry">
   <form type="lemma">
      <orth>СПАСНИКЪ</orth>
   </form>,
   <gramGrp>
      <gen>м.</gen>
   </gramGrp>
   <sense n="0">
      <def>
         <defTemplate class="agent">Тот, кто</defTemplate>
         несет <ref target="#спасение">спасе¬<lb/>ние</ref>, охраняет.
      </def>
      <cit type="example">
         <quote>Да сице убо ст҃ии добрыимъ<lb/>чл҃комъ и бл҃говѣрьныимъ звѣзды суть и<lb/>
               <oRef type="pl">сп҃сьници</oRef>, якоже рече Гь҃ къ нимъ: вы есте<lb/>
               свѣтъ мира сего и соль земльная
               <gloss lang="gr" class="ex">(σωτήριοι)</gloss>.
         </quote><lb/>
         <bibl>
            <title level="a" id="izb1073-2">Изб.Св. 1073 г.<hi rend="sup">2</hi></title>,
               <biblScope unit="page">572</biblScope>.
         </bibl>
         <date class="hidden" when="1073"/>
      </cit>
   </sense>
</entry>
```

Figure 1: A simplified TEI-based representation of the entry *спасникъ* ‘savior’.



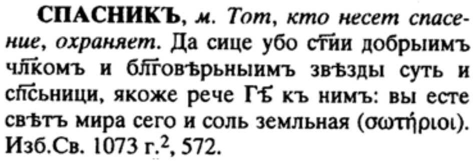
СПАСНИКЪ, *м.*

Тот, кто несет спасение, охраняет.

1073 **Да сице ѹбо ст҃ии добрыимъ чл҃комъ и бл҃говѣрьныимъ ꙁвѣꙁды сѹть и сп҃сьници, якоже рече Гь҃ къ нимъ: вы есте свѣтъ мира сего и соль ꙁемльная** (σωτήριοι).

Изб. Св. 1073 г.², 572

Figure 2: The browser view for the entry *спасникъ* ‘savior’.



СПАСНИКЪ, *м. Тот, кто несет спасение, охраняет.* Да сице убо ст҃ии добрыимъ чл҃комъ и бл҃говѣрьныимъ звѣзды суть и сп҃сьници, якоже рече Гь҃ къ нимъ: вы есте свѣтъ мира сего и соль земльная (σωτήριοι). Изб.Св. 1073 г.², 572.

Figure 3: The pdf view for the entry *спасникъ* ‘savior’.

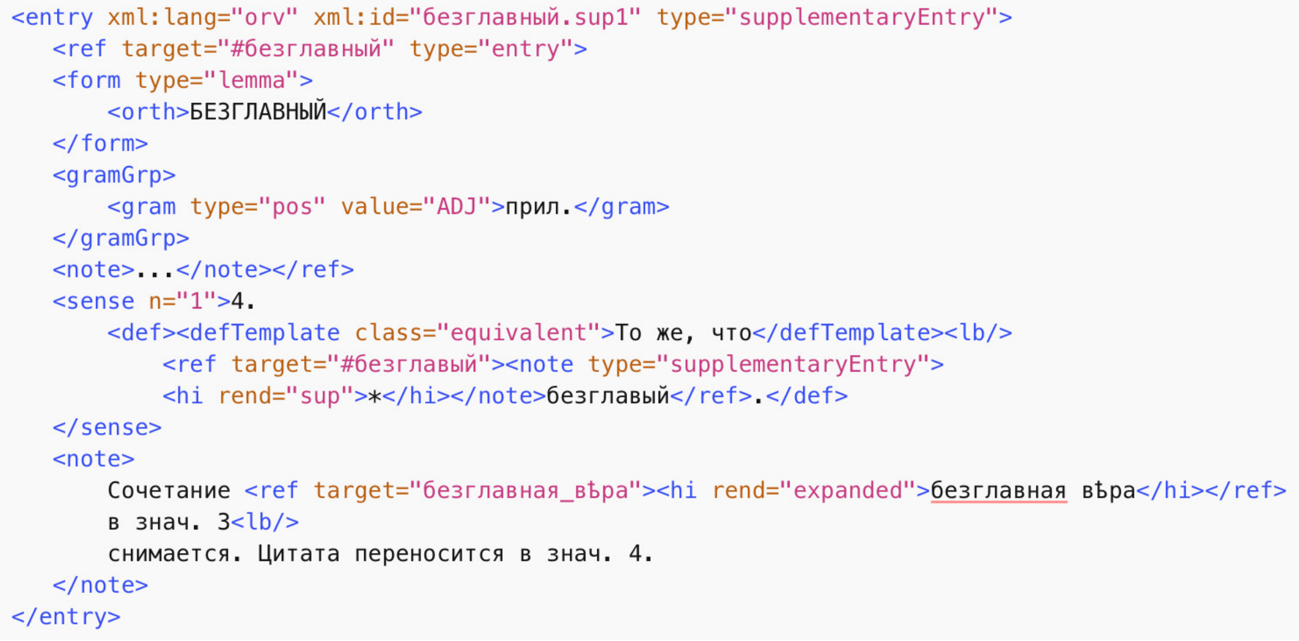
5. TEI-based approach to track the revision changes

In the online searchable version (i), the revised entry is represented as one merged entry that incorporates data from both Source (entry published in an earlier volume) and Revision (entry published as addendum in later or supplementary volumes). As none of them presents the correct content of the entry in full, the TEI-based representation of the Target should be generated from both. In the retro-digitized version (ii), Source and Revision are considered as two (or more) separate entries, with different metadata. However, it is useful to link these entries to each other, to facilitate access to the related content. Figure 4 shows the Revision entry of the word *безглавный* ‘headless’ (see the TEI tag <entry type="supplementaryEntry">), which refers to the Source entry using the TEI tag <ref>.

While compiling the electronic version of the dictionary, we face a challenging task to represent the various kinds of additions and corrections made, over time, by the editors of the DRL11–17. At the same time, advanced users may have access to the revision history of a given entry and to deleted entries. So, the history of changes should also be reflected in the representation.

Taking into account the typology of revisions provided in section 3, we distinguish among:

- adding and deleting the whole entry, reference entry, particular senses;
- transferring data from one entry or subentry (sense) to another, and from one field to another;
- changing the order of certain elements;
- replacing data within a particular entry field.



```
<entry xml:lang="orv" xml:id="безглавный.sup1" type="supplementaryEntry">
   <ref target="#безглавный" type="entry">
   <form type="lemma">
       <orth>БЕЗГЛАВНЫЙ</orth>
   </form>
   <gramGrp>
       <gram type="pos" value="ADJ">прил.</gram>
   </gramGrp>
   <note>...</note></ref>
   <sense n="1">4.
       <def><defTemplate class="equivalent">То же, что</defTemplate><lb/>
           <ref target="#безглавый"><note type="supplementaryEntry">
           <hi rend="sup">*</hi></note>безглавый</ref>.</def>
   </sense>
   <note>
       Сочетание <ref target="безглавная_вѣра"><hi rend="expanded">безглавная вѣра</hi></ref>
       в знач. 3<lb/>
       снимается. Цитата переносится в знач. 4.
   </note>
</entry>
```

Figure 4: A simplified TEI-based representation of corrections to the entry *безглавный* ‘headless’ in the supplementary volume.

Complex changes are decomposed into the components listed above. Technically, transferring and changing the order of elements is encoded as adding and deleting in the TEI-based encoding.

There is no uniform convention on how to represent the revision history in dictionaries. On the one hand, the tags <add>, <del>, <subst> are good candidates to reflect simple changes, but they are defined to encode changes made in the same primary source (i.e. one tangible medium) (TEI P5 2021). Certainly, this does not hold true in our case, since we deal with multiple, though related, media that carry the changes. On the other hand, tags such as <revisionDesc> and <listChange> are intended to summarize the history of changes in the header, rather than throughout the body of the document. We choose the third option, namely, using the critical apparatus module of TEI that is intended to represent related texts found in different physical witnesses (TEI P5 2021, Section 12).

To illustrate the use of the critical editing tags, let us move back to the previous example. In the Revision entry of the adjective *безглавный* ‘headless’, sense 4 is added provided with a definition only. The editors add a note that the section of the sense 3 that contains a multi-word expression is deprecated; however, the citation from this section should be transferred to sense 4. The reason for restructuring the entry is that the new adjective *безглавый* ‘headless’, of roughly the same morphological structure and with the same meaning as in the multi-word expression, is added to the dictionary.

In order to document all the changes, the list of all DRL11–17 volumes is declared as witnesses in the <listWit> element of the TEI header. In the entry, the revised sections are enclosed in the <app> tag (see Figure 5).

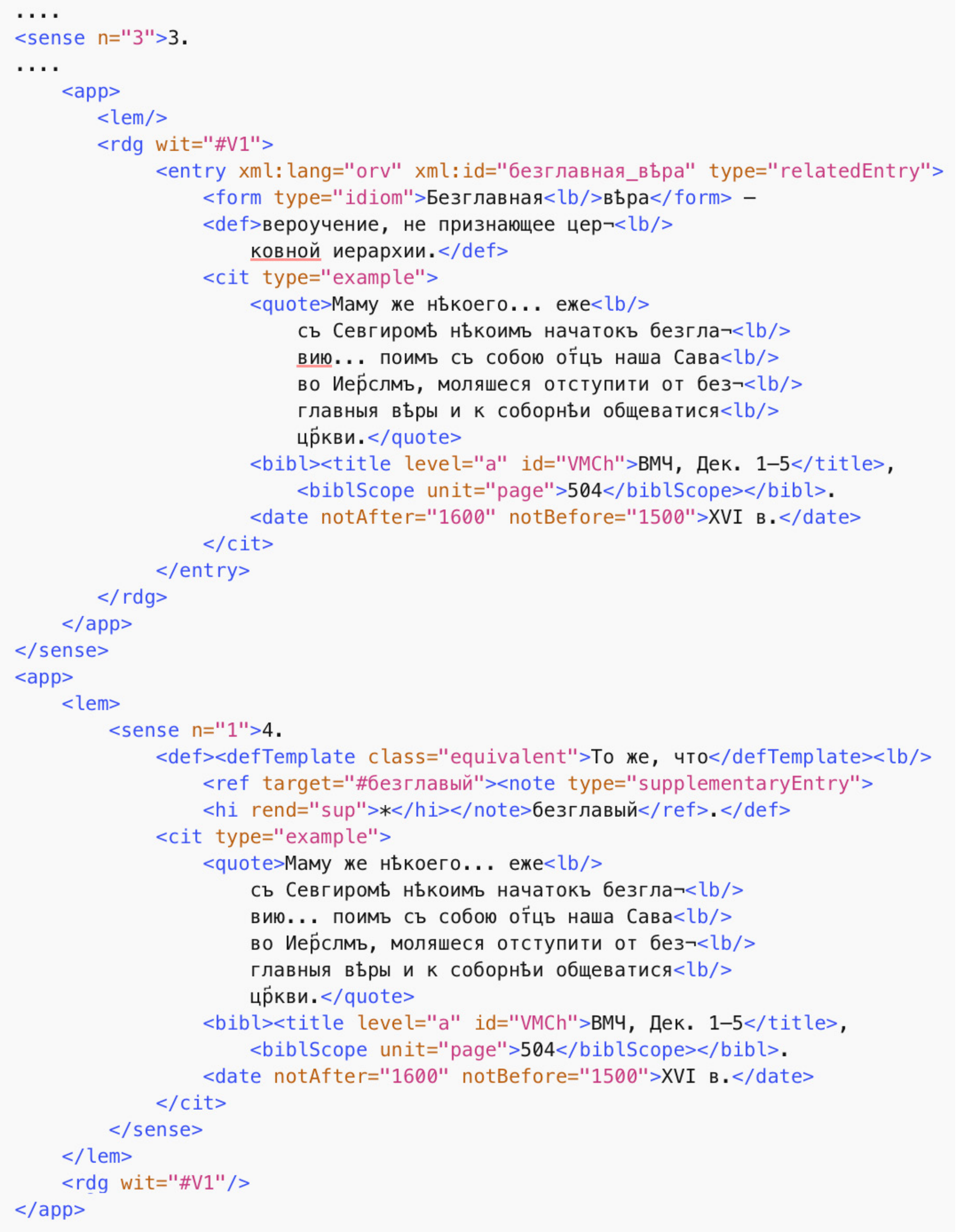
Since the entries have been revised only once by now, each <app> section contains a <lem> element (a “preferred”, corrected reading) and one <rdg> element (a reading from an earlier edition identified by the attribute *wit*). In the case of the subentry transfer, there are two <app> sections corresponding to deletion and addition, respectively. The first <app> element contains an empty <lem/> element (deletion) and the subentry for the multi-word expression from the Source entry enclosed in the <rdg wit="#V1">…</rdg> tags. The second <app> element contains a lemma with both new sense and its definition explicitely represented in the Revision entry and the transferred citation. The <rdg> element is empty.

Note that in deviation from most common practice, the Source and Revision witnesses are not different versions of the same work, rather, the Revision can be considered as a commentary and supplementary material that substitues the content of the Source representing the final version of the lexicographic material. Another limitation of the current TEI-based representation is that content provided by the editors and generated content are not explicitely distinguished in the <lem> elements. Yet, in the absence of better solutions, this takes a step towards encoding the “genetic relationships among documents” (Barney 2018).

6. Conclusion

Digitizing the Dictionary of Russian Language of the 11th – 17th centuries not only makes the published volumes of this long-lasting project more accessible but also gives researchers a more flexible and powerful resource to work with. Our approach offers a new vision of the critical electronic edition for the multi-volume historical dictionaries. This implies the parallel handling of an online, searchable version, enriched with linguistic and textological information, and a retro-digitized version that preserves the layout of the published volumes.

In this paper, we addressed challenges arising from inconsistencies in editorial practice, strategies to report added, deleted, corrected, and restructured entries. Following the developed typology of simple and complex revisions, we adopted a TEI-compatible XML scheme to represent corrections, content restructuring and transfer, made by editors over many years. However, more efforts to standardize the tracking of longitudinal changes in multi-volume dictionaries are needed.



```
....
<sense n="3">3.
....
    <app>
       <lem/>
       <rdg wit="#V1">
            <entry xml:lang="orv" xml:id="безглавная_вѣра" type="relatedEntry">
                <form type="idiom">Безглавная<lb/>вѣра</form> –
                <def>вероучение, не признающее цер¬<lb/>
                    ковной иерархии.</def>
                <cit type="example">
                    <quote>Маму же нѣкоего... еже<lb/>
                        съ Севгиромѣ нѣкоимъ начатокъ безгла¬<lb/>
                        вию... поимъ съ собою от̆цъ наша Сава<lb/>
                        во Иер̆слмъ, моляшеся отступити от без¬<lb/>
                        главныя вѣры и к соборнѣи общеватися<lb/>
                        цр̆кви.</quote>
                    <bibl><title level="a" id="VMCh">ВМЧ, Дек. 1–5</title>,
                        <biblScope unit="page">504</biblScope></bibl>.
                    <date notAfter="1600" notBefore="1500">XVI в.</date>
                </cit>
            </entry>
       </rdg>
    </app>
</sense>
<app>
    <lem>
        <sense n="1">4.
            <def><defTemplate class="equivalent">То же, что</defTemplate><lb/>
                <ref target="#безглавый"><note type="supplementaryEntry">
                <hi rend="sup">*</hi></note>безглавый</ref>.</def>
            <cit type="example">
                <quote>Маму же нѣкоего... еже<lb/>
                    съ Севгиромѣ нѣкоимъ начатокъ безгла¬<lb/>
                    вию... поимъ съ собою от̆цъ наша Сава<lb/>
                    во Иер̆слмъ, моляшеся отступити от без¬<lb/>
                    главныя вѣры и к соборнѣи общеватися<lb/>
                    цр̆кви.</quote>
                <bibl><title level="a" id="VMCh">ВМЧ, Дек. 1–5</title>,
                    <biblScope unit="page">504</biblScope></bibl>.
                <date notAfter="1600" notBefore="1500">XVI в.</date>
            </cit>
        </sense>
    </lem>
    <rdg wit="#V1"/>
</app>
```

Figure 5: A fragment of the generated TEI-based representation of the *безглавный* ‘headless’.

7. References

Aksyonov, K., Bobrik, M., Krivko, R., Orekhov, B., Novosyolova, A., & Vechkaeva A. (2015). Dictionary of Russian Language of the 11th—17th Centuries as a Database: Information Retrieval and Research Perspectives. In *Approaches to the Editing of Slavonic Texts. Tradition and Innovation in Palaeoslavistic Ecdotics’* (*ATTEST*), Regensburg, 11-12 December 2015. Book of Abstracts. Accessed at: https://www.uni-regensburg.de/sprache-literatur-kultur/slavistik/medien/aktuelles/krivko_abstract.pdf [20/02/2020]

Barney, B. (2018). TEI, the Walt Whitman Archive, and the Test of Time. In Proceedings of JADH 2018, p. 251.

Chernysheva, M.I. (2013). Iz istorii izmemenija koncepcii Slovarja russkogo jazyka XI–XVII vv. In *Acta linguistica petropolitana. Proceedings of the Institute for Linguistic Studies RAS.* Vol. IX (2). *Russkaja istoričeskaja* *leksikologija i leksikografija XVII–XIX vv*. [*Russian historical lexicology and lexicography*]. Saint-Petersburg: Nauka, pp. 192–205.

Chernysheva, M.I. (2006). *Dictionary of Russian language of the 11th–17th centuries. Addenda et corrigenda*. Issue 1 (А-Б). Moscow: Nauka.

Chernysheva, M.I. (ed.) (2020) Resource materials to the dictionary of Russian language of the 11th–17th centuries. Index of sources. Word index (direct). Moscow: Leksrus.

DRL11–17, vol. 1: *Dictionary of Russian language of the 11th–17th centuries. Volume 1*: (А–Б). Editor-in-chief S.G. Barhudarov. Moscow: Nauka, 1975.

DRL11–17, vol. 27: *Dictionary of Russian language of the 11th–17th centuries. Volume 27*: (*Спасъ–Старицынъ*). Editor-in-chief V.B. Krysko. Moscow: Nauka, 2006.

DRL11–17, vol. 28: *Dictionary of Russian language of the 11th–17th centuries. Volume 28*: (*Старичекъ–Сулебный*). Editor-in-chief V.B. Krysko. Moscow: Nauka, 2008.

DRL11–17, vol. 29: *Dictionary of Russian language of the 11th–17th centuries. Volume 29*: (*Сулегъ–Тольмиже*). Editor-in-chief V.B. Krysko. Moscow: Nauka, Azbukovnik, 2011.

DRL11–17, vol. 30: *Dictionary of Russian language of the 11th–17th centuries. Volume 30*: (*Томъ–Уберечися*). Editor-in-chief R.N. Krivko. Moscow–Saint-Petersburg: Nestor-Istorija, 2015.

DRL11–17, vol. 31: *Dictionary of Russian language of the 11th–17th centuries. Volume 31*: (*Убивание–Улокъ*). Editor-in-chief R.N. Krivko. Moscow: Leksrus, 2019.

Krysko, V.B. (2007). Russkaja istoričeskaja leksikografija (XI–XVII vv.): problemy i perspektivy [Russian historical lexicography of the 11th — 17th]. In *Voprosy Jazykoznanija*, 1, pp. 103–118.

Shapoval, V. V. (2016). *Teorija i praktika verifikacii slovarnyh dannyh na osnove istočnikov* [*Theory and practice of the lexicographic data verification based on sources*]. Doct. Sc. dissertation. Moscow.

Vechkaeva, Anna (2016). *Dictionary of Russian Language of the 11th — 17th Centuries as a Database: Information Retrieval and Research Perspectives*. BA thesis. Moscow, National Research University Higher School of Economics.

TEI P5 Guidelines, Chapter 11: Representation of Primary Sources. Version 4.0.0. Last updated on 1st March 2021. Accessed at: https://tei-c.org/release/doc/tei-p5-doc/en/html/PH.html [20/03/2021].

Acknowledgements

The work has been supported by the Ministry of Science and Higher Education of the Russian Federation, Agreement No 075-15-2020-793.

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Papers

Lexicography and Corpus Linguistics

A Lexicographic platform for migration terminology: problems and methods

Chiari I.[1]

Dipartimento di Lettere e Culture Moderne, Sapienza University, Rome, Italy
isabella.chiari@uniroma1.it

Abstract

"Language on the Fly" is a lexicographic resource for the domain of migration. The peculiarity of migration lexicon is due to scope (often geographical and institutional) and time. The language of migration is found on an international level where it is defined, for example, by institutions, like the EU regulations (both legal and administrative); on a national level, where general international procedures are modified and adapted to the specific country administrative and general migration policies and, finally, on an ordinary level which is interlinked to issues that migrants have to face in their interactions with institutions (social security, health, education, administrative issues). This paper focuses on corpus-based procedures used to build the second version made of a set of 2,094 entries and collocations starting from Italian language corpora specifically built to represent the three levels of lexicon, and further translated in 5 EU languages and 10 non-EU languages. The translation process also involves corpus-based techniques and multilingual corpora. Building the lemma list on three - specially built - Italian corpora using keyword extraction techniques, the glossary also uses corpus-based techniques to extract glosses that are further rewritten using controlled language in Italian in order to facilitate the use in cultural mediation contexts.

Keywords: migration lexicon; corpora; glossary; multilingual corpora; lexicography

1 Introduction

The "Language on the fly" project is an online platform which aims to provide reliable and updated linguistic, lexicographical, and documentary information for orientation language in the first reception of migrants, immigrants, and asylum seekers upon arrival in an EU country. The multimedia platform, accessible online also from mobile devices such as smartphones and tablets, and downloadable offline, will allow to frame the terminology useful for access, orientation, and reception of the country in three versions: a lexicographic reference resource; guide for humanitarian workers; guide for migrants. The prototype of the platform, to be released by the end 2021, covers basic local, national, and international terminology in the relating field in Italian, English, Arabic and French, while the general resource is built in order to cover many more target languages especially among the languages of countries sources of migration.

The project also addresses methodological questions about the development of linguistic-terminological resources related to the domain of migration, and investigation of linguistic and psycho-social needs for the orientation of the recipients of the platform. The peculiarity of migration lexicon is due to different aspects, mainly depending on the scope (often geographical and institutional) and time. More specifically the language of migration has an international or transnational level where it is defined for example by institutions, like the EU regulations (both legal and administrative); there is a further national level, where general international procedures are modified and adapted to the specific country administrative and general migration policies and a final ordinary level that is interlinked to issues that migrants have to face in their interactions with institutions (social security, health, education, administrative issues). All these aspects and more specifically the first two are furthermore constantly changing with frequent modifications in the legal framework on this subject (Dublin and its revisions, in Italy for example Security decrees that have changed asylum typologies in the last years, etc.). Furthermore, some of the terminology is also used in a non-technical sense in common media generating potentially dangerous ambiguity (*illegal and legal immigration, refugee, economic migrant*, etc.). This complexity needs to be addressed both for the updating of general lexicographic works, but also for the use of cultural mediators that are assisting migrants on their path in the new country and, finally, directly to migrants as beneficiaries of reference glossaries and guides.

The project stems from the need for inclusion that is a basic right especially for those who make significant sacrifices and face danger to reach a safe country and who are often penalized by the insufficient linguistic and legal information when facing the procedures of asylum and residence permit acquisition. This tool aims to fill the gap in reference tools in the hands of legal assistants and operators by providing a reliable resource, accessible and constantly updated and enriched with new languages.

The goal of the project is to bring together different types of operational, scientific, and psychosocial skills in a way to produce a model of linguistic guidance that is both synthetic and characterized by precision and translation accuracy, and usable for different types of publics (lexicographers, mediators and humanitarian workers, migrants). The constant interlocutors of the research group are in fact the civil society organizations operating in the humanitarian field in Italy and Europe, the operators, and the guests of the reception centres of Lazio, the Immigration Offices of the Provinces and the mediators that operate there, the voluntary associations operating in the assistance to the migrants. The prototype of the lexical resource is in fact based on the Italian case, providing a starting point for a quadrilingual resource (in its first version) with Italian, Arabic, English, and French languages.

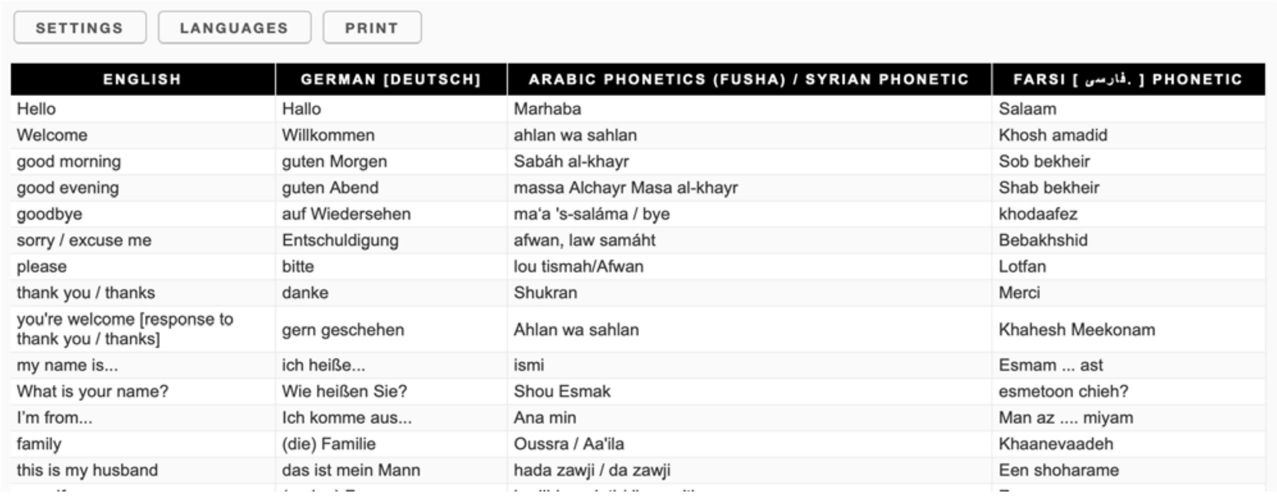
The methodology used is particularly innovative as it combines essential aspects for the development of quality resources, complete and immediately usable for users. In particular, the first phase of experimentation related to the needs of the guests in reception centres through a questionnaire that has been proposed to guests of reception centres in the Rome area and the use of corpus and computational linguistic techniques for the selection of the terminology to be treated.

The first release of the "Language on the Fly" Italian glossary (multilingual in output) was based on three corpora representing the international, national, and ordinary levels of description of the migration lexicon, further elaborated to extract the relevant terminology with different statistical techniques and compared with previously available reference glossaries. The paper will illustrate the methodological structure and procedure to produce the lexicographic resource, its dynamic corpus-based methodology and some of the specific challenges that this kind of terminology poses to lexicographic description and its multiple uses.

The paper is organized in the following way: §2 provides a background on resources that address similar issues from different perspectives and that motivate the need of a new approach for the glossary that is proposed; §3 is a reflection on the different addressee of glossaries on the migration network and on ways to integrate focus on those differences in a unique resource; §4 illustrated the structure of the monolingual section of the glossary; §5 illustrated the cycles by which the SL lemma list is constantly enriched and checked; §6 describes the macro-structure of the construction of the TL section of the dictionary illustrating the asymmetrical relations in content.

2 The Background: Migration Lexicons and Phrasebooks

Lexicography strives to provide translations and cultural reference points, and the dissemination of 'emergency' resources such as online glossaries and phrasebooks, which are often -de facto- prepared by non-professionals without the necessary control of accuracy and in-depth analysis. Examples of this kind are *The Refugee Phrasebook* (Paul Feigelfeld 2016) that is a set of Google Doc Sheets including useful phrases (a general set of 600 phrases, 150 phrases for helpers) that are said to be covering 28 languages. The last update seems to date 2017. The approach is to cover general language basic needs as can be seen in Figure 1. There is also a specific small portion of 150 phrases dedicated to issues defined as generally *juridical* (Feigelfeld 2016).

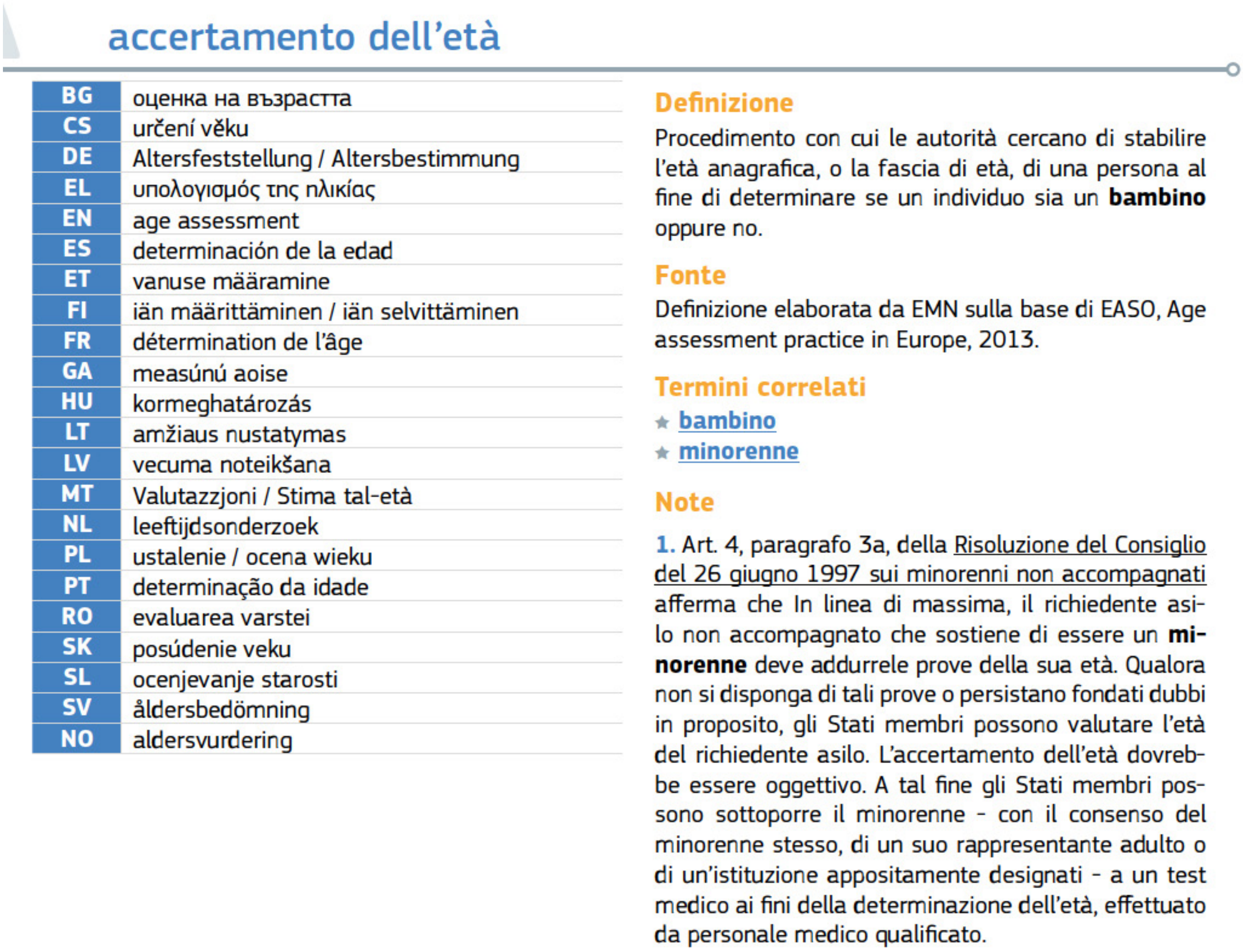


ENGLISH	GERMAN [DEUTSCH]	ARABIC PHONETICS (FUSHA) / SYRIAN PHONETIC	FARSI [فارسی.] PHONETIC
Hello	Hallo	Marhaba	Salaam
Welcome	Willkommen	ahlan wa sahlan	Khosh amadid
good morning	guten Morgen	Sabáh al-khayr	Sob bekheir
good evening	guten Abend	massa Alchayr Masa al-khayr	Shab bekheir
goodbye	auf Wiedersehen	ma'a 's-saláma / bye	khodaafez
sorry / excuse me	Entschuldigung	afwan, law samáht	Bebakhshid
please	bitte	lou tismah/Afwan	Lotfan
thank you / thanks	danke	Shukran	Merci
you're welcome [response to thank you / thanks]	gern geschehen	Ahlan wa sahlan	Khahesh Meekonam
my name is...	ich heiße...	ismi	Esmam ... ast
What is your name?	Wie heißen Sie?	Shou Esmak	esmetoon chieh?
I'm from...	Ich komme aus...	Ana min	Man az .... miyam
family	(die) Familie	Oussra / Aa'ila	Khaanevaadeh
this is my husband	das ist mein Mann	hada zawji / da zawji	Een shoharame

Figure 1: Example from *The Refugee Phrasebook.*

The same kind of approach that is mainly focused on Arabic for Syrian refugees is Gorsau (2015) and contains 200 Arabic sentences, questions and phrases translated into 26 European language, based on the idea of basic needs such as directions, food, accommodation, transport, but not covering any legal or administrative terminology regarding migration.

On the other hand, authoritative multilingual resources are mainly focused on legal and administrative aspects related to EU regulations or best practices such as those of the Council of Europe, and therefore not considering the deep changes that intervene in the 'systems' of reception of individual countries, or their specificities terminologies (European Migration Network 2018, International Organization for Migration (IOM) 2019, Perruchoud 2004), more specifically on Italian language (Programma Integra n/a, Anci, Cittalia & SPRAR 2015). Worth of a specific mention is the Glossary of the European Network *Asylum and Migration Glossary*, started in 2014 and with the last release in 2018 (6.0). It describes fully, with relevant referring sources, 256 entries in 23 languages of the EU. The aim of the glossary is institutional and is aimed at members of the EU and policy makers.



accertamento dell’età

BG	оценка на възрастта
CS	určení věku
DE	Altersfeststellung / Altersbestimmung
EL	υπολογισμός της ηλικίας
EN	age assessment
ES	determinación de la edad
ET	vanuse määramine
FI	iän määrittäminen / iän selvittäminen
FR	détermination de l’âge
GA	measúnú aoise
HU	kormeghatározás
LT	amžiaus nustatymas
LV	vecuma noteikšana
MT	Valutazzjoni / Stima tal-età
NL	leeftijdsonderzoek
PL	ustalenie / ocena wieku
PT	determinação da idade
RO	evaluarea varstei
SK	posúdenie veku
SL	ocenjevanje starosti
SV	åldersbedömning
NO	aldersvurdering

Definizione

Procedimento con cui le autorità cercano di stabilire l’età anagrafica, o la fascia di età, di una persona al fine di determinare se un individuo sia un **bambino** oppure no.

Fonte

Definizione elaborata da EMN sulla base di EASO, Age assessment practice in Europe, 2013.

Termini correlati

★ bambino

★ minorenne

Note

1. Art. 4, paragrafo 3a, della Risoluzione del Consiglio del 26 giugno 1997 sui minorenni non accompagnati afferma che In linea di massima, il richiedente asilo non accompagnato che sostiene di essere un **minorenne** deve addurrele prove della sua età. Qualora non si disponga di tali prove o persistano fondati dubbi in proposito, gli Stati membri possono valutare l’età del richiedente asilo. L’accertamento dell’età dovrebbe essere oggettivo. A tal fine gli Stati membri possono sottoporre il minorenne - con il consenso del minorenne stesso, di un suo rappresentante adulto o di un’istituzione appositamente designati - a un test medico ai fini della determinazione dell’età, effettuato da personale medico qualificato.

Figure 2: Example of entry *accertamento dell'età* in EMN Glossary (Italian version, 6.0, 2018).

Most of the existing resources are generally centered on EU languages and do not contain useful elements for interfacing with the languages of the beneficiaries of the reception and this provides a strong limit to the use of the tools themselves in the real application field by translators, humanitarian mediators, and operators. A similar approach but with a broader scope covering general EU terminology is that of (EUROVOC 2010, IATE 2018, UNHCR 2006), containing, as a section, sets of entries pertaining to the domain of migration.

Of different character are works dedicated to the words describing immigration as associated to racism and xenophobia, also relevant and interesting for the topic but slightly out of focus for the purpose of this research (Bolaffi, Gindro & Tentori 1998, Bhopal 2004). Also, a different focus is that of Małgorzata (2016) that bears educational aims trying to describe visually concepts related to migration, forced migration, torture and related issues.

3 A Resource for Whom?

As mentioned above, and shown by the diversity of approaches and addressees of previous works, the challenges posed by the migration lexicon are based on its internal stratification depending on factors such as: national, regional, and local differences, diversity in possible audiences (from institutional international stakeholders to aid workers and finally migrants themselves) and from the intrinsic non correspondence of different migration responses along with rapid changes in legislation for each recipient country taken into consideration.

At a macro level we can organize the migration lexicon into three large sub-sets or domains differentiated by scope:

a) An international or transnational level, here identified with the institution of the European Union’s regulations (legal and administrative/institutional and its migration approach principles, e.g., Dublin regulations and its revisions). This domain is generally best described since it needs to be standardized at least for all EU languages (although not taking into consideration the languages of migrants which are seldom corresponding to any of the above-mentioned languages) – a selection of terms of this typology is that provided by the (European Migration Network 2018) that describe terms such as *Accordo di Cotonou* (‘Cotonou Agreement), *adozione fittizia* (‘adoption of convenience’), *lavoratore migrante* (‘migrant worker’), *minore non accompagnato* (‘non accompanied minor’), *paese di transito* (‘country of transit’), *permesso di soggiorno* (‘residence permit’), etc. that are horizontal to the documentation found in specific EU National regulations;

b) The second macro sub-set is national in scope, and it regards procedures, regulations, adaptations, and additions that are modified and proposed by each specific country administrative and general migration policies. Each country does in fact implement and define regulations and implementations in individual and not cross-nationally comparable ways. These regulations and implementing decrees are constantly changing depending on government direction and overturns and on public opinion stances (e.g., in Italy for example Security decrees that have deeply changed in time asylum typologies in the last few years, etc.). The second layer is country-specific and often does not offer any official or non-official translation of its content and terminology being of national, regional, or local interest. In the case of Italian, examples are *soccorso civile* ('civil aid'), *maggiorenne* ('adult'), *lettera di assunzione* ('hiring contract'), *profugo* ('refugee'), *a tempo determinato* ('fixed term'), *Decreto Sicurezza* ('Security Decree'), etc.;

c) The last sub-set of the lexicon relevant to migration management process concerns activities that are interlinked to aspects that migrants must face in their interactions with institutions (social security, health, education, administrative issues).

For all these sub-strata (which in actual use can also overlap, especially between level A and B) there is an urgent need to develop strategies and tools that cover significant gaps both at international and national levels.

The basic needs regarding glossary enrichment concern: methodology (corpus-based, corpus design and updating); inclusion of languages of migrants; inclusion of the overall stratification of domains pertaining migration not focusing only on institutional and regulative texts. Furthermore, the migration domain is affected by a widespread media usage of terms in a non-technical sense generating potentially dangerous ambiguity (*illegal and legal immigration, refugee, economic migrant*, etc.). This complexity needs to be addressed both for the updating of general lexicographic works, but also for the use of cultural mediators that are assisting migrants in their path in the new country and finally directly to migrants as beneficiaries of reference glossaries and guides.

Thus, a general resource on migration terminology serves multiple beneficiaries: general lexicographers, policy makers, aid workers and mediators, migrants themselves. This aim can be achieved by devising a source language description that is accurate but understandable, with explicit references and a focus that includes as target languages the languages of migration specifically connected to a single reception country. Furthermore, the second macro-level (national or country-based) requires the glossary to be monodirectional and poses great challenges for translation equivalents since terminology is often not internationally or transnationally shared. The "Language on the Fly" project is by design monodirectional, since it aims to consider especially the lexicon that is used in country specific national, regional, and local texts on migration and that often do not conform to international standards. The prototype itself can nevertheless be applied to any source country (and language) as a methodology and procedure for collection and description of data.

The concept behind the project and its novelty is both in trying to address a very complex stratification of lexicon that is not standardized and that, by definition includes languages that are not only typologically different from most EU languages but that are also characterized by administrative and legislative as well as social differences that are challenged in translation of official documentation. Furthermore, in some cases terminology is completely absent since most TL are not countries that do not have a significant incoming tradition in immigration so can lack sufficient corpus materials to provide comparable terminology and need compromises, paraphrases and glosses in order to be properly used. So, lexicological analysis and translation require consistent, explicit approaches and a full and comprehensive knowledge of the underlying context of the countries that represent the TL (a point particularly critical in the case of Arabic).

From a macro-structural point of view the main innovations in concept, compared to the available resources are:

a) Diverse language inclusion with a focus of TL of countries of origin of migration – not only EU languages;

b) Taking into account the time variable both at national and international level that often requires a temporal specification for definitions, since legislation on the subject has constantly changed; this also require accessible strategies to represent those changes;

c) A wider audience that varies from EU policy makers on the one hand to migrants themselves on the other, requiring complex choices in accessibility and levels of representation, and that are motivated both by linguistic and civil purposes of the work.

This prototype suggests some significant theoretical points regarding multilingual lexicography. One of the main issues regards language directions and the relationship between languages and countries where the language is spoken, since in the case of domains other than general language of more standardized portion of the lexicon, the relationship with terminology to the overall country management, and legislative and administrative apparatus is very tight. So, a relevant element of reflection regards the nature of linguistic resources themselves linked to areas similar to that of migration, that has a crucial international role, needs multilingual resources, but resources cannot be conceived as traditional *linguistic* and *translation* issues that can be represented bi-directionally without posing significant threats to the domains represented.

This theoretical point is not specific to the domain of migration, but emerges in a striking way in this domain since migration requires tools and linguistic resources that cover languages and cultures that come into contact and conflict in terms of responsibility and humanity of the addressee of documents and cultural mediation, adding to a known asymmetry the burden of bearing strong consequences on the lives of people accessing texts and asking to have their rights recognized and granted.

4 Structure of the Monolingual Section (SL)

The prototype model proposed, to be released by the end 2021, covers basic local, national, and international terminology in the relating field in Italian (as source language) and English and Arabic as target languages. Working versions are now being developed in six EU languages (Italian, English, French, Spanish, Portuguese, German) and 10 non-EU languages (Arabic, Azerbaijani, Serbian, Pashto, Russian, Persian, Albanian, Turkish, Chinese, Norwegian). Priority is given to languages that are most represented as countries of origin of migration in Italy.

The LoF project is organized as a processing cycle starting from the source language (SL), which in the prototype is Italian, and the further linking to target languages (TL) linguistic data, which in the first release will be Arabic and English. The phases of processing are organized in cycles to benefit from the empirical approach and to take into account updates in relevant documents and procedures to be considered.

For the monolingual description of lemmas (whose selection process is described in the next paragraph) the scheme of the prototype is exemplified in Figure 2.

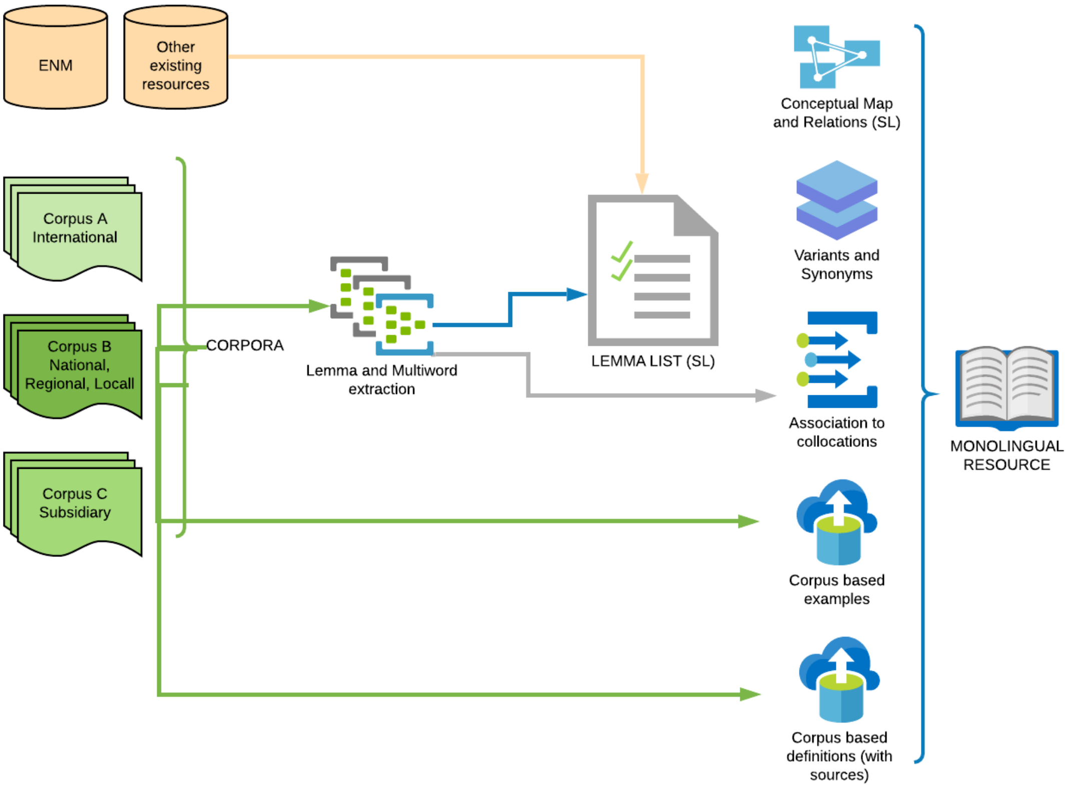

ENM
Other existing resources
Corpus A International
Corpus B National, Regional, Locall
Corpus C Subsidiary
CORPORA
Lemma and Multiword extraction
LEMMA LIST (SL)
Conceptual Map and Relations (SL)
Variants and Synonyms
Association to collocations
Corpus based examples
Corpus based definitions (with sources)
MONOLINGUAL RESOURCE

Figure 3: Monolingual SL Resource Map.

5 Building the Lemma List

As can be observed in Figure 3, the first step in glossary building is the definition of a starting lemma list in the SL. The lemma list is further increased by progressive cycles deriving from semi-automatic terminology extraction by the three Italian corpora, which are also continuously updated. As shown, the starting point is a set of lemmas already collected in previous general glossaries including Italian as SL or TL (564 lemmas).

The core of the new lemma list extraction is corpus based. Three corpora representing the three levels of stratifications have been constructed and are constantly updated. An overview of current characteristics from the corpora is:

Corpus Title	Area	Occurrences	Text Typologies
A.1 Corpus of International EU Migration (Italian)	EU	ca 5,000,000	Documentation from the European Parliament, The Council of Europe, The Court of Justice, The European Commission on migration issues
B.1 Corpus of National Migration 1 (Italian)	Italy	ca 5,000,000	Italian legislation and documentation of migration management specific to the Italian context
C.1 Corpus of Subsidiary National Migration 1 (Italian)	Italy	ca 5,000,000	Documentation for social security, health, education, administrative issues etc.

Table 1: Lof Italian SL Corpora.

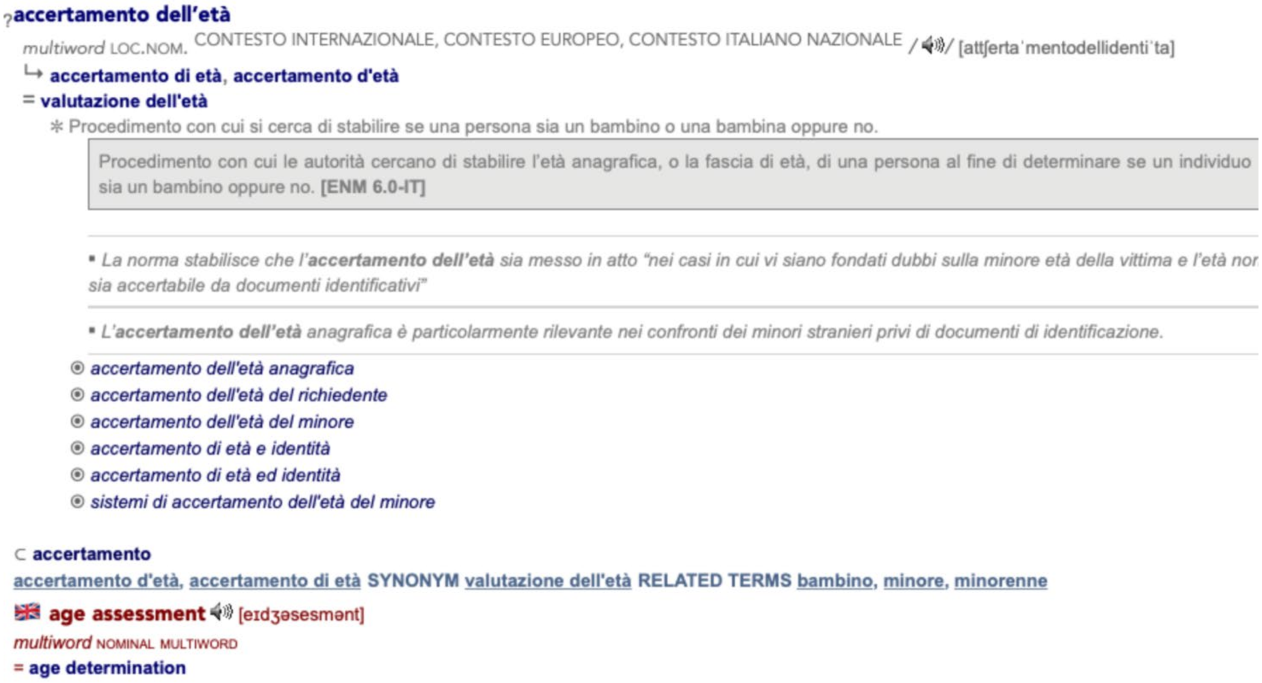
Lemmas and multiwords (including collocations) are extracted from each corpus, checked by concordance, and inserted into the LoF lemma list (which is defined by a version number depending on cycles of corpus monitor). The release of the

LoF resource is (provisionally) composed of 2,094 entries (of which 1,558 multiwords, and 89 named entities).

Following the definition of the entry list a set of procedures are applied for the description of the SL entry properties:

a) Identification of relationships (semantic and conceptual) among lemmas;
b) Identification of formal variants and synonyms;
c) Identification and association of collocations and idioms;
d) Extraction and selection of usage examples from corpora;
e) Extraction of primary source definition for key concepts with relative source.

Steps b) c) d) and e) are all performed by accessing build corpora for SL. The result of these operations is converted in a relational database form that produces a monolingual resource of terminology that will be the starting point for the processing for all target languages. For pure exemplification the entry for the SL has the following structure (but multiple layouts depending on the device and purpose of the applications), see Figure 4.



?**accertamento dell'età**

multiword LOC.NOM. CONTESTO INTERNAZIONALE, CONTESTO EUROPEO, CONTESTO ITALIANO NAZIONALE / 🔊/ [attʃerta'mentodellidenti'ta]

↳ **accertamento di età**, **accertamento d'età**

= **valutazione dell'età**

✻ Procedimento con cui si cerca di stabilire se una persona sia un bambino o una bambina oppure no.

> Procedimento con cui le autorità cercano di stabilire l'età anagrafica, o la fascia di età, di una persona al fine di determinare se un individuo sia un bambino oppure no. **[ENM 6.0-IT]**

▪ *La norma stabilisce che l'**accertamento dell'età** sia messo in atto "nei casi in cui vi siano fondati dubbi sulla minore età della vittima e l'età nor sia accertabile da documenti identificativi"*

▪ *L'**accertamento dell'età** anagrafica è particolarmente rilevante nei confronti dei minori stranieri privi di documenti di identificazione.*

◉ *accertamento dell'età anagrafica*
◉ *accertamento dell'età del richiedente*
◉ *accertamento dell'età del minore*
◉ *accertamento di età e identità*
◉ *accertamento di età ed identità*
◉ *sistemi di accertamento dell'età del minore*

⊂ **accertamento**

accertamento d'età, accertamento di età **SYNONYM** valutazione dell'età **RELATED TERMS** bambino, minore, minorenne

age assessment 🔊 [eɪdʒəsesmənt]

multiword NOMINAL MULTIWORD

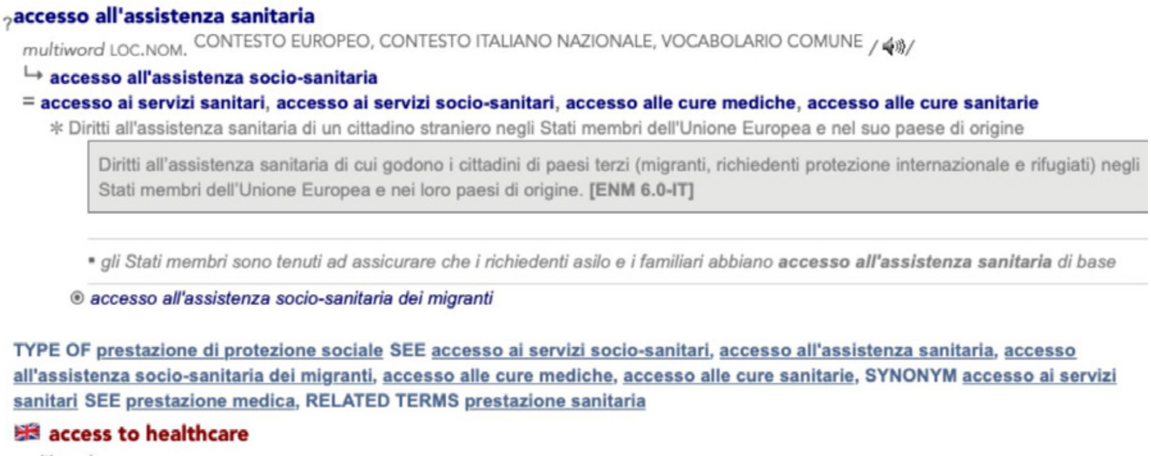
= **age determination**

Figure 4: Example *accertamento dell'età* in Monolingual SL resource database.

The provisional layout provides grammatical classes, information about the general scope (international, national, common language), pronunciation both as playable sound files and in phonetic transcription. The structure of the glossary further provides formal variants (such as for *accertamento dell'età, accertamento di età, accertamento d'età).* Formal variants have been introduced not only to take into account variability but also to enable the resource to be used potentially in text mining and in computational tools. Where available, synonyms have been provided (as in this case *valutazione dell'età*). The definition preceded by the ✻ symbol is provided in simple controlled language to be fully understandable by average users, using a prototypical approach, while the technical definition – where available - is provided in a box along with reference to the original source. Examples are all corpus based and are not manipulated in any way but chosen in order to represent common context usages of the entry, see example 1.

(1) L'accertamento dell'età anagrafica è particolarmente rilevante nei confronti dei minori stranieri privi di documenti di identificazione. [*age assessment is particularly relevant for foreign minors without identification documents*]

Finally, a list of collocations found in the corpus is given along with semantic relations and related terms, as can also be observed in Figure 5.



?**accesso all'assistenza sanitaria**

multiword LOC.NOM. CONTESTO EUROPEO, CONTESTO ITALIANO NAZIONALE, VOCABOLARIO COMUNE /🔊/

↳ **accesso all'assistenza socio-sanitaria**

= **accesso ai servizi sanitari**, **accesso ai servizi socio-sanitari**, **accesso alle cure mediche**, **accesso alle cure sanitarie**

✻ Diritti all'assistenza sanitaria di un cittadino straniero negli Stati membri dell'Unione Europea e nel suo paese di origine

Diritti all'assistenza sanitaria di cui godono i cittadini di paesi terzi (migranti, richiedenti protezione internazionale e rifugiati) negli Stati membri dell'Unione Europea e nei loro paesi di origine. **[ENM 6.0-IT]**

- *gli Stati membri sono tenuti ad assicurare che i richiedenti asilo e i familiari abbiano* ***accesso all'assistenza sanitaria*** *di base*

◉ *accesso all'assistenza socio-sanitaria dei migranti*

TYPE OF prestazione di protezione sociale SEE accesso ai servizi socio-sanitari, accesso all'assistenza sanitaria, accesso all'assistenza socio-sanitaria dei migranti, accesso alle cure mediche, accesso alle cure sanitarie, SYNONYM accesso ai servizi sanitari SEE prestazione medica, RELATED TERMS prestazione sanitaria

access to healthcare

Figure 5: Example *accesso all'assistenza sanitaria* in Monolingual SL resource database.

6 Building the TL Macro-Structure

The following steps regarding the work on TL involves different procedures depending on languages and availability of corpus resources to rely on for the translations, mapping, examples and definitions. For Corpus A (International EU) we have built parallel corpora at least for the EU languages included in the project and comparable corpora for non-EU languages, where possible, facing a number of challenges. The sub-set of the lexicon extracted from Corpus A is generally more widespread and standardized, so the process of extracting TL entries, synonyms, examples and definitions from parallel or comparable corpora poses less of a challenge. To this set belong entries such as: *Accordo di Cotonou, beneficiario di protezione internazionale, cittadino di un paese terzo, Convenzione di Dublino, discriminazione diretta, migrante di seconda generazione*, etc.

One of the main challenges remains the absence of a common migration framework that grant good quality translations and do not introduce potentially risky ambiguities. In this respect glosses are provided for translations in cases where the cultural and legislative background demands it.

Processing of Corpus B (National, Regional and Local Scope) is the biggest challenge since it contains terminology, concepts and entities that are country specific and that derange significantly even from the EU given framework. This peculiarity pushes to rely on new terms in TL in order not to generate ambiguity with close or similar terms common to the international terminology. Examples of entries extracted from the B corpus of Italian are: *Agenzia del demanio, ente territoriale, certificazione sanitaria, ufficio G.I.P., SPRAR, abuso della libera circolazione, associazioni del Terzo Settore, autonomie locali*, etc.

Corpus C (Subsidiary) concerns additional aspects in migration management which are not directly linked to the process of asylum and are related to issues like social security, health, education, and administrative issues. From the point of view of corpus collection and comparative or parallel corpora build it can be considered of medium complexity since in many countries forms and infos about these subject matters are also available in different language and are generally less sensible in content and nature (e.g., *carta di identità, stato civile, certificate di matrimonio, tessera sanitaria*, etc.).

At the moment of writing the quantitative data about entries description is the following:

Languages (ISO codes)	# entries	of which # multiwords	variants	synonyms
IT	2,094	1,558	455	970
EN	2,094	1,609	519	1,196
AR	1,009	764	0	22

Table 2: Lof entries described for SL and TLs.

The structure of TL entries linked to the SL entries is the same as the monolingual side. This uncovers many differences in word usages and helps identify asymmetric relations among entries, both simple lemmas and multiwords. For example, observing the English collocations associated with *age assessment* we find many that are not represented in the Italian corpus and in the entry such as *errors in age assessment; means of age assessment; to make age assessment; age assessment procedures; to carry out the age assessment; age assessment guidance; the process of age assessment; integrated age assessment*. We also find different synonyms to be taken into account such as *age determination*. A similar situation happens with the correspondent Arabic where *accertamento dell'età* can be translated by two multiwords تقدير السنّ or تقدير العمر. In this case there is no substantial difference in the two and the terminology is fairly standardized in Arabic. A slightly different case is that of *permesso di soggiorno* that in English is the standard *residence permit* while in Arabic, although strongly associated to تصريح الإقامة in a general sense, can also be found in corpora a form less standardized as إذن الإقامة that would be "permission to reside" or "authorization of residence" along with *residence* *permit*.

7 Challenges

The process of building a corpus based (country-based) glossary of migration terminology has proven to provide many challenges in the linking of data and especially in working on languages that possess a completely different administrative, legislative, and generally cultural background. The case of Arabic is particularly significant since it includes dozens of linguistic varieties being spoken by around 422 million speakers in 25 different countries. The challenge is not only cultural but specifically linguistic since some expressions are peculiar only to some areas and not to other, although are still considered Standard Arabic (and not dialectal forms). The linking between the monolingual SL structure to the TLs structures is still to be perfected to assure the user the capability of moving themselves through the different information provided in the entry and their translation equivalents.

The most significant base for new entries extraction has been Corpus B, since it is the corpus that is strongly related to the migration process description but nevertheless contains terminology, which is not shared in EU or international documentation, since it depends on the country choices in legislation and administration and at the same time very rarely possesses official translations in any language other than the SL (both EU and non-EU languages).

Working on this kind of terminology further exposes significant problems in the difference in policies that EU countries adopt as reflected by the language chosen and by the political implication of administrative and legislative determinations. Thus, working on migration terminology has proven to be a challenge that needs to be further taken into account to provide better services and to assure the democratic access to basic human rights, which are also guaranteed by linguistic choices and accessible description.

8 References

Bhopal, R. (2004). Glossary of terms relating to ethnicity and race: for reflection and debate. In *Journal of Epidemiology & Community Health*, 58(6), pp. 441-445.

Bolaffi, G., Gindro, S. and Tentori, T. (1998). *Dizionario della diversità: le parole dell'immigrazione, del razzismo e della xenofobia*. Firenze: Liberal Libri.

European Migration Network (2018). *Asylum and Migration. Glossary 6.0*. Belgium: European Migration Network (EMN).

EuroVoc Thesaurus - The EU's Multilingual Thesaurus. Accessed at: https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc [18/04/2021].

Feigelfeld, P., (2016) *Refugee Phrasebook - Juridicial Phrases for Refugees*. Accessed at: https://www.refugeephrasebook.de/ [18/04/2021].

Glossario. Accessed at: http://www.programmaintegra.it/wp/risorse/glossario/ [18/04/2021].

Gorsau, H. (2015). *Special dictionary for Syrian refugees, migrants and asylum seekers travelling towards Europe*. St-Orens, France: Editions Goursau.

Interactive Terminology for Europe. Accessed at: https://iate.europa.eu/home [18/04/2021].

International Organization for Migration (IOM) (2019). *International Migration Law N°34 - Glossary on Migration*. Interactive Terminology for Europe: International Organization for Migration (IOM).

Małgorzata, T. (2016). *Visual Icon Dictionary on Migration*. Poland: One World Association

Perruchoud, R. (2004). *International migration law: Glossary on migration*. International Organization for Migration.

The Refugee Phrasebook. Accessed at: https://en.wikibooks.org/wiki/Refugee_Phrasebook [25/03/2021].

UNHCR (2006). *UNHCR Master of Glossary of Terms*. Geneva 2, Switzerland: U. N. H. C. f. Refugees.

Using verb patterns to find recurrent metaphors in corpus

Renau I.

Pontificia Universidad Católica de Valparaíso
irene.renau@gmail.com

Abstract

In this study, we examine the possibility of finding regularities in combinations of verb patterns, and if these regularities can be used to find recurrent metaphors in discourse. As the source of the data, we used Verbario, a database of 227 Spanish verbs that were annotated with the Corpus Pattern Analysis technique (Hanks 2004, 2013). We restricted our analysis to transitive patterns in order to have identical syntactic structures and be able to focus our analysis to semantic types only. Given a verb pattern such as *[[Humano]] guarda [[Objeto Físico]] ([[Human]] keeps [[Physical Object]]),* the base pattern is *[[Human]] ~ [[Physical Object]],* a syntacto-semantic structure that can be found also in verbs other than *guardar.* 177 verbs from the database (78%) had 2 or more transitive structures and were included in the study. Results show how a small number of semantic types and combinations of verb patterns are linked to most of the verbs. Additionally, many pairs of base patterns are connected to each other through metaphors. The study is of interest for lexicographic tasks involving corpus analysis and is a contribution to corpus-based studies of metaphor.

Keywords: Corpus Pattern Analysis; metaphor; polysemy; semantic type; Spanish

1 Introduction

Metaphors have often been classified as irregular polysemy, especially in contrast with metonymy, which is usually linked to regular polysemy (Apresjan 1974). While polysemy by metonymy is more systematic and predictable, polysemy by metaphor is more idiosyncratic and accidental, and it is not predictable. However, the theory of metaphor (Lakoff & Johnson 1980) postulates that metaphor, as well as metonymy, are cognitive (not only linguistic) resources that people exploit to categorize the world and communicate: thus, metaphors have to hold a certain degree of regularity in order to be used and understood. A metaphor such as "Humans are Machines" is not predictable nor systematic, but it lays beyond many linguistic expressions such as *My mind does not work well today, I have too many memories in my hard drive,* etc. This is precisely what conceptual metaphors are. The same happens with conceptual metonymies: a metonymy such as "Plant Part for the Plant" is a cognitive resource that can be exploited, but we cannot predict when a plant part is going to be labeled with the name of the plant, and the same happens with colors/flours, products/plants, etc. (Renau 2021). Hence, there is probably not a sharp distinction between regular polysemy, irregular polysemy and homonymy, but a gradual distinction (Moldovan 2019).

In this study, we propose a method to find regularities in combinations of verb patterns which could be used to find recurrent metaphors in discourse. We take the concept of verb pattern from Hanks' Corpus Pattern Analysis, CPA (Hanks 2004, 2013, among others). The author (Hanks 2004: 87) states that word meaning is associated to "syntagmatic patterns with which words in use are associated". Thus, in real-life discourse, these patterns, consisting of the basic valency structure and other semantic and syntactic features, are the ones carrying the meaning of the verb, and not the verb in isolation. For our investigation, we use a database of Spanish verbs (Renau et al. 2019) that were annotated following the CPA principles. Given two verb patterns, we examine the possibility of finding an equivalent pair of patterns in other verbs. Observe the following examples:[1]

(1) Verb *ensanchar* 'to widen'
Pattern 1: [[Humano]] ensanchar [[Objeto Físico]] *([[Human]] widens [[Physical Object]])*
Example: "En verano ensancharon el camino" *('In the summer they widened the road').*

Pattern 2: [[Humano]] ensanchar [[Entidad Abstracta]] *([[Human]] widens [[Abstract Entity]])*
Example: "[Ustedes] sabrán preservar y ensanchar nuestra armónica convivencia" *('You will know how to preserve and widen our harmonious coexistence').*

(2) Verb *guardar* 'to keep'
Pattern 1: [[Humano]] guarda [[Objeto Físico]] *([[Human]] keeps [[Physical Object]])*
Example: "Compró el libro y lo guardó en el bolsillo del abrigo" *('[She/he] bought the book and kept it in the pocket of her/his coat').*

[1] See the complete, original analysis of all Spanish patterns in the Verbario database (http://www.verbario.com). For clarity's sake, in the paper we simplified some of the patterns. All examples are from the EsTenTen corpus, available in the Sketch Engine (Kilgarriff et al. 2014).

Pattern 2: [[Humano]] guarda [[Entidad Abstracta]] *([[Human]] keeps [[Abstract Entity]])*
Example: "Ya están grandes y saben lo que implica guardar un secreto" *('They are grown up enough to know what it means to keep a secret').*

In (1), pattern 1 and pattern 2 of *ensanchar* 'to widen' are linked by a metaphor whereby an [[Abstract Entity]] is categorized as a [[Physical Object]] that can be "widen". We could formulate this metaphor as "Abstract Entities are Physical Objects", which is one of the most used conceptual metaphors. The interesting is that we can find the same relation in other verbs, such as *guardar* 'keep' in (2). Pattern 1 and pattern 2 of both verbs share the same semantic types in the argument structure. These common patterns could be expressed as follows:

(3) *Base pattern 1:* HUMAN ~ PHYSICAL OBJECT
Base pattern 2: HUMAN ~ ABSTRACT ENTITY

We call *base pattern* to the abstract pattern consisting of the semantic types and argument structure only, without the verbs and with no link to any specific meaning. (Base patterns are indicated with small capitals.)
Following this rationale, we wonder if we can find similar associations such as the ones shown in examples (1) and (2) by extracting base patterns such as (3) from the CPA patterns that we already have in our database. Do all these associations have a metaphorical nature? Are they relatively stable? Having a base pattern X, can we predict that a base pattern Y is going to appear in the same verb?
This study is of interest for lexicography by contributing to automatic techniques to interrogate a corpus for lexicographic purposes (Kosem 2016). Specifically, it could be of help regarding the so-called "pre-lexicographic" or "preliminary" tasks of the dictionary-making process (Atkins & Rundell 2008, Hartmann 2001), particularly for collecting and analyzing corpus data. Corpus analysis is necessary as the empirical basis of the information offered in a dictionary, but it is still a very time-consuming and complicated task which has to be executed manually to a large extent. If we can find that two base patterns are regularly found together in verbs, we can help the lexicographer by making suggestions while she/he is annotating the corpus. This study contributes to a more enriched and complex corpus annotation in which the system can help to find semantic regularities instead of a list of concordances with no inter-connection. This proposal could also be of help to providing clues for meaning differentiation and ordering in the dictionary entry (Jiang and Chen 2017).

2 Theoretical and Methodological Framework

The CPA patterns such as the ones shown in examples (1) and (2) are pieces of phraseology that are found frequently in a corpus. Each meaning of the verb is linked to one or more patterns of use. Observe the following example:

(4) [[Human]] keeps [[Physical Object]] in [[Location]]

A pattern such as (4) is mapped to the meaning 'to store' of the verb *to keep.* This shows that the different meanings of the verb *to keep* can be disambiguated in context by analyzing the argument and syntactic structure of the verb and categorizing the arguments with semantic types. This theoretical and methodological line of research has its roots in a number of authors who observe that word meanings are disambiguated by context, e.g., Malinowski (1923), Firth (1935), Sinclair (1998) and Pustejovsky (1995), among others (see Hanks 2013 for a more detailed approach to the theoretical background of this technique). Firth (1935: 7) early states that "the complete meaning of a word is always contextual, and no study of meaning apart of a complete context can be taken seriously". According to the same author (Firth 1935: 7), this principle is what makes "systematic use of quotations or context" in dictionaries a crucial element for lexicographic representation of meaning. In the same way, CPA is a proposal for systematic corpus analysis of words, in which syntagmatic context of a verb in real discourse is analyzed and mapped into a meaning.
As already stated, verb patterns consist of basic valency structure, but an appropriate semantic categorization of each argument is also necessary. In (4), many words or phrases can be the subject of the verb (e.g., *student, you, Veronica, the new owners, we,* etc.), and all of them are unified under the same semantic type, [[Human]]. The same happens with [[Physical Object]] and [[Location]]. Semantic types are semantic categories that, in CPA, connect each other in an ontology of around 250 labels. Thus, while the verb *to keep* is ambiguous in isolation, patterns are unambiguous. In the present proposal, we pay special attention to semantic types and how they play a role in the configuration of verb meanings. For example, we can observe that pattern 1 and 2 in examples (1) and (2) have the same syntactic structure (transitive), but there is a variation in the semantic type of the direct object: in both examples (1) and (2), pattern 1 has [[Physical Object]] as direct object, while pattern 2 has [[Abstract Entity]]. This variation alone allows to differentiate the patterns, which are mapped onto different meanings.
Normal patterns such as (4) can be exploited for creative purposes or for fitting a specific communicative situation. For example, a sentence such as "There will be a large freezer for keeping food" can be considered a normal use of pattern (4), but a sentence such as "There will be a large freezer for keeping your pleasure" is not so common, but one can understand that there is a game of words in which the pleasure ([[Abstract Entity]]) a person gets by eating the food ([[Physical Object]]]) stored in the freezer is materialized as something one can eat. Both norms and exploitations usually have their origin in a metaphor, and can be understood because, as stated in the introduction, metaphors are cognitive devices which are shared by the members of a community.
CPA is being used to build the *Pattern Dictionary of English Verbs* (Hanks, online) and the Verbario database for Spanish verbs (Renau, http://www.verbario.com). So far, Verbario contains 227 verbs, 1,233 patterns and 84,227 manually analysed concordances which are linked to the patterns. All CPA projects use the same method and ontology, which makes the data compatible (as was shown in Baisa et al. 2016) and allows to test the present proposal in other languages. For the purposes of the present study, the limitation of the technique lays on the fact that it is basically manual in spite of some attempts to

automatize certain parts of the task (Renau et al. 2019). Therefore, the analysis could be biased by the different annotators, and we do not possess data on a large-enough scale to generalize our results. Hence, the present study is explorative, and we expect to address these limitations in the future work.

3 Methodology

As already explained, we used Verbario's database as the source of our study units. For this preliminary study, we restricted our analysis to transitive patterns, that is, we included in our analysis only those verbs with 2 or more transitive patterns, with argument 1 as subject and argument 2 as direct object. We included in our study those patterns having a 3rd argument (e.g., adverbial or indirect object), but argument 3 was excluded from the study, because we were looking for identical syntactic structures in order to have semantic types as the only variable. We leave for future work to compare other possible structures and arguments, such as intransitive patterns *(Argument 1 + verb)* or trivalent structures with direct object or adverbial as argument 3. According to this, the first operation was to delete all the patterns with less than 2 transitive patterns, and for the remaining patterns, to keep the structure *Argument 1 + verb + Argument 2* and delete the rest of the pattern.
The second step was to transform the patterns into base patterns by deleting the verb and keeping the semantic types, for example:

(5) *Pattern 1 of "ensanchar":* [[Humano]] ensanchar [[Objeto Físico]]
Pattern 1 of "guardar": [[Humano]] guarda [[Objeto Físico]]
} *Base pattern 1:* Humano ~ Objeto Físico

(6) *Pattern 2 of "ensanchar":* [[Humano]] ensanchar [[Entidad Abstracta]]
Pattern 2 of "guardar": [[Humano]] guarda [[Entidad Abstracta]]
} *Base pattern 2:* Humano ~ Entidad Abstracta

(See (1), (2) and (3) for clarification.)

In case of semantic alternations–that is, when more than one semantic type alternate in the same argument (Hanks, 2013: 176-180), we split the verb pattern in each of the semantic types involved and created one base pattern for each semantic type. Observe the following example:

(7) Verb *acarrear* 'to carry'

Pattern 1: [[Humano | Vehículo]] acarrear [[Objeto Físico]] *([[Human | Vehicle]] carries [[Physical Object]])*

Vertical line between [[Human]] and [[Vehicle]] in example 7 means that both semantic types can be the subject of the verb in pattern 1 of *acarrear* 'to carry'. This alternation does not change the meaning of the pattern, which in this case is 'to take something somewhere'. In this respect, in example 8, we can observe a sentence in which we have [[Human]] as subject (*los niños* 'the children') and another one in which we have [[Physical Object]] as subject (*una caravana de camiones* 'a caravan of trucks'). Both sentences are linked to pattern 1 in example (7).

(8) *Example (for [[Human]] as subject):* "*Los niños* enfrentan riesgos de seguridad y de salud al tirar y acarrear cargas pesadas" ('*Children* deal with security and health risks when pulling and carrying heavy loads').

Example (for [[Vehicle]] as subject): "*Una caravana de camiones* se encarga de acarrear las provisiones" ('*A caravan of trucks* is responsible for carrying the provisions').

Hence, in our study, for the [[Human]] variant of the pattern, we created the base pattern Human ~ Physical Object, and for the [[Vehicle]] variant, we created the base pattern Vehicle ~ Physical Object.
Once the base patterns were extracted, we created a matrix where each column is a base pattern and each row a verb. We added the number of the pattern to the cell when there was a match (see table 1 for examples).

	Eventuality ~ State of Affairs	Eventuality ~ Cognitive State	Eventuality ~ Event	Eventuality ~ Human Group	Eventuality ~ Human	Physical Object ~ Physical Object
abrasar 'to burn'	0	0	0	0	3	2
abrigar 'to wrap up'	0	0	0	0	0	0
abrir 'to open'	0	16	0	0	0	0
acarrear 'to carry'	3	0	0	0	0	0
aconsejar 'to advise'	0	0	0	0	0	0
acortar 'to shorten'	0	0	0	0	0	0
acosar 'to harass'	0	0	0	0	3	0

acostar ‘to lay down’	0	0	0	0	0	0
activar ‘ to activate’	0	0	0	0	0	0
agrietar ‘to crack’					2	1

Table 1: Fragment of the matrix for illustration.

In table 1 we observe how, in this fragment of the matrix, there are 7 matches between verbs and base patterns. For example, pattern 16 of *abrir* ‘to open’ has EVENTUALITY ~ COGNITIVE STATE as verb pattern.

After this operation, we calculated the frequency in which a base pattern X appeared together with base pattern Y in the same verb, and in how many verbs we have the same coincidence. For example, in table 1, *abrasar* and *acosar* share the base pattern EVENTUALITY ~ HUMAN, but these verbs do not share any other pattern –nor in table 1, which is a fragment for illustration, nor in any other part of the whole matrix. This means that this combination of two verbs is not a candidate to find possible pairs such as the ones shown in (7) and (8). Conversely, *abrasar* and *agrietar* ‘to crack’ do share two combinations: PHYSICAL OBJECT ~ PHYSICAL OBJECT and EVENTUALITY ~ HUMAN. These combinations are the target of our study. In this case, for example, these pairs of base patterns exhibit a metaphorical relation in which [[Eventuality]] behaves with [[Human]] the same way as [[Physical Object]] with another [[Physical Object]]. Table 2 shows this example in more detail: we observe how the event of ‘burning something physical’ (in the case of *abrasar* ‘to burn’) is transferred to ‘emotionally burning a person’. In parallel, the event of ‘cracking something physical’ (in *agrietar* ‘to crack’) is transferred to ‘morally cracking a person’. The same metaphor underlies both verbs in the same way.

Base pattern	Verb	Pattern	Implicature	Example
PHYSICAL OBJECT ~ PHYSICAL OBJECT	*abrasar* ‘to burn’	*Pattern* 2 [[Objeto Físico]] abrasar [[Objeto Físico]] *([[Physical Object]] burns [[Physical Object]])*	*[[Physical Object]] makes that [[Physical Object]] is very hot.*	…los hierros abrasando la carne. *(…the irons burning the flesh.)*
	agrietar ‘to crack’	*Pattern 1* [[Objeto Físico]] agrietar [[Objeto Físico]] *([[Physical Object]] cracks [[Physical Object]])*	*[[Physical Object]] makes that cracks appear in [[Physical Object]].*	Una bola de granizo agrieta tu cristal. *(A hail ball cracks your glass.)*
EVENTUALITY ~ HUMAN	*abrasar* ‘to burn’	*Pattern 3* [[Eventualidad]] abrasar [[Humano]] *([[Eventuality]] burns [[Human]])*	*[[Eventuality]] has an strong and negative effect on [[Human]].*	Seguir pensando [en que] no sé qué hacer con mi vida me abrasaba. *(Continuing thinking that I do not know what to do with my life burned me.)*
	agrietar ‘to crack’	*Pattern 2* [[Eventualidad]] agrietar [[Humano]] *([[Eventuality]] cracks [[Human]])*	*[[Eventuality]] weakens [[Human]], makes her/him lose her/his power or strength.*	Esa tarde, otra vez lo agrietó el descreimiento. *(That afternoon, disbelief cracked him again.)*

Table 2: An example of a match of base patterns in two verbs: *abrasar* ‘tu burn’ and *agrietar* ‘to crack’ (see table 1, green cells). The implicatures are paraphrases of the patterns which explain their meanings (Hanks, 2013).

Finally, we also calculated an association coefficient which indicated the grade of reciprocity in which a pair of base patterns appears in the same verb. To do this, we applied the following formula:

$$\frac{f\,(i\;j)}{\sqrt{f\,(i)}\cdot\sqrt{f\,(j)}}$$

i and *j* are 2 base patterns appearing in the same verb. Frequency of *i* and *j* appearing together is divided by the total frequency of *i* per the total frequency of *j.* Square root is used to mitigate the difference between the highest and lowest numbers. Numbers were multiplied by 100 in order to avoid decimals.

Results of frequency and association were displayed in a table.

4 Results and Discussion

Of the total 227 verbs in the database, we found 177 with 2 or more transitive patterns (78% of the verbs), that is to say, 78% of the verbs were included in the study (in the matrix, they were displayed in the rows, as shown in table 1). Of these verbs, we obtained 510 base patterns (in the matrix, they were displayed in the columns, as shown in table 1). 32 of the base patterns (6,27%) associate with another base pattern creating 77 pairs. These pairs appear in 111 of the 177 verbs (62,7%). This means that only a few base patterns combine with another base pattern two or more times, but at least one of these combinations can be traced in most verbs. For example, the combination HUMAN ~ EMOTION → HUMAN ~ HUMAN is very frequent (n = 5, 71%): thus, in 71% of the verbs in which we find the base pattern HUMAN ~ EMOTION, we also find the base pattern HUMAN ~ HUMAN. Combinations are not commutative, for example, the combination HUMAN ~ HUMAN → HUMAN ~ EMOTION is very rare: only in 5% (n = 5) of the verbs in which we find the base pattern HUMAN ~ HUMAN the base pattern HUMAN ~ EMOTION is present too. This is normal, as HUMAN ~ HUMAN is much more frequent per se than HUMAN ~ EMOTION.

Figure 1 shows the semantic types which were found more frequently (> 10) in the base patterns of the data sample:

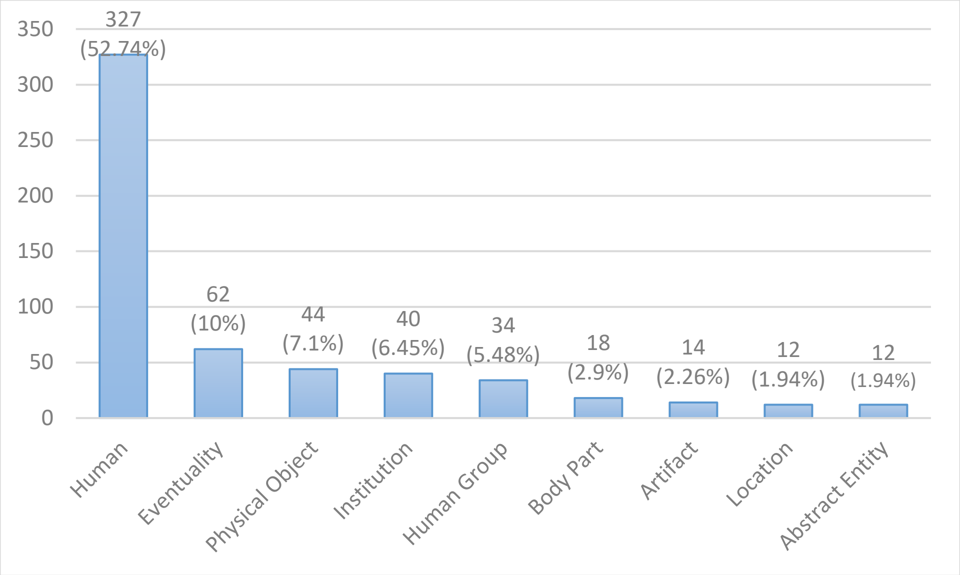


Figure 1: The most frequent semantic types taking part in recurrent combinations of base patterns.

It is not surprising to find these semantic types in figure 1 because they are common categories found in general in CPA patterns. In addition, we can observe that they take part in the formation of many metaphors, e.g. “Humans are Artifacts”, “Abstract Entity is Physical Object”, “Physical Object is Human”, etc.

Figure 2 shows the most frequent (> 5) base patterns found in a recurrent association with another base pattern:

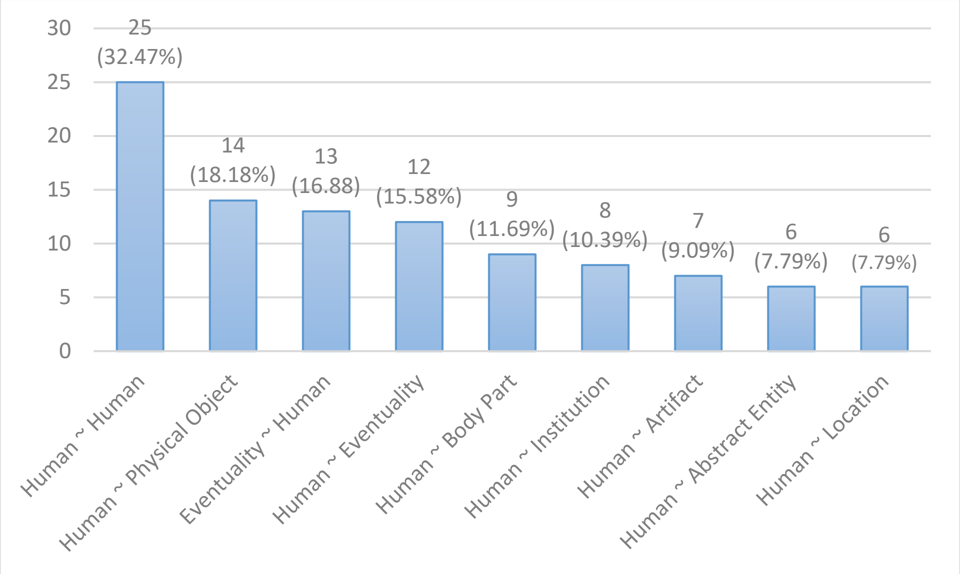


Figure 2: The most frequent base patterns found in association with other base patterns in the data sample.

Considering the data shown in figure 2, it is normal to find HUMAN as a frequent semantic type in table 2 as well. We find

it usually as the subject of the sentence, probably as agent of the event which is directed to abstract or concrete entities, or eventualities.
Table 3 shows the most frequent (> 5) associations of base patterns in the sample:

Base pattern 1 →	**Base pattern 2**	**Verbs (n)**	**Verbs (%)**
Human ~ Human	Human ~ Physical Object	19	10,73
Human ~ Eventuality	Human ~ Human	14	7,91
Eventuality ~ Human	Human ~ Human	11	6,21
Human ~ Body Part	Human ~ Human	11	6,21
Human ~ Artifact	Human ~ Human	10	5,65
Human ~ Abstract Entity	Human ~ Physical Object	9	5,08
Human ~ State of Affairs	Human ~ Human	9	5,08
Human ~ Event	Human ~ Human	9	5,08
Human ~ Human	Human ~ Location	9	5,08
Human ~ Human	Physical Object ~ Physical Object	9	5,08

Table 3: The most frequent associations of base patterns in the sample. Percentages are in relation with the total 177 verbs of the sample.

These results do not necessarily show that the association has a metaphorical nature, but coincide with types of metaphors which have been reported in the literature, and we could hypothesize that they are clues to identify metaphors. For example, in the pair Human ~ Human → Human Physical Object, the underlying metaphor could be "Humans are Physical Objects". The logic under these associations would be that certain events (represented by the verb) could be the source domain for other types of events which, as target domain, are characterized as having certain similarities with the source domain. In section 5 we will present a case study to try to show this rationale.
Table 4 shows the frequency in which a base pattern appears in a verb together with another base pattern:

Base pattern 1 →	**Base pattern 2**	**Verbs (n)**	**Verbs (%)**
Eventuality ~ Eventuality	Human ~ Human	6	100
Human ~ Anything	Human ~Eventuality	6	100
Physical Object ~ Human	Human ~ Human	5	100
Eventuality ~ Psychological Trait	Human ~ Human	4	100
Institution ~ Event	Human ~ Human	4	100
Human ~ Stuff	Human ~ Human	7	85
Eventuality ~ Eventuality	Eventuality ~ Human	5	80
Human ~ Emotion	Human ~ Human	7	71
Human ~ Anything	Human ~ Human	6	66
Human ~ Information	Human ~ Human	11	63
Human Group ~ Human	Institution ~ Institution	7	57
Human Group ~ Human	Institution ~ Human	7	57
Human ~ Stuff	Human ~ Abstract Entity	7	57
Human Group ~ Human	Human ~ Institution	7	57
Human ~ Stuff	Eventuality ~ Human	7	57
Human ~ Abstract Entity	Human ~ Physical Object	17	52
Human ~ State of Affairs	Human ~ Human	17	52

Table 4: Frequency of combinations of base patterns (> 50% verbs).

Table 4 shows how in 17 of the 77 combinations of patterns (22,07%) we can find a frequent combination. In particular, in 5 of the verbs (6,5%) the association between base patterns 1 and 2 covers 100% of the cases, that is, each time that we find base pattern 1, we also find base pattern 2. These results show that base patterns in column 1 are good predictors for the existence of base patterns in column 2 in the data sample. If these results were corroborated by studies with larger data, the method would be appropriate as an assistance to detect new meanings in corpus.
Finally, table 5 shows the highest (> 30) association coefficients of the pairs.

Base pattern 1 →	**Base pattern 2**	**Verbs (n)**	**Coefficient association**	**Verbs (%)**
Human Group ~ Human	Institution ~ Institution	4	47	57%
Institution ~ Institution	Human ~ Human	4	47	40%
Human Group ~ Human Group	Institution ~ Institution	4	44	50%
Institution ~ Institution	Human Group ~ Human Group	4	44	40%
Human ~ Anything	Human ~ Eventuality	6	41	100%
Human ~ Eventuality	Human ~ Anything	6	41	17%
Human Group ~ Human	Institution ~ Human	4	39	57%
Institution ~ Human	Human Group ~ Human	4	39	26%
Human ~ Stuff	Human ~ Abstract Entity	4	36	57%
Human Group ~ Human Group	Institution ~ Human	4	36	50%
Institution ~ Human	Human Group ~ Human Group	4	36	26%
Human ~ Abstract Entity	Human ~ Stuff	4	36	23%
Eventuality ~ Eventuality	Eventuality ~ Human	4	32	80%
Human ~ Abstract Entity	Human ~ Physical Object	9	32	52%
Human ~ Physical Object	Human ~ Abstract Entity	9	32	20%
Eventuality ~ Human	Eventuality ~ Eventuality	4	32	12%
Institution ~ Institution	Human ~ Artifact	5	31	50%
Human ~ Physical Object	Human ~ Human	19	31	43%
Human ~ Human	Human ~ Physical Object	19	31	22%
Human ~ Artifact	Institution ~ Institution	5	31	19%

Table 5: Coefficient association > 30 for the base pattern pairs. Frequency is shown for reference.

Table 5 shows that these pairs exhibit a strong association. The association is not reciprocal, though: for example, while the base pattern Human ~ Anything appears together with Human ~ Eventuality in the same verb 100% of the times, it is not as frequent that Human ~ Eventuality appears together with Human ~ Anything.

5 Case Studies

In this section, we give a more detailed description of results regarding some of the pairs of base patterns in the data sample. We want to observe if, as we stated in section 1, we could find potential sources for metaphors. Table 6 shows results for the pair Human ~ Human / Physical Object ~ Human.

Base pattern 1 →	**Base pattern 2**	**Verbs (n)**	**Association coefficient**	**Frequency**	**Verbs**
Human ~ Human	Physical Object ~ Human	5	24	5%	*aplastar, cubrir, dañar, estorbar, estremecer* ('to crush, to cover, to harm, to hinder, to shake')
Physical Object ~ Human	Human ~ Human			100%	

Table 6: Frequency and coefficient association for the pairs Human ~ Human → Physical Object ~ Human and for Physical Object ~ Human → Human ~ Human.

Table 6 shows that the association Human ~ Human → Physical Object ~ Human is very infrequent (5%), while the association Physical Object ~ Human → Human ~ Human takes place 100% of the times. The verb patterns are the following ones:

(9) Verb *aplastar* 'to crush'
Pattern 1: [[Objeto Físico]] aplasta a [[Humano]] *([[Physical Object]] crushes [[Human]])*
Example: "Una mujer de 54 años falleció ayer aplastada por un vehículo de limpieza" *('A 54-year-old woman died yesterday crushed by a cleaning vehicle').*

Pattern 2: [[Humano]] aplasta a [[Humano]] *([[Human]] crushes [[Human]])*
Example: "Esta joven de progresión imparable aplasta a sus rivales sin compasión" *('This young woman of*

unstoppable progression crushes her rivals without compassion').

(10) Verb *cubrir* 'to cover'
Pattern 1: [[Objeto Físico]] cubre a [[Humano]] *([[Physical Object]] covers [[Human]])*

Example: "Detestan las pieles que las cubren [a las mujeres ricas]" *('They hate the furs that cover them [the rich women]').*

Pattern 2: [[Humano]] cubre a [[Humano]] *([[Human]] covers [[Human]])*
Example: "Tengo miedo de que todavía se estén cubriendo unos a otros" *('I am afraid they are still covering for each other').*

(11) Verb *dañar* 'to harm'
Pattern 1: [[Objeto Físico]] daña a [[Humano]] *([[Physical Object]] harms [[Human]])*
Example: "Esta potente escopeta puede dañar a varios enemigos" *('This powerful shotgun can harm several enemies').*

Pattern 2: [[Humano]] daña a [[Humano]] *([[Human]] harms [[Human]])*
Example: "Sabe expresar sus emociones como las siente sin dañar a los demás" *('She/he knows how to express her/his emotions without harming others').*

(12) Verb *estorbar* 'to hinder, to disturb'
Pattern 1: [[Objeto Físico]] estorba a [[Humano]] *([[Physical Object]] hinders [[Human]])*
Example: "Esquivaba los pocos autos que le estorbaban en el camino" ('She/he dodged the few cars that stood in her/his way').

Pattern 2: [[Humano]] estorba a [[Humano]] *([[Human]] bothers [[Human]])*
Example: "Me dirigí al jardín para no estorbar a los adultos" *('I went to the garden to not disturb the adults').*

(13) Verb *estremecer* 'to shake'
Pattern 1: [[Objeto Físico]] estremece a [[Humano]] *([[Physical Object]] makes [[Human]] shake)*
Example: "Ese cuadro la estremece como ninguna otra cosa" *('That painting makes her shake like nothing else').*

Pattern 2: [[Humano]] estremece a [[Humano]] *([[Human]] harms [[Human]])*
Example: "El predicador la estremecía con sus emociones personales" *('The preacher shook her with his personal emotions').*

All examples (9) to (13) exhibit a metaphorical relation between base pattern 1 and 2. Metaphors are based on the categorization of [[Humans]] as [[Physical Objects]] that can make actions to other [[Humans]] which are similar to the ones that [[Physical Objects]] can do to [[Humans]]. We could formalize the metaphors underlying patterns (9) to (13) as follows:

- (9) "Somebody morally crushing somebody is an object physically crushing her/him".
- (10) "Somebody covering for somebody is an object covering her/him".
- (11) "Somebody morally harming somebody is an object physically harming somebody".
- (12) "Somebody disturbing somebody is an object physically disturbing her/him".
- (13) "Somebody emotionally shaking somebody is an object physically shaking her/him".

As we can observe, this formulation is rich in information and explain the semantic and cognitive relation between two meanings of a verb, and between different verbs. All of them share the basic idea that events caused by a [[Physical Object]] and experimented by a [[Human]] can be used to understand moral or emotional events. These preliminary findings are promising in the sense that they could be empirical linguistic data to corroborate the theory of conceptual metaphor.
It is interesting, though, that there are cases in which the two base patterns do not have a metaphorical relation, because both are metaphors of another base pattern. Observe the following example taken from the pair Human ~ Human / Human ~ Physical Object:

(14) Verb *comer* 'to eat, win, crash'

Pattern 1: [[Humano]] (se) come a [[Humano]] *([[Human]] beats [[Human]])*
Example: "No podemos dejar que nos coman, hemos de imponernos" *('We cannot let them beat us, we have to impose ourselves').*

Base pattern 1: HUMAN ~ HUMAN

Pattern 2: [[Humano]] (se) come [[Objeto Físico]] *([[Human]] crashes against [[Physical Object]])*
Example: "Fidel se comió un escenario a raíz de un traspié" *('Fidel crashed against a stage as a result of a stumble')*

Base pattern 2: HUMAN ~ PHYSICAL OBJECT

In (14), both patterns, 1 and 2, have their origin in the pattern *[[Humano | Animal]] (se) come [[Comida]] ([[Human | Animal]] eats [[Food]]),* which is the most common, literal meaning. This pattern originates different figurative meanings of the verb, such as the ones shown in (14), which take the act of 'eating food' to refer to the act of 'clearly winning somebody in a competition' (pattern 1) (as 'devouring somebody'), or 'violently contacting something' (a humorous way of describing a crash when a person collides head-on with an object) (pattern 2). Hence, there is no metaphorical relation between base patterns 1 and 2, but between these two patterns and the one mapped to the literal meaning 'to eat'.
Finally, another aspect of the qualitative analysis of the data is the fact that the method cannot predict the direction of the metaphor. In all cases from (9) to (13) the source domain is the base pattern PHYSICAL OBJECT ~ HUMAN, and HUMAN ~ HUMAN is the target domain. However, observe the following examples of the pair HUMAN ~ HUMAN / HUMAN ~ PHYSICAL OBJECT:

(15) Verb *albergar* 'to give accommodation, to store'

Pattern 1: [[Humano]] alberga a [[Humano]] *([[Human]] gives accommodation to [[Human]])*
Example: "...venteros socarrones como el que alberga a Don Quijote y Sancho" *('...sardonic inn owners such as the one who gives accommodation to Don Quijote and Sancho').*

Base pattern 1: HUMAN ~ HUMAN

Pattern 2: [[Humano]] alberga [[Objeto Físico]] *([[Human]] stores [[Physical Object]])*
Example: "En su casa albergaba una colección de libros" *('At her/his home she/he stored a book collection').*

Base pattern 2: HUMAN ~ PHYSICAL OBJECT

(16) Verb *cortar* 'to cut, to interrupt'

Pattern 1: [[Humano]] cortar [[Objeto Físico]] *([[Human]] cuts [[Physical Object]])*
Example: "Seguí cortando la leña" *('I kept cutting the firewood').*

Base pattern 2: HUMAN ~ PHYSICAL OBJECT

Pattern 2: [[Humano]] corta a [[Humano]] *([[Human]] interrupts [[Human]])*
Example: "Me cortó con una sequedad que me dejó desorientado" *('She/he cut me with such a brusqueness that left me disoriented').*

Base pattern 2: HUMAN ~ HUMAN

We can observe that, in (15), HUMAN ~ HUMAN is the source domain for HUMAN ~ PHYSICAL OBJECT, as [[Physical Objects]] are categorized as such valuable things that are like persons that one hosts in a place. In contrast, in (16), the source domain is HUMAN ~ PHYSICAL OBJECT and the target domain is HUMAN ~ HUMAN, as conversation is categorized as something that can be "cut" or interrupted as a [[Physical Object]]. These findings, together with the ones exemplified in (14), lead us to believe that the method could be used to automatically detect these association of base patterns in corpus, but it would be necessary for human analysis to corroborate that the associations are really metaphors.

6 Conclusions and Future Work

This proposal is a preliminary attempt to find regularities in combinations of verb patterns with the same syntactic structure, using semantic types only. The purpose was to find out if these regularities have a metaphorical origin. The method is very simple and it can be used in all CPA projects or similar types of corpus annotation. Results seem promising, but they have to be tested with more data. It is also necessary to have a better theoretical articulation between types of metaphors and our corpus-driven findings, which could allow us to refine our results.
This study leads us to the following preliminary conclusions:

- Patterns of usage of different verbs share common features, particularly when focusing on semantic types. This common, general semantic information, that we called *base patterns,* can be easily extracted from verb patterns and traced in all verbs.
- A small number of base patterns form combinations which cover most of the verbs in the sample, which leads to consider that the proposed method could be appropriate to find new patterns of usage in corpus, linked to new meanings, and to organize the lexicographic information in the entry.

- Results also show that a small number of semantic types cover most of the base patterns, and a small number of base patterns connect to create pairs. This Zipfian tendency allows us to consider that further studies with this small group of semantic types and pairs could be cost-effective and cover many other cases.
- We do not have enough data to confirm that this method is appropriate to find metaphors in corpus, but the preliminary results show that in many cases we find a metaphorical origin in the connections between base patterns. Thus, at the moment, results show that base patterns linked via metaphors are common.
- While, certainly, most of the analysed pairs exhibit a metaphorical relation, the source domain of the metaphor is not predictable, because many semantic types can work as source or target domain. This is not new in the theory of metaphor, but the present analysis provides us with types of source domain / target domain combinations that may not be very frequent or prototypical. These data could bring new insights regarding the typology of metaphors.
- Similarly, combinations of base patterns provide us with information which could potentially enrich the usual formulation of conceptual metaphors. For example, a usual metaphor is “Events are Physical Objects”, such as in “Economic crisis hit the industry” (like a hammer), “Their relationship cracked” (like a wall), etc. Our results could add information to this formulation, such as we proposed in section 5, e.g. “An event cracking is an object physically cracking”.

There are different lines for future research. The most important next task would be to apply the same method to a more extensive group of verbs and see if results were consistent with the present ones. Another possible line of work could be to replicate the same procedure by converting semantically specific semantic types into more general labels, e.g. converting [[Illness]] into [[Eventuality]], [[Emotion]] into [[Abstract Entity]], etc. This can be done because semantic types are organized in an ontology and linked to each other via IS-A relations. This way it is possible to find more recurrent, general metaphors and this would allow us to establish a corpus-driven taxonomy of metaphors. This operation can be easily automatized since we already have a machine-readable version of the CPA Ontology. Finally, as already mentioned, the procedure can be replicated in other CPA projects such as the *Pattern Dictionary of English Verbs* (Hanks, online).

7 References

Apresjan, J. D. (1974). Regular Polysemy. In *Linguistics,* 12(142), pp. 5-32.

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Baisa, V., Može, S. & Renau, I. (2016). Multilingual CPA: Linking Verb Patterns across Languages. In T. Margalitadze, G. Meladze (eds.), *Proceedings of the XVII Euralex International Congress.* Tbilisi: Tbilisi University Press, pp. 410-417.

Firth, J. R. (1935[1957]). The Technique of Semantics. In J. R. Firth, *Papers in Linguistics.* Oxford: Oxford University Press, pp. 7-33.

Hanks, P. (2004). Corpus Pattern Analysis. In G. Williams, S. Vessier (eds.), *11th Euralex International Congress. Proceedings.* Lorient: Université de Bretagne-Sud, pp. 87-97.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations.* Cambridge, Ma: MIT Press.

Hanks, P. (dir.). (Online). *Pattern Dictionary of English Verbs.* Accessed at: http://www.pdev.org.uk

Hartmann, R.R.K. (2001). *Teaching and Researching Lexicography.* Harlow, UK: Pearson Education.

Jiang, G. & Chen, Q. (2017). A Micro Exploration into Learner’s Dictionaries: A Prototype Theoretical Perspective. In *International Journal of Lexicography,* 30(1), pp.108-139.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The sketch engine: ten years on. In *Lexicography*, 1(1), pp. 7-36.

Kosem, I. (2016). Interrogating a Corpus. *The Oxford Handbook of Lexicography.* Oxford: Oxford University Press, pp. 76-93.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By.* Chicago: The University of Chicago Press.

Malinowski, B. (1923). The Problem of Meaning in Primitive Languages. In C. K. Odgen I. A. Richards (eds.). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism.* Cambridge: Cambridge University Press, pp. 296-336.

Moldovan, A. (2019). Descriptions and Tests for Polysemy. In *Axiomathes*, *31(3),* pp. 1-21.

Renau, I. (2021). Algunos datos lexicográficos y de corpus para la representación de la polisemia regular en los diccionarios. *Boletín de Filología,* Anexo, pp. 905-925.

Renau, I., Nazar, R., Castro, A., López, B. & Obreque, J. (2019). Verbo y contexto de uso: un análisis basado en corpus con métodos cualitativos y cuantitativos. In *Revista Signos,* 52(101), pp. 878-901.

Renau, I. (dir.). (Online). *Verbario.* Accessed at: http://www.verbario.com

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, Ma: MIT Press.

Sinclair, J. (1998[2004). The Lexical Item. In J. Sinclair, *Trust the Text. Language, Corpus and Discourse.* Amsterdam: Routledge, pp. 131-148.

Acknowledgements

This study was funded by Fondecyt project nr. 1191204, Agencia Nacional de Investigación y Desarrollo, Chile government. Many thanks to the evaluators for their useful comments.

Les termes des arts dans les dictionnaires de la tradition française et dans les corpus de dernière génération : une relation d'inclusion réciproque ?

Zotti V.

University of Bologna Alma Mater Studiorum, Italy
valeria.zotti@unibo.it

Abstract

Dans cette contribution nous illustrons d'abord comment les termes des arts sont traités dans quelques dictionnaires de langue française de référence, pour ensuite vérifier dans quelle mesure trois corpus disponibles pour la langue française fournissent des informations complémentaires. Nous montrons ensuite, à travers notre exploration et une enquête menée auprès d'étudiants en lexicographie, que les données les plus intéressantes pour l'enrichissement des dictionnaires généraux existants proviennent de sous-corpus lexicographiques contenant des dictionnaires spécialisés sur l'art.

Keywords: lexicographie, corpus, terminologie, art, architecture, beaux-arts

1 Introduction

Depuis la révolution des corpus décrite par Rundell & Stock (1992), une multitude d'études soulignent les atouts de l'utilisation des corpus comme source d'information pour compléter les informations lacunaires données par les dictionnaires, qu'il s'agisse de dictionnaires de langue ou de dictionnaires bilingues (Bertels et al. 2009; Bertels & Verlinde 2011; Loock 2016; Granger 2018; entre autres). Les résultats obtenus de l'analyse de différents types de corpus montrent qu'ils contiennent, en général et toutes distinctions faites, des indications précieuses pour enrichir les descriptions lexicographiques traditionnelles. Sans vouloir nier cette constatation, confirmée par un grand nombre de recherches, nous voudrions, dans le cadre de cette contribution, attirer l'attention sur le fait que cet acquis général n'est pas toujours valable si on a affaire à certains domaines du lexique. Nous nous pencherons notamment sur celle qui est traditionnellement appelée "la langue des Arts", dont font partie les arts majeurs (l'architecture, la peinture et la sculpture), qui sont aussi nommés Beaux-Arts.

Après avoir exposé brièvement les caractéristiques de cette langue, qui se pose à mi-chemin entre la langue générale et la langue spécialisée, nous illustrerons, dans la première partie de notre contribution, comment les termes des arts sont traités dans quelques dictionnaires de langue française de référence : le *Dictionnaire de l'Académie française* (DAF), le *Trésor de la Langue française* (TLF), et le *Grand* et le *Petit Robert* (GR et PR). Dans la deuxième partie de ce travail, nous nous tournerons vers les corpus pour vérifier si et dans quelle mesure trois corpus disponibles pour la langue française (Frantext, French Web 2017 et LBC Français) fournissent en effet des informations nouvelles ou complémentaires par rapport à celles recensées dans les trois dictionnaires examinés, concernant la description de cette part de lexique qui se distingue par sa pluridisciplinarité considérable (Cetro & Zotti 2020). Nous avons testé dans ces corpus, aux caractéristiques bien différentes, un échantillon de termes des Arts qui seraient les plus intelligibles possibles pour un public de semi-experts et de non-experts.

En dernier ressort, c'est en nous fondant aussi sur une expérience menée auprès d'étudiants en lexicographie que nous montrerons ce que révèlent les résultats de l'exploration des corpus, tout en essayant de répondre à la question suivante : quelles données tirées des corpus sont les plus séduisantes pour l'enrichissement des dictionnaires généraux existants ? Nous soutiendrons qu'il existe entre dictionnaires et corpus une relation d'inclusion réciproque qui met en question certains acquis du débat en cours mentionné plus haut concernant l'impact de la linguistique de corpus sur la dictionnairique, et ce pour le domaine de la langue des arts notamment.

2 La langue des arts

Sous plusieurs points de vue, la nature de la langue des Arts n'est pas facilement saisissable (Casale & D'Achille 2004), car elle présente différentes facettes et se situe à mi-chemin entre sciences humaines et sciences exactes (Cetro & Zotti 2020: 83). Si, d'une part, elle est sans doute une langue spécialisée, c'est-à-dire une langue naturelle employée par les initiés "pour rendre compte techniquement de connaissances spécialisées" (Lerat 1995: 21), d'autre part, elle se projette sur la langue commune, en sorte que des mots employés par les artistes, comme *toile*, *pinceau*, *couleur*, *arcade*, *figure*, sont aussi des mots de la vie quotidienne. De plus, étant donné que les Arts ont été longtemps liés à la littérature et aux milieux cultivés, un bon nombre de ces mots passent naturellement dans la langue soutenue, et sont fréquemment employés de manière métaphorique dans des contextes non spécifiques (ex. *dresser un tableau*, *être la clef de voûte*).

Confronté à la question de la transmission des savoirs et notamment à la "description des arts", Diderot avait déjà illustré, dans l'*Encyclopédie* (ART, 1751), les difficultés et les problèmes qui se posaient alors à lui pour saisir cette langue qu'il trouvait imparfaite à cause de l'abondance des synonymes. Cela revient à dire que, dans ce domaine, les intersections entre langue commune et langue de spécialité sont telles et si nombreuses que l'inclusion des "termes des arts et des sciences" dans les dictionnaires de langue généraux a, depuis toujours, constitué une question épineuse pour les lexicographes. Faut-il inclure les noms des outils dont se servent les artistes, des matières qu'ils manient, des techniques qu'ils appliquent, etc. ? Les préfaces des dictionnaires de langue française les plus savants pullulent de réflexions sur les critères à adopter concernant l'intégration de vocabulaires spéciaux et techniques.

Nous avons établi un échantillon d'analyse, composé de n. 20 mots/termes se situant précisément dans ce continuum entre langue spécialisée et langue générale (Lerat 1995; Resche 2001), que nous avons analysés sur le plan quantitatif (présence ou absence de la nomenclature des dictionnaires et nombre d'attestations dans les corpus) et qualitatif (présence de citations littéraires ou techniques, nature et autorité des sources attestées). Nous avons écarté les mots très polysémiques qui auraient compliqué l'analyse.

- auvent
- arcade
- balustrade
- camaïeu
- clocher
- colonne
- coloris
- coupole
- dôme
- façade
- fresque
- gouache
- gravure
- loge
- marqueterie
- médaillon
- porche
- portail
- toile
- voussure

Par souci de brièveté dans le cadre de cet exposé, nos exemples ne porteront que sur quelques lexies, représentatives de domaines différents (beaux-arts et peinture, sculpture, architecture).

3 Les dictionnaires français de référence

En 1985, en comparant différents panoramas lexicographiques européens, Hausmann (1985) avait relevé une véritable "passion dictionnairique" en France. La France dispose en effet d'une particularité culturelle dans le domaine lexicographique et, comme l'a remarqué Pruvost (2000: 9), revendique une grande richesse sur le plan tant quantitatif que qualitatif :

> Le nombre de dictionnaires français publiés, plusieurs dizaines de milliers si l'on se réfère au simple intitulé dictionnaire utilisé comme générique pour toutes sortes d'ouvrages dès que l'ordre alphabétique les structure, et leur qualité enviée, surprennent effectivement de grandes nations étrangères.

La primauté de la France dans ce domaine encore aujourd'hui, même à l'époque des dictionnaires électroniques (De Schryver 2003), est le produit des progrès accomplis dans le travail lexicographique depuis des siècles, grâce à la coexistence d'une lexicographie institutionnelle et de nombres d'entreprises lexicographiques privées, les deux donnant lieu à des dictionnaires de la tradition très solides. Nous illustrerons dans les paragraphes suivants comment les termes des arts sont traités dans les trois dictionnaires français de référence.

3.1 Le Dictionnaire de l'Académie Française (DAF)

L'Académie Française, fondée en 1635 par le cardinal Richelieu, eut la tâche de répéter l'expérience italienne du *Vocabolario degli Accademici della Crusca* (1612) et, jusqu'en 1694, date de parution de la première édition de son dictionnaire, se prodigua, comme l'indique son Statut (1635), pour "travailler avec tout le soin et toute la diligence possible à donner des règles certaines à notre langue et à la rendre pure, éloquente et capable de traiter les arts et les sciences" (article XXIV). À cet effet, le même Statut établit qu'"il sera composé un dictionnaire, une grammaire, une rhétorique et une poétique" (article XXVI), et seront édictées pour l'orthographe des règles qui s'imposeront à tous (article XLIV). L'ambition de donner à la langue française les moyens de parvenir à rendre compte des sciences et des arts, qui constituent, avec les lettres, l'"une des plus glorieuses marques de la félicité d'un État" (Statut AF 1635), est donc centrale dès la fondation de cette institution.

Si, en principe, les termes des sciences et des arts sont initialement exclus de la 1^{e} édition du *Dictionnaire de l'Académie française* (DAF, Préface 1694), seulement "ceux qui sont extrêmement connus & d'un grand usage" seront inclus progressivement dans ses éditions successives (DAF, Préface, de la 2^{e} éd. 1718 à la 8^{e} éd. 1932). À titre d'exemple, parmi les 20 mots présents dans notre échantillon, 10 d'entre eux figurent dans la 1^{e} édition (*arcade*, *auvent*, *balustrade*, *clocher*, *coloris*, *dôme*, *façade*, *marqueterie*, *porche*, *portail*), les 10 autres n'apparaîtront que dans la 4^{e} (*camaïeu*, *coupole*, *fresque*, *gouache*, *gravure*, *voussure*) ou dans la 5^{e} (*colonne*), à l'exception de *médaillon* (mot assez polysémique, dont l'acception architecturale ne figurera que dans la 6e, alors que l'acception relative à la peinture n'apparaîtra que dans la 8e), de *toile* (qui figure dès la 1e édition dans le sens général de tissu, mais dont le sens de "toile peinte" ne figurera qu'à partir de la 4e), et de *loge* (dont l'acception architecturale figurera dans la 6e).

La 9e édition en ligne du DAF (en voie d'achèvement) mérite un développement distinct, parce qu'elle introduit des changements importants concernant et sa modalité de consultation, et le regard, sans doute renouvelé, porté sur la langue. Cette édition est accessible sur un nouveau portail numérique "innovant et sans équivalent, qui permettra la consultation dynamique de toutes les éditions de son Dictionnaire" (AF 2019) et qui ouvre, pour la première fois, des passerelles vers d'autres ressources externes, comme la *Base de Données Lexicographique Panfrancophone* (BDLP), qui appelle ainsi le DAF "à devenir une nouvelle référence en matière de dictionnaires dans l'espace numérique francophone" (AF 2019), et comme la base de données *FranceTerme*, recensant les mots scientifiques et techniques officiellement recommandés dans le cadre du dispositif d'enrichissement de la langue française, dont les termes des Arts. La consultation des entrées de certains mots de notre échantillon, notamment *façade*, *fresque*, *gravure* et *portail* donne accès, par le biais d'un lien hypertextuel, aux fiches terminologiques publiées au Journal Officiel de : *élévation*, *fresque vidéo*, *gravure* et *portail de messagerie*, respectivement dans les domaines (Habitat et construction > Architecture), (Arts > Audiovisuel), (Électronique > Composants électroniques) et (Informatique > Internet).

L'examen de notre échantillon dans cette 9e édition révèle aussi des changements importants concernant la description des sens spécialisées des mots à l'intérieur des articles du dictionnaire: des marques de domaine ou terminologiques ont été apposées pour cerner le champ d'application de chaque mot et/ou sens. À titre d'exemple, l'étiquette Architecture est antéposée à la définition des mots suivants (*balustrade*, *colonne*, *coupole*, *dôme*), l'étiquette Peinture à l'entrée *coloris* et l'étiquette Beaux-Arts entre dans ce dictionnaire et précède la définition de *fresque*.

Cette double ouverture, l'accès à la ressource externe *FranceTerme* et l'introduction d'étiquettes terminologiques dans la microstructure, confirment que, même dans un dictionnaire normatif de la grande tradition française, "il n'y a plus de distinction nette entre terminographie et lexicographie, ces deux disciplines s'étant rapprochées jusqu'à faire converger leur méthodologie et leurs procédures de travail" (Cabré 2018: 38). Sur le plan de l'exemplification, cette édition reste fidèle à la vocation de l'Académie d'être la gardienne du bon usage, ce pourquoi elle n'enregistre aucune nouveauté concernant les exemples qui sont encore forgés, en sorte que, pour la compréhension de l'emploi en contexte des termes du domaine des Beaux-Arts en l'occurrence, ce dictionnaire n'est pas toujours satisfaisant.

3.2 Le Trésor de la Langue Française (TLF)

Le deuxième dictionnaire de référence qui fait l'objet de notre analyse est un dictionnaire qui a été pionnier en matière de corpus : le *Trésor de la Langue Française* (TLF), dictionnaire extensif (plus de 100 000 entrées) qui, de par sa nature, intègre un nombre élevé de lexiques des langues spécialisées. Le lexique des Beaux-Arts y est largement représenté et illustré par des citations d'auteurs dont l'autorité est reconnue dans ce domaine. Nombreuses sont par exemple les entrées dans lesquelles sont attestées des citations tirées d'ouvrages spécialisés de célèbres historiens de l'art et architectes des XVIIe et XVIIIe siècles, tels que André Félibien (cité dans 398 entrées), Augustin Charles-D'Aviler (44 entrées), Eugène Viollet-Le Duc (491 entrées) et Jules Adeline (136 entrées), ainsi que de nombreux articles de l'*Encyclopédie* de Diderot et D'Alembert (1751-1772). Aussi, il est intéressant de remarquer la présence de 794 entrées qui contiennent des informations tirées du *Journal* du peintre Eugène Delacroix.

Ce qui nous importe dans le cadre de cette démonstration, c'est de montrer la portée et la pertinence de ces citations pour la description de la langue des Beaux-Arts. Le tableau 1 rapporte, à côté de quelques entrées de notre échantillon, des citations, tirées de textes techniques (traités et manuels d'histoire de l'art et de critique de l'art) et de dictionnaires spécialisés, qui illustrent l'emploi de ces termes avec une ouverture en diachronie sur leur histoire culturelle. Les citations littéraires, qui sont pour la plus grande partie des attestations présentes dans ce dictionnaire, n'ont pas été prises en compte.

Entrée	*Citation*	*Source*
auvent	Lorsqu'on renonça aux chaises à porteurs pour ne plus se servir que des carrosses, ceux-ci ne pouvant pénétrer dans les vestibules, il fallut modifier le programme des entrées d'honneur ; établir des *auvents* formant saillie en dehors de ces vestibules, afin de préserver les arrivants de la pluie et des bourrasques ; ce qui fut fait. On donna à ces *auvents* le nom de marquises.	VIOLLET-LE-DUC, *Entretiens sur l'archit.*, t. 2, 1872, p. 260.
balustrade	Le sanctuaire se distinguait du chœur, (...) par une *balustrade* ou chancel particulier, placé antérieurement au maître-autel et formant la table de communion.	A. LENOIR, *Archit. monastique*, t. 2, 1856, pp. 253-254.
camaïeu	Plus simple et plus prenant peut-être [que les autres portraits d'Ingres] est le portrait en *camaïeu* de sa première femme.	L. RÉAU, *L'Art romantique*, 1930, p. 80.
coloris	Vous savez que chaque artiste a son style (...). Si c'est un peintre, il a son *coloris*, riche ou terne, ses types préférés, nobles ou vulgaires, ses attitudes, sa façon de composer	TAINE, *Philos. de l'art*, t. 1, 1865, p. 2.
dôme	La Descente de Croix peinte par Baroche pour le *Dôme* de Pérouse, fut une des premières imitations de Volterra en Italie.	MÂLE, *L'Art relig. après le Concile de Trente*, 1932, p. 280.

gravure	La *gravure* est un art qui s'en va, mais sa décadence n'est pas due seulement aux procédés mécaniques avec lesquels on la supplée, ni à la photographie ni à la lithographie...	DELACROIX, *Journal*, 1857, p. 30.
porche	À Albi, au flanc de la forteresse de brique, les derniers gothiques ont attaché un porche léger fouillé comme une dentelle, comme une pièce d'orfèvrerie.	HOURTICQ, Hist. art, Fr., 1914, p.105.
toile	Alain Fournier préférait Laprade ; il insistait peu sur les règles de la composition et ne tenait à distinguer, dans les toiles de ce peintre délicat, que la fragilité, la translucidité, le mystère des êtres qui peuplaient ses paysages aux verts exténués.	LHOTE, Peint. d'abord, 1942, p. 30.

Tableau 1 : Attestations d'ouvrages spécialisés dans l'art dans le TLFi.

Ce qui frappe, c'est aussi la présence parmi les sources documentaires de dictionnaires spécialisés du domaine des Beaux-Arts ou de l'Architecture qui sont employés pour définir le sens de certaines acceptions spécialisées. C'est le cas de l'entrée *gravure* dans laquelle l'acception qui relève du domaine de la peinture est illustrée par une définition reprise du célèbre ouvrage de référence *Lexique des termes d'art* de Jules Adeline (1885):

GRAVURE, subst. fém.
A. Vx. Sillon. (Dict. XIXe et XXe s.).
ARCHIT. "Ornements indiqués par des tailles en creux dont on fait grand usage dans le style néo-grec pour agencer des rinceaux autour de fleurons en relief et dont la tradition remonterait à l'architecture égyptienne" (ADELINE, *Lex. termes art*, 1884). […]

Aussi, des dictionnaires qui font autorité dans le domaine des Beaux-Arts, comme le *Dictionnaire d'architecture* d'André Félibien (1676-1690), figurent dans le TLF même dans la section consacrée à l'étymologie et histoire, comme pour le cas de l'entrée *balustrade*:

BALUSTRADE, subst. fém. […]
ÉTYMOL. ET HIST. [Mil. XVIe s. selon Pt ROB.] 1653 *balaustrata* (Oud. d'apr. DG); 1654 (LORET, 26 sept., 98 dans BRUNOT t. 3, p. 220 : Adieu la chambre à balustrade) [*balustre*, au sens de « balustrade » 1676 (FÉLIBIEN, *Dict. d'archit*. dans GAY : *balustre* signifie aussi la balustrade qui environne le lit des rois et des princes)]. […]

Ces différents types d'informations sont issus d'ouvrages qui ont été numérisés à partir des années 1960 et qui peuvent être consultés dans la base de données Frantext, un corpus de référence pour la langue française à base littéraire sur lequel nous reviendrons plus loin.

3.3 Le Grand et le Petit Robert (GR et PR)

Nous nous focalisons à présent sur l'apport offert par les dictionnaires alphabétiques et analogiques de la langue française le *Grand Robert* (GR) et le *Petit Robert* (PR) (éd. 2020) pour la description du lexique artistique. Les dictionnaires de langue Le Robert ont bénéficié au cours des siècles de la contribution de leurs illustres prédécesseurs, en particulier du *Dictionnaire de la langue française* d'Émile Littré (1863-1872) pour sa portée historique et son exhaustivité. Dans une entrée du PR sont en effet condensés les progrès accomplis par la recherche dans les sciences du langage (phonétique, étymologie, sémantique, philologie, stylistique) au cours des siècles jusqu'aux années '70 (Zotti 2008: 70). Comme l'a remarqué Pruvost (2006: 69):

> Au-delà d'une présentation en arborescence de chaque article, avec des indications étymologiques qui intègrent la datation des sens, les deux caractéristiques majeures restent cependant, d'une part, le riche corpus de citations sur lequel s'appuie le dictionnaire, avec de nombreux auteurs contemporains, et, d'autre part, la mise en réseau sémantique des mots selon le programme annoncé en sous-titre, les mots et les associations d'idées.

Ces différents types d'informations sont issus d'ouvrages qui ont été numérisés à partir des années 1960 et qui peuvent être consultés dans la base de données Frantext, un corpus de référence pour la langue française à base littéraire sur lequel nous reviendrons plus loin.

Nous nous arrêtons donc sur ces deux aspects, le riche corpus de citations d'une part, et d'autre part la mise en réseau sémantique des mots, autrement dit la présence de renvois analogiques à l'intérieur des entrées qui jettent un éclairage sur la « circulation des sens des mots » (Rey-Debove et Rey 1993 : XVII, Préface) à l'intérieur du lexique. Grâce à l'introduction de la dimension onomasiologique dans un dictionnaire alphabétique, reprise du grand *Dictionnaire alphabétique et analogique* de Paul Robert en 7 volumes (1978), le PR permet en effet de découvrir, parmi les renvois analogiques, une large gamme de synonymes partiels comprenant différents degrés de spécialisation, ce qui peut s'avérer très utile lorsqu'on s'aventure dans un domaine complexe et hybride, tel que celui des Beaux-Arts, où les croisements entre différentes disciplines, entre l'ingénierie et l'architecture par exemple, sont fréquents, comme nous l'avons évoqué au début de cette étude.

Nous donnerons quelques exemples significatifs, relatifs à l'architecture, afin d'éclairer ces acquis. Les exemples *arcade* et *balustrade*, dont nous rapportons ci-dessous des extraits tirés du PR (2020), nous semblent particulièrement révélateurs:

arcade nom féminin […]
1 ARCHIT. Ouverture en arc ; ensemble formé d'un arc et de ses montants ou points d'appui (souvent au plur.). Les arcades d'un aqueduc, d'un cloître, d'une galerie (→ **arcature**). Les arcades de la rue de Rivoli, du Palais-Royal. Arcade aveugle, feinte, simulée.

Arcade profonde. → 2. **arche**. Arcades en plein cintre, en ogive. → **1. arc**, **archivolte**.
balustrade nom féminin […]
2 Clôture à hauteur d'appui et à jour. → **Garde-corps**. La balustrade d'une terrasse, d'une galerie, d'un balcon, d'une passerelle (→ **Rambarde**), d'un escalier (→ **Rampe**), d'un pont (→ **Garde-fou**, **parapet**). Une petite balustrade. → **Balustre**. Entourer d'une balustrade. → **Balustrer** (vx). Être accoudé à la balustrade. Enjamber la balustrade.

Le système sémiotique complexe mis en place par l'introduction de la démarche onomasiologique dans un dictionnaire de langue à base sémasiologique prouve qu'il est possible de saisir et reproduire dans un dictionnaire le jeu d'échos, la série infinie de réactions en chaîne que la langue provoque. Ici l'entrée *arcade* renvoie à *arcature*, *arche*, *arc* et *archivolte*, des mots analogues, que l'on trouvera dans des textes spécialisés de ce domaine pour désigner de manière très précise les différents types d'*arcades*. Aussi, l'entrée *balustrade* renvoie à *garde-corps*, un synonyme de l'entrée, à *rambarde*, *rampe*, *garde-fou* et *parapet*, c'est-à-dire des mots hyponymes gravitant autour d'une notion donnée, donc appartenant au même champ associatif. Mis à part l'utilité reconnue de ce système pour la production, cela est particulièrement utile lorsque on a affaire à une langue spécialisée, parce qu'il éclaire l'utilisateur sur la différence de sens entre des mots presque synonymes (ex. entre une *balustrade* et une *rambarde*). Comme l'a affirmé Heinz (1993:111), en effet "c'est grâce aux renvois […] que la lexicographie réussit à résoudre bon nombre des problèmes concernant la représentation des relations inter-lexicales". Aussi, cette manière de concevoir les choses permet de trouver dans le PR des articles extrêmement riches et donne une couleur encyclopédique à ce dictionnaire essentiellement linguistique (Veyrat 1995 : 191).

Concernant le corpus textuel à la base des dictionnaires de langue Le Robert, nous avons constaté que le corpus d'attestations est essentiellement littéraire, plus à jour bien évidemment par rapport à celui du TLF, étant donné qu'ici même la littérature contemporaine est prise en compte. Cependant, nous avons remarqué que peu de sources techniques sont utilisées pour attester l'emploi de la langue des Beaux-Arts dans le GR qui, en tant que dictionnaire extensif, devrait contenir un apparat d'exemples beaucoup plus riche que le PR. Lorsque quelques attestations tirées de textes spécialisées y figurent (cf. tableau 2), il s'agit plutôt de développements encyclopédiques, comme dans le cas de l'entrée *gravure.*

Entrée	*Citation GR*	*Source*
colonne	2. Elles *(les âmes du moyen âge)* aspirent au gigantesque (…) amoncellent les colonnes en piliers monstrueux (…)	TAINE, *Philosophie de l'art*, I, II, VI, 4.
fresque	On appelle *peindre à fresque*, l'opération par laquelle on emploie des couleurs détrempées avec de l'eau, sur un enduit assez frais pour en être pénétré. En italien on exprime cette façon de peindre par ces mots *dipingere a fresco*, peindre à frais. C'est de là que s'est formée une dénomination qui, dans l'orthographe française, semble avoir moins de rapport avec l'opération, qu'avec le mot italien dont elle est empruntée.	WATELET, in *Encyclopédie* (DIDEROT), art. *Fresque* (1751).
gouache	Le charme particulier de l'aquarelle, auprès de laquelle toute peinture à l'huile paraît toujours rousse et pisseuse, tient à cette transparence continuelle du papier ; la preuve c'est qu'elle perd de cette qualité quand on gouache quelque peu ; elle la perd entièrement dans une *gouache.*	E. DELACROIX, *Journal*, 6 oct. 1847.
gravure	*Gravure*. La gravure est un art qui s'en va, mais sa décadence n'est pas due seulement aux procédés mécaniques avec lesquels on la supplée, ni à la photographie, ni à la lithographie, genre qui est loin de la suppléer, mais plus facile et plus économique (…) La gravure est une véritable traduction, c'est-à-dire l'art de transporter une idée d'un art dans un autre (…) La langue étrangère du graveur (…) ne consiste pas seulement à imiter par le moyen de son art les effets de la peinture, qui est comme une autre langue. Il a, si l'on peut parler ainsi, sa langue à lui qui marque d'un cachet particulier ses ouvrages (…)	E. DELACROIX, *Journal*, 25 janv. 1857.
porche	Nous avons eu à signaler l'importance du *porche*, né de l'église-porche carolingienne, dans beaucoup d'édifices du XIe siècle. Tantôt ils se rattachent (…) au type du clocher de façade. Tantôt ils composent des espèces de portiques, soit ouverts, comme à Saint-Benoît-sur-Loire, et portés par des colonnes, soit bâtis sur de fortes masses murales percées de baies, comme à Ébreuil. Tantôt ils ont les dimensions d'églises annexes, précédant la nef (…) les Bourguignons restèrent fidèles à ce parti. Le narthex de Vézelay, le porche gothique de Cluny (…) montrent leur constance à cet égard.	Henri FOCILLON, *l'Art d'Occident*, I, II, 2, p. 70.

Tableau 2 : Attestations d'ouvrages spécialisés dans l'art dans le GR

Le corpus de citations savantes qui a été rassemblé pour les dictionnaires Le Robert nous paraît donc moins exhaustif pour ce qui relève de la description du domaine des Arts.

4 Les corpus textuels

Dans cette partie, nous nous tournons vers les corpus pour vérifier si et dans quelle mesure trois corpus disponibles pour la langue française fournissent en effet des informations nouvelles ou supplémentaires par rapport à celles recensées dans les trois dictionnaires examinés, concernant la description de "la langue des Arts". Nous avons sélectionné :

– un corpus textuel de référence, le corpus *Frantext*, composé pour la plupart de textes littéraires et philosophiques, mais aussi scientifiques et techniques (environ 10%), qui vont de 1180 à 2013, développé au sein de l'ATILF-CNRS dans les années '70 afin de fournir des exemples pour le *Trésor de la Langue Française*, et qui, une fois le dictionnaire terminé, continue à évoluer ;

- le corpus *French Web 2017* (frTenTen17), un corpus très vaste (5,7 milliards de mots) constitué de textes collectés automatiquement sur la Toile et qui décrit aussi différentes variétés diatopiques de la langue française (européennes, canadiennes et africaines) ;
- le corpus *LBC Français* (*Lessico Beni Culturali*), un corpus monolingue comparable ouvert, qui continue d'être alimenté et qui a atteint à ce jour la taille de 3,5 millions de mots, rassemblant des textes en français sur le patrimoine artistique italien qui sont représentatifs de différents niveaux de technicité.

Nous avons testé dans ces corpus, aux caractéristiques bien différentes, notre échantillon de termes des Arts. Nous n'en rapportons ici qu'un cas de figure représentatif : l'exploration du mot *portail*.

4.1 Frantext

La recherche du mot *portail* dans l'intégralité du corpus Frantext donne 2 259 résultats. La plupart des occurrences font référence à *portail* en tant que "grille" ou en tant qu'entrée d'une "maison / édifice non religieux". Les extraits qui contiennent le pivot, pour la grande majorité littéraires, sont surtout tirés de romans dans lesquels on mentionne le portail, souvent en fer forgé, d'un jardin / parc / maison qui s'ouvre et se referme. Le portail en tant qu'"entrée monumentale avec une porte d'un édifice religieux / église", qui est l'acception architecturale selon le DAF (9^e éd.), est très peu attesté. Une exploration plus poussée de ce corpus donne finalement accès à des occurrences relatives à cette acception ("portail à triple rang de fenêtres gothiques", "portail à colonnes", "portail à arceaux surbaissés", "portail à deux travées", "portail en ogive", le plus fréquent) qui constituent des informations complémentaires à celles attestés dans les trois dictionnaires consultés dans la première partie de notre analyse. Dans Frantext, corpus à dominante littéraire, on ne trouve bien évidemment aucune occurrence de *portail* en tant que site d'accès internet (domaine de l'informatique et des télécommunications) qui est en revanche très attesté dans le corpus frTenTen17.

4.2 frTenTen17

Le corpus frTenTen17, un corpus issu de la Toile et pour cela très à jour, contient beaucoup d'attestations de *portail* en tant que "grille automatisée / électrique / télécommandée / coulissante" et, dans le domaine de l'informatique, en tant que "portail multimédia / web / dimensionnel". Le *word-sketch* lancé sur Sketch-Engine montre de manière évidente que la plupart des attestations de *portail* se révèlent peu pertinentes aux fins de notre analyse. Presque aucune des occurrences relevant du domaine artistique ("entrée monumentale qui comporte une porte d'une église") n'y figurent. Une exploration plus poussée, et coûteuse en termes de temps, de ce corpus permet de repérer quelques occurrences liées au domaine de l'architecture et du bâtiment qui sont cependant très limitées en nombre ("portail nord", "grand portail", "tympan du portail", etc.). Elles n'enrichissent en rien notre recherche, car il s'agit de collocations qui sont déjà présentes dans Frantext et attestées dans les trois dictionnaires de langue consultés.

4.3 LBC Français

Le corpus LBC Français, en tant que corpus spécialisé de la langue française dans le domaine des arts, permet évidemment de repérer seulement des occurrences pertinentes qui concernent le domaine des Beaux-Arts, ce qui rend sa consultation plus rapide et plus aisée au regard des objectifs de ce travail. Ces attestations sont toutes tirées de textes en français sur le patrimoine artistique italien qui sont représentatives de quatre différentes typologies textuelles (textes littéraires, de vulgarisation, techniques et dictionnaires spécialisés) et de divers niveaux de spécialisation (Cetro & Zotti 2020). En termes quantitatifs, même si les résultats affichés pour le mot *portail* sont beaucoup plus restreints (231 occurrences) par rapport à Frantext et surtout à frTenTen17, elles sont toutes acceptables et satisfaisantes. Ainsi, dans ce corpus spécialisé, on trouve, outre les collocations qu'on avait déjà trouvées dans les autres ressources, bien d'autres collocations et syntagmes spécialisés concernant le mot *portail* ("portail de la nef", "portail de la façade", etc.) et les différentes parties constitutives d'un portail dans son acception architecturale ("ébrasements du portail", "voussures du portail", "pignon du portail", "archivoltes du portail", etc.).

Ces collocations proviennent d'une source qui fait autorité parmi les spécialistes, présente en version intégrale dans le corpus LBC, à savoir le *Dictionnaire raisonné de l'architecture française du XIe au XVIe siècle* de l'architecte français Eugène Viollet-Le-Duc (1875). Sur les 231 occurrences de *portail* attestées dans le corpus LBC Français, 139 sont tirées de ce dictionnaire. Cette donnée permet d'entrevoir une caractéristique, qui est aussi une limite à l'état actuel, de ce corpus, à savoir le fait qu'il n'est pas équilibré. Les articles du dictionnaire spécialisé de Viollet-le-Duc comptent plus de 20% du corpus, ce qui rend cet auteur et cette source surreprésentés (Farina et Sini 2020 : 9).

En dépit de cet aspect problématique, qui fera l'objet d'une intervention pour obtenir un corpus plus équilibré et qui concernera tous les sous-corpus des différentes typologies textuelles prises en compte, le corpus comparable LBC Français contient des échantillons des variétés qui caractérisent la langue de l'art, et couvre la période qui va de la Renaissance à l'époque contemporaine. On trouve donc dans ce corpus la même hétérogénéité discursive et lexicale qui caractérise la langue de l'art et, pour cette raison, il est possible de le considérer comme un corpus représentatif du lexique de l'art et du patrimoine artistique italien en français. En dépit de sa petite taille, si comparé aux grands corpus de référence, la variété textuelle, lexicale, chronologique et de registres de ce corpus permet d'avoir une photographie fiable des deux phénomènes que le projet LBC veut étudier à travers le corpus : la langue de l'art pour décrire le patrimoine florentin et la langue spécialisée des domaines de la peinture, de l'architecture et de la sculpture. A ce propos, nous partageons ce qui a été observé par Loock (2016) en citant O'Keffe (2007):

> […] en matière de corpus, ce n'est pas la taille qui compte mais la qualité de ce que l'on trouve qui peut permettre à un petit échantillon d'être suffisamment représentatif […] la taille d'un corpus dépend des informations que l'on souhaite y chercher ; ainsi pour un registre spécialisé, un petit corpus suffit.

5 Remarques sur la complémentarité entre dictionnaires et corpus

L'examen des dictionnaires pris en compte, le TLF, le DAF et le PR et GR, montre que les collocations et les syntagmes de *portail* qui y sont attestés sont bien évidemment les plus connus et les plus fréquents. En fait, ces dictionnaires de langue enregistrent les informations les plus saillantes pour comprendre le sens, l'emploi et le domaine d'application d'un mot donné, de *portail* en l'occurrence. Nous reconnaissons que les lexicographes ont sélectionné les données les plus pertinentes pour décrire ce mot de manière à en présenter dans leurs entrées une synthèse structurée et, toutes distinctions faites entre les quatre dictionnaires en question, de consultation aisée. En fait, les informations les plus significatives intégrées dans les dictionnaires consultés correspondent en général à celles plus fréquentes qui ont été extraites des trois corpus examinés (ex. *portail* : *d'une cathédrale*, *de l'église*, *royal*, *roman*, *gothique*, *nord*, *sud*, *méridional*, *central*, etc.). D'autre part, les corpus ont permis d'intégrer plusieurs attestations complémentaires qui relèvent surtout de la terminologie propre au domaine de l'architecture (outre les collocations mentionnées plus haut tirées du dictionnaire de Viollet-Le-Duc dans le corpus LBC Français, nous mentionnons aussi *contreforts du portail*, *soubassement du portail*, *trumeaux du portail*, *bossage du portail*, *portail* : *à arceaux surbaissés* / *nu en ogive* / *sévillan* / *brodé*, repris des trois autres corpus confondus).

Nous en concluons que l'exploration des trois corpus pour compléter les descriptions offertes par les dictionnaires s'est avérée utile, bien que les caractéristiques intrinsèques de ces corpus, qui en sont les limites constitutives, aient rendu le repérage des informations relatives à la langue de l'art plus ardue et moins efficace sur le plan pratique, notamment : la surreprésentation du genre littéraire dans le corpus Frantext ; la taille démesurée (plusieurs milliards de mots) du corpus frTenTen17 qui a fait en sorte qu'on se heurte au problème de la polysémie du mot recherché et que l'on ait dû trier énormément d'informations non pertinentes ; et, pour finir, le manque d'équilibre du corpus LBC Français qui, tout en fournissant des données très congrues pour la description du lexique artistique, n'est pas fiable en termes de fréquence et de représentativité des attestations pour les différents genres textuels qui y sont intégrés.

Nous en arrivons ainsi au cœur de notre démonstration et, pour ce faire, il nous faudra faire appel encore à quelques données obtenues de l'analyse de notre échantillon. L'exploration d'un autre mot, *clocher*, dans le corpus LBC Français a confirmé que les informations les plus pertinentes pour compléter les descriptions des dictionnaires existants dérivent du *Dictionnaire* de Viollet-Le-Duc (cf. tableau 3).

Patron syntaxique	Collocation en lien avec l'art	Exemple d'occurrence	Source	T-score	*n* de cooccurrences
[N+A]	*clocher* central	Avant la construction du *clocher central* de Vernouillet [...]	Viollet Le Duc, *Dictionnaire raisonné de l'architecture française du XIe au XVIe siècle*, charnier-console, 1854.	7,269	53
[N+A]	*clocher* normand	[...] des grands *clochers normands* élevés [...]		3,153	10
[V+N]	*clocher* s'élève	[...] *s'élève un grand clocher* sur une base épaisse [...]		5,672	33
[N+N]	couronnement du *clocher*	(…) son étage supérieur octogonal sous la flèche nous rappelle les *couronnements des clochers* de Brantôme [...]		2,812	8
[A+N]	grand *clocher*	Nous ne possédons pas un seul *grand clocher* complet [...]		2,872	11

Tableau 3. Synthèse des occurrences de *clocher* dans le corpus LBC Français.

La recherche d'un autre mot de notre échantillon, *fresque*, dans le même corpus restitue beaucoup de résultats, à savoir sept pages de concordances pour un total de 1 084 occurrences de ce mot pour les seuls textes en français langue originale, dont des collocations adjectivales (épithètes : *visionnaire*, *admirable*, *charmante*, *magnifique*, etc.) et verbales (*décorer*, *travailler*, *faire à -* ; aussi avec PP : *peinte*, *exécutée*, *tirée*, etc.) qui étaient absentes des dictionnaires consultés. Les données les plus séduisantes pour l'enrichissement des dictionnaires généraux existants qui montrent une fréquence élevée d'emploi dérivent en particulier de textes écrits par des spécialistes (bien qu'ils soient indiqués dans le corpus comme littéraires, ce qui sera revu dans la prochaine mise à jour du corpus), ce que nous avions déjà remarqué en analysant le corpus Frantext, à savoir :

– des historiens de l'art, des critiques d'art, ainsi que des artistes (cf. références bibliographiques) : Robert Moran (1994), Léon Palustre (1892), Michel Feuillet (2009), Élie Faure (1924), Robert de la Sizéranne (1910), Émile Michel (1901), Georges Lafenestre (1882), etc.
– ainsi que des dictionnaires spécialisés, encore une fois le *Dictionnaire d'architecture* de Viollet-Le-Duc (1854-1868) et de l'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* de Diderot et d'Alembert, précisément l'article « École florentine » de Chevalier Jaucour, qui présente 4 occurrences de

fresque.

L'intérêt de la présence de dictionnaires d'autorité dans le domaine des Beaux-Arts est confirmé encore ailleurs au cours de l'analyse de notre échantillon. Une autre source lexicographique qui fait autorité dans ce domaine, *Des principes de l'architecture, de la sculpture, de la peinture, et des autres arts qui en dépendent. Avec un Dictionnaire des termes propres à chacun de ces arts*, d'André Félibien (1676), qui est présent partiellement à ce jour à l'intérieur du corpus LBC, fournit un grand nombre d'attestations précieuses du mot *colonne* (40 occurrences) ainsi que de deux autres mots présents dans notre échantillon (*dôme* et *toile*).

En définitive, les résultats de notre exploration relèvent que les données les plus intéressantes pour l'enrichissement des dictionnaires généraux existants proviennent de sous-corpus lexicographiques contenant des dictionnaires spécialisés sur l'art à l'intérieur des corpus examinés. Cela nous amène à soutenir qu'il faudrait parler d'une relation d'inclusion réciproque et circulaire entre dictionnaires et corpus, où les premiers nourrissent les seconds pour que les premiers soient de plus en plus performants.

6 Conclusions

Une enquête menée auprès de deux groupes d'étudiants en lexicographie, issus de deux formations distinctes en France et en Italie, respectivement le Master LTTAC (Lexicographie, Terminographie, Traitement Automatique des Corpus) de l'Université de Lille et le Master international LSC (Language, Society & Communication) de l'Université de Bologne nous a permis de confirmer nos constats. Les étudiants, qui ont testé d'autres mots du lexique d'art dans les mêmes ressources lexicographiques et textuelles présentées dans cette contribution, ont observé que la consultation des corpus s'est avérée utile jusqu'à un certain point, notamment pour vérifier la fréquence d'emploi et surtout pour détecter les collocations les plus récurrentes, voire plus pertinentes. Cependant, ils n'ont pas considéré l'exploration des corpus comme indispensable au regard des informations déjà repérées dans les dictionnaires de langue générale, ce pourquoi ces corpus ne sont pas toujours ciblés pour la tâche accomplie. En outre, ils ont relevé que les dictionnaires contiennent certes moins de contextualisations que les corpus, mais que celles-ci ont été triées par des experts, les lexicographes, qui possèdent des compétences pour réaliser une analyse linguistique fine, alors que les données de corpus sous forme brute s'avèrent indigestes et difficiles à interpréter, étant donné que les collocations les plus intéressantes sont réservées à des textes et à des domaines extrêmement précis et spécialisés. Étudier un corpus très conséquent de façon minutieuse peut sans conteste fournir des résultats plus complets, dans certains domaines. C'est le cas du corpus LBC qui nous a amené à plaider pour une relation d'inclusion réciproque entre les dictionnaires spécialisés dans le domaine des arts et les dictionnaires de la grande tradition lexicographique française.

Nous conclurons par une réflexion issue de cette enquête mais aussi de plusieurs années de recherche en lexicographie et en linguistique de corpus, ainsi que d'une passion pour les dictionnaires qui nous amène à lutter pour la reconnaissance de la valeur de ces derniers. Comme nous l'avons évoqué au début de cette contribution, une multitude d'études se penchent aujourd'hui sur les atouts des corpus pour compléter les informations lacunaires données par les dictionnaires. Sans vouloir nier leur grande utilité, pour ce qui concerne en particulier les possibilités offertes par les corpus d'accéder à un grand nombre d'exemples d'une unité lexicale dans différents contextes, ici nous entendons nous prononcer en faveur des dictionnaires, pour défendre leur valeur ajoutée à un moment de l'histoire où leur survie est sérieusement mise en danger. Nous soutenons en fait que, depuis la naissance de la lexicographie, la tâche du lexicographe, un spécialiste de la langue, est de condenser et d'agencer dans une entrée de dictionnaire toutes les informations linguistiques nécessaires pour guider l'utilisateur dans la compréhension d'une unité lexicale sous toutes ses facettes. Aujourd'hui cette analyse fine préalable de données linguistiques est léguée à l'utilisateur du concordancier qui n'est pas toujours préparé ni formé convenablement pour accomplir cette tâche complexe. Autrement dit, en demandant à l'utilisateur d'analyser toutes les concordances, au lieu d'accélérer et de faciliter sa tâche, on lui donne du travail supplémentaire qui était auparavant effectué par le lexicographe professionnel. Si l'on pense aux progrès accomplis dans l'histoire de la lexicographie française (du *Littré* au *Petit Robert* en passant par Hatzfeld et Darmester par ex.), peut-on soutenir qu'une véritable (r)évolution en lexicographie s'est affirmée à la suite de l'avènement de la linguistique informatique ? Ou bien s'agit-il plutôt d'une involution ? Nous plaidons pour l'affirmation d'une relation d'inclusion réciproque qui est particulièrement patente dans certains domaines du lexique, comme dans la langue des Beaux-Arts. En dernier ressort, nous soutenons ici que l'analyse fine donnée par des dictionnaires, de bons dictionnaires bien entendu, ne pourra jamais rivaliser avec des corpus et que ces derniers ne pourront pas mettre en question la survie de ces mêmes dictionnaires.

7 Références bibliographiques

Académie Française (2019). *L'Académie française met son Dictionnaire à la disposition du public grâce à un portail numérique en accès libre et gratuit*, Communiqué de presse, Paris, le 7 février 2019. https://www.dictionnaire-academie.fr/lancement [18/04/2021].

Bertels, A., Fairon, C., Tiedemann, J. & Verlinde S. (2009). Corpus parallèles et corpus ciblés au secours du dictionnaire de traduction. In *Cahiers de lexicologie*, 94(1), pp. 199-219.

Bertels, A., Verlinde, S. (2011). La lexicographie et l'analyse de corpus : nouvelles perspectives. In *Meta*, 56, pp. 247-265.

Cabré, T., (2018). *Terminology: Theory, methods and applications.* John Benjamins Publishing.

Casale, V., D'Achille, P. (2004). *Storia della lingua e storia dell'arte. Atti del III Convegno ASLI*, Firenze: Cesati.

Cetro, R., Zotti, V. (2020). Les corpus et la base terminologique LBC. Des ressources pour la traduction du patrimoine artistique. In Mangeot, M., Tutin, A. (eds.) (2020). Lexique(s) et genre(s) textuel(s) : approches sur corpus. Actes de la conférence 11e Journées du réseau Lexicologie, Terminologie, Traduction. Paris: Editions des archives contemporaines, pp. 81-98.

De La Sizeranne, R. (1910). Les Masques et les Visages-Portraits de Florentine, le long de la Seine et de l'Arno I. XVe siecle. In *Revue des Deux Mondes*, 60, pp. 160-190.

De Schryver, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. In *International Journal of Lexicography*, 16(2), pp.143-199.

Farina, A., Sini, L. (2020). Il corpus LBC Francese. In Billero R., Farina A. et Nicolás Martínez M. C. (eds.), I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali. Firenze University Press, pp. 77-99.

Faure, E. (1924). *Histoire de l'art III: l'art renaissant,* Paris: G. Cres et Cie.

Feuillet, M. (2009). *L'art italien*, Paris: PUF.

Granger, S. (2018). Has lexicography reaped the full benefit of the (learner) corpus revolution?. In Čibej, J., Gorjanc, V., Kosem; I., Krek, S. (eds.), *Proceedings of the XVIII EURALEX International Congress - Lexicography in Global Contexts, 17-21 July 2018,* Ljubljana, pp. 17-24.

Hausmann, F. J. (1985). Trois paysages dictionnairiques: la Grande-Bretagne, la France et l'Allemagne. Comparaisons et connexions. In *Lexicographica*, 1, pp. 24-50.

Heinz, M. (1993). *Les locutions figurées dans le « Petit Robert ». Description critique de leur traitement et propositions de normalisation.* Tübingen: Max Niemeyer Verlag.

Lafenestre, G. (1882). Chapitre 1. La sculpture italienne aux XIIIe et XIVe siècles. In *Maitres anciens: études d'histoire et d'art*, Paris: H. Loones, pp. 1-19.

Lerat, P. (1995). *Les langues spécialisées*, Paris: PUF.

Loock, R. (2016). *La traductologie de corpus*. Villeneuve D'Ascq: Presses universitaires du Septentrion.

Pruvost, J. (2000). *Dictionnaires et nouvelles technologies*. Paris: PUF.

Pruvost, J. (2006). *Les dictionnaires français outils d'une langue et d'une culture*. Paris: Ophrys.

Michel, E. (1901). Le Dessin chez Leonard de Vinci. In *Revue des Deux Mondes*, 1, 5e periode, tome, pp. 342-375.

Moran, R. (1994). *Secrets de peintres: apprêts, marouflage, médiums, pigments, glacis, vélatures, dorure, vernis*, Paris: Fleurus.

Palustre, L. (1892). Chapitre III. In *L'Architecture de la Renaissance*, Librairies-imprimeries réunies, pp. 93-134.

Resche, C. (2001). Réflexions sur la frontière entre langue générale et langue spécialisée. In M. Mémet et M. Petit (eds.), L'anglais de spécialité en France. Bordeaux : Geras Editeur, pp. 37-46.

Rundell, M. & Stock, P. (1992). The Corpus Revolution. *English Today*, 30/31/32.

Veyrat, Ch. (1995). *L'intelligence du Petit Robert. Anatomie d'un dictionnaire*. Québec: Les éditions Logiques.

Zotti, V. (2008). I dizionari in Francia e in Italia: due tradizioni a confronto al servizio dell'apprendimento linguistico. In: *Lessicografia e metalessicografia francese e inglese oggi*. Fasano-Parigi: Schena-Alain Baudry et Cie, pp. 59-77.

Dictionnaires

Adeline, J. (1885). *Lexique des termes d'art*, Paris: Quantin Editeur.

Dictionnaire de l'Académie Française, 9^{e} édition : www.dictionnaire-academie.fr

Diderot, J. & D'Alembert, J.-B. (1751-1765). *L'Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*, Paris: Briasson.

Félibien, A. (1676-1690). *Des principes de l'architecture, de la sculpture, de la peinture, et des autres arts qui en dépendent. Avec un Dictionnaire des termes propres à chacun de ces arts*, Paris: J.-B. Coignard.

Hatzfeld, A. et Darmester, A. (1926). *Dictionnaire général de la langue française du commencement du 17e siècle jusqu'à nos jours*, 2 voll., Paris: Delagrave.

Littré, E. (1883). *Dictionnaire de la langue française*, 5 voll., Paris: Hachette.

Rey, A. et Rey-Debove, J. (dir.) (1993). *Le nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue française*, Paris, Le Robert.

Rey, A, (dir., 1984). *Le Grand Robert de la langue française. Dictionnaire alphabétique et analogique de la langue française de Paul Robert*, 9 vol., Paris : Le Robert [version électronique 2020].

Rey, A. et Rey-Debove, J. (dir.). *Le nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue française*, Paris: Le Robert [version électronique 2020].

Trésor de la langue française. Dictionnaire de la langue du XIX^{e} et XX^{e} siècle (1789-1960), 1971-1994, édité par P. Imbs (vol. 1-10), Paris, CNRS, et par B. Quemada (vol. 11-16), Paris: Gallimard ; version informatisée: http://atilf.atilf.fr/

Viollet Le Duc E.-E. (1854-1868). *Dictionnaire raisonné de l'architecture française du XIe au XVIe siècle*, Paris: Bance éditeur.

Corpus

Frantext https://www.frantext.fr

FrTenTen17: https://www.sketchengine.eu/frtenten-french-corpus

LBC Français: http://corpora.lessicobeniculturali.net/

Acknowledgements

Nous tenons à remercier le Département de Langues, Littératures et Cultures Modernes (LILEC) de l'Université de Bologne en Italie qui nous a accordé un financement pour la participation à ce colloque dans le cadre du projet d'excellence "DIVE-IN Diversity & Inclusion".

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Lexicography and Language Technologies

Towards Automatic Definition Extraction for Serbian

Stanković R.[1], Krstev C.[1], Stijović R.[2], Gočanin M.[2], Škorić M.[1]

[1] University of Belgrade, Serbia, 2Institute for the Serbian Language of SASA, Serbia
ranka.stankovic@rgf.bg.ac.rs

Abstract

The paper presents preliminary results of the automatic extraction of candidates for dictionary definitions from unstructured texts in the Serbian language with the aim of accelerating dictionary development. Definitions in the Serbian Academy of Sciences and Arts (SASA) dictionary were used to model different definition types (descriptive, grammatical, reference-based and synonym-based) having different syntactic and lexical features. The research corpus consists of 61,213 definitions of nouns, which were analysed using Serbian morphological e-dictionaries and local grammars implemented as finite state transducers in an open-source corpus processing suite Unitex. The 21 models developed up to the present moment cover 57% of dictionary definitions, 83% of which were fully recognized. The analysis has shown that many definitions have a structure that can be modelled, as evidenced by the statistics of definitions grouped by type. These models were used to retrieve noun definitions from a 1.4-million-word corpus containing 25 primary and secondary school textbooks covering various domains. The obtained results were thoroughly analysed, and guidelines were offered for their improvement.

Keywords: definition modelling; definition extraction; Serbian; automatization of dictionary-making; local grammar

1 Introduction

In the age of electronic lexicography, in which the goal is fast production of dictionaries and the products derived from them, special attention is paid to automatic or semi-automatic performance of some tasks. The use of electronic dictionaries can speed up the work on dictionary production by automatically associating some grammatical information to lemmas, namely, word class, word forms and different types of markers (Krstev 2008). Since the research related to extraction of dictionary examples has shown that information extraction from a corpus can be used to speed up the work on the Serbian Academy of Sciences and Arts Dictionary of Serbo-Croatian Literary and Folk Language (SASA Dictionary) and other dictionaries (Stanković et al. 2019; Kitanović et al. 2021), we focused our present research on the extraction of the sentences contained in the definition. The extraction also implies recognition of paradigmatic lexical relations, e.g. synonyms, antonyms, hypernyms, hyponyms. The problem of automatic extraction of definitions from the text has not been thoroughly researched for the Serbian language so far. The assumption is that its solution could significantly contribute to the acceleration of the development of the SASA dictionary, as well as other dictionaries.

Definition extraction is a relevant task in different areas. In addition to dictionary writing, it is also important in the domain of question answering, namely responding to ‘What-is’ questions, as well as for ontology development. The context in which we will apply definition extraction is semi-automatic creation of dictionaries. In this paper, we present an approach for definition modelling and extraction, relying on the existing Serbian dictionaries (morphological and descriptive), as well as the results of the preliminary experiments in automatic extraction of definitions from unstructured Serbian text.

Definition extraction task can be formalized either as a sentence classification task (i.e., containing term-definition pairs or not) or a sequential labelling task (i.e. identifying the boundaries of terms and definitions). The previous work on definition extraction can be classified as follows: 1) the rule-based approach, with linguistic rules and templates that capture patterns to express term-definition relations; 2) the feature engineering approach relying on the statistical machine learning models with syntax and semantic features; 3) the deep learning approach, using word embeddings via multiple layers of neural networks (Veyseh et al. 2020).

Barnbrook (2002) claimed that definition is a basic activity of language, of particular importance to linguists because of its use of language to describe itself. He described the subset of general language used in definition sentences and the development of a taxonomy of definition types, a grammar of definition sentences and parsing software which can extract their functional components. We were inspired in our work by his definition type taxonomy (structural pattern) and definition language grammar.

The second section of this paper discusses the problem of dictionary definition modelling and methods for automatic extraction of candidates for dictionary definitions. The third section provides an analysis of lexical and syntactic characteristics of dictionary definitions in Serbian based on the corpus of definitions of nouns from the five digitized volumes of the SASA dictionary and presents the models developed for noun definitions taking the form of local grammars that can be used for recognition and extraction. These models were applied to a corpus consisting of textbooks and the results achieved in definition extraction are presented in the fourth section, with examples of integration of definitions into different dictionaries. The achieved results are analysed and plans for further research presented in the concluding section.

2 Methodologies for Definition Extraction

According to the standard "ISO 1951:2007, Presentation/representation of entries in dictionaries — Requirements, recommendations and information" a definition is "A statement that describes a concept and permits its differentiation from other concepts within a system of concepts.". According to the standard "ISO 1087:2019 (en) Terminology work and terminology science — Vocabulary", a definition is "representation of a concept by an expression that describes it and differentiates it from related concepts". This standard distinguishes intentional definition, that conveys the intention of a concept by stating the immediate generic concept and the delimiting characteristic(s), and extensional definition, that enumerates all the subordinate concepts of a superordinate concept under one criterion of subdivision. Intentional definitions are preferable to other types of definitions because they clearly reveal the characteristics of a concept within a concept system: they should be used whenever possible.

A lexicographic definition is the identification of the semantic content of a certain lexeme, with relevant elements of realization. The semantic content consists of an archiseme, which carries information about the lexeme belonging to a wider lexical-semantic group, and a lower-ranking seme, which carries information about individual characteristics of a lexeme, based on which one lexeme differs from another in the same lexical-semantic group. The basic types of lexicographic definition are descriptive, synonymous, and combined - descriptive and synonymous. A descriptive definition without synonyms is used to define the basic meanings of simple, non-derived words (which have no synonyms), and a synonymous one, which consists of words of close or the same semantic structure, when there is a near synonym of that lexeme or when that token is marked with wider uses (Gortan-Premk 2014: 131–132). Combined definitions are used in large descriptive dictionaries - the descriptive part identifies the term directly and lists the elements of meaning, and the synonym refers to the semantic content indirectly (Gortan-Premk 1980: 111–112; Gortan-Premk 1983).

The problem of automatic definition extraction is a relevant task in different areas of natural language processing including Question Answering systems that synthesize answers to questions of the type "What is…" based on one or more sources, and dictionary writing and ontology development (Navigli & Velardi 2010). The approaches to solving this problem are often based on the development and application of lexical-syntactic patterns. For example, in English, the following patterns are often found in definitions (Jin et al. 2013):

<term> defined (as|by) <definition>
define(s)? <term> as <definition>
definition of <term> <definition>
<term> a measure of <definition>
<term> is DT <definition> (that|which|where)
<term> comprise(s)? <definition>
<term> consist(s)? of <definition>
<term> denote(s)? <definition>
<term> designate(s)? <definition>
<definition> (is|are|also) called <term>
<definition> (is|are|also) known as <term>
"part of/in/on/…", "type of", "is <hypernym> …", …

However, Navigli and Velardi point out that the application of such patterns can give a weaker response (they do not recognize enough definitions) and less precision (they recognize sentences that are not definitions), due to very variable syntactic structures that define the terms. For text definition modelling, they suggest using WordClass Lattices (WCLs), a generalization of word grids and an alternative to lexical-syntactic patterns for which they believe both precision and recall are improved. The lattice is a directed acyclic graph, a subclass of finite automata, aiming to preserve the main differences between different sequences, while at the same time removing redundant information. Pattern comparison is based on the use of an asterisk (wildcard *), which facilitates the clustering of sentences. Each sentence class is then generated by a grid of word classes (each class is either a high-frequency word or a word's part-of-speech). A key feature of the approach is the ability to identify definitions and isolate hypernyms.

The task of automatic definition extraction can be formalized in different ways. One of the possibilities is to consider it as a problem of classifying sentences into those that are potential candidates for defining a term and those that are not, namely, as a problem of determining whether a sentence contains a term-definition pair or not. Automatic extraction of definitions can also be regarded as an annotation task where it is required to identify the boundaries of concepts and their definitions. For example, in the sentence "**Virus** je *program ili kôd koji se sam replikuje u drugim datotekama s kojima dolazi u kontakt.…*" (A virus is a program or code that replicates itself in other files it comes in contact with) the headword and its definition would be marked with XML tags as follows: *<headword>Virus</headword> je <def> je program ili kôd koji se sam replikuje u drugim datotekama s kojima dolazi u kontakt. </def>*

Morphosyntactic patterns applied to the computational linguistics corpus are used to extract candidates for definitions for Slovenian and English (Pollak et al. 2012) by automatic recognition of terminology and semantic annotation of WordNet meanings. The tool uses patterns of the form "NP is NP", "NP refers to NP", "NP denotes NP", where NP refers to the noun phrase. The authors also use the Slovenian WordNet to detect definitions: sentences that potentially contain definitions begin with a word from the WordNet and at least one more word belonging to the same chain of hypernym/hyponym relations. The system passes information about the preferred position of the term for which a definition is sought: whether it must be at the beginning of the sentence, after predefined initial forms, which is the default choice, or whether it can be anywhere in the sentence.

Spala et al. (2019) associate a detailed annotation scheme with the corpus in order to explore diverse structures of term definitions in free and semi-structured texts. In addition to the basic concept (Term) and its main definitions (Definition), sentence segments containing pseudonyms or additional names (Alias Term) are also annotated and associated with the basic term. Likewise, noun phrases (Referential Term) that refer to the previously marked term, secondary definitions (Secondary Definition) with additional information, qualifiers (Qualifier) that specify any conditions under which the definition applies and alike are also annotated. The tagged segments are connected by appropriate relations.

Ristić et al. (2018) analyze vertical chaining of concepts present in their lexicographic definitions using the thematic field 'house, building' as an example, pointing to a series from the highest levels of conceptual categorization, complex and simple primitives, to lower levels of hypernyms (PLACE> AREA) [WHERE], AREA > SPACE [SOMETHING THAT EXTENDS (BOUNDLESSLY) IN ALL DIRECTIONS]). They also state that this type of research could contribute to introducing advanced search of digitized editions of descriptive dictionaries of Serbian.

The context in which the definitions for automatic extraction are analyzed and formalized in our paper is the support of dictionary drafting (Kilgarriff & Rychlý 2010), which implies the development based on corpora. The results were achieved by using electronic dictionaries of the Serbian language and local grammars developed based on the results of some previous similar research, particularly (Barnbrook 2002), analysis of definitions in existing dictionaries and previous research into Serbian (Krstev et al. 2015).

3 Analysis and Recognition of Definitions in the SASA Dictionary

The first step in our research was a thorough analysis of various lexical and syntactic features of definitions in the Serbian Academy of Sciences and Arts (SASA) dictionary; this part of a dictionary entry is presented in italics. The definition structure in the SASA dictionary is informally outlined in the Guidelines for Dictionary Processing. The guidelines suggest that the descriptive definition comes first, followed by the definition by substitution with related words; ex: **безвлашће**… *стање без државне власти, анархија, безакоње* (powerlessness, ... state without power of a state, anarchy, lawlessness). The closest senses can be separated just by semicolons, thus representing a single definition; eg: **годишњак**…*периодична публикација која излази једанпут годишње; штампани годишњи извештај неке установе.* [yearbook … periodical published once a year; printed annual report of an institution].

We will illustrate a few models given in the guidelines: if a noun is to be defined by a relative clause, the definition should begin with the expression "онај који ..." ("one who ..."), as opposed to adjectives where definitions begin with "који..." ("which ..."). For abstract nouns ending with *-ост*, *-ство*, *-ота*, *-оћа*, *-ина*, etc. first a general definition is given in the form: osobina, stanje ili svojstvo (an attribute, a state or a feature) *онога који је* (of one who is) */ онога што је* (which is), followed by the adjective from which the abstract noun was derived. This general definition may occasionally be supplemented by two or at most three synonyms.

For the construction of the system for automatic extraction of definitions, it was necessary to formalize models of definition types, for which the Guidelines for Dictionary Processing alone were not enough. It was necessary to create a corpus that would provide the means for conducting an analysis of examples of definition structures. Inspired by the endeavours of large lexicographic houses and research centres that have recognized the possibilities provided by lexicographic databases, especially in supporting modern dictionary use (Rundell 2014), we formalized the microstructure of a SASA dictionary entry (Stijović and Stanković 2017). This made possible the design of a lexical database that can store a structured record of a dictionary article in a relational structure, and the development of a software solution that transforms the unstructured text of a Word document into a relational database (Stanković et al. 2018).

These preliminary steps led to the construction of a research corpus consisting of definitions from the five digitized volumes of the SASA dictionary (SASA Dictionary 2019; Stanković et al, 2018). The corpus contains 61,213 noun definitions, 19,708 verb, 12,312 adjective, 2,564 adverb definitions and 2,967 definitions of other word classes. The main aim of developing models of clauses used in SASA dictionary definitions is to enable formalization of definitions for some future dictionaries - to establish which models are the most frequently used, which are preferable, and which definitions unnecessarily deviate from the common patterns.

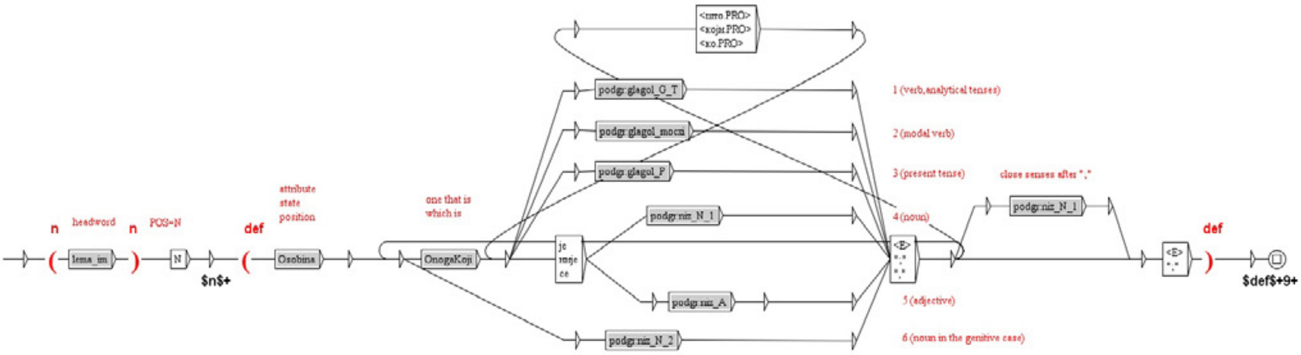
The definitions of nouns from the SASA dictionary were analysed using Serbian morphological e-dictionaries and local grammars in the form of Finite-State Transducers (Krstev 2008) and implemented in the Unitex corpus processing suite[1]. Electronic morphological dictionaries of Serbian intended for automatic processing have been undergoing development for many years now, and their current size and content ensure successful use in different real-world applications in the field of Serbian language processing. These dictionaries contain automatically generated word forms of more than 200,000 lemmas[2], and their content covers both general vocabulary and proper names - personal, geopolitical, names of organizations and the like. Moreover, the dictionary contains multi-word units, which are recorded in traditional dictionaries as syntagms or phrases. The basic unit of these dictionaries is a word form associated with its lemma (usually the headword of a traditional dictionary entry), Part-Of-Speech, markers that describe its syntactic, semantic, usage, dialectal and domain properties, followed by the codes of its possible morphosyntactic realization.

Finite state transducers (FSTs) are abstract mathematical constructions that allow modelling of local grammars to describe some linguistic constructions, for example, noun phrases. A finite state transducer "passes" through the text it analyses to compare a text chunk with the model it represents. In the case of successful recognition, a final state transducer produces some result, which can be a modification of the source text by adding tags for types of recognized words or a recognized syntactic structure (Vitas & Krstev 2012). Finite state transducers are visualized by graphs for

[1] Unitex/GramLab - Lexicon-Based Corpus Processing Suite (https://unitexgramlab.org/).

[2] A part of this lexicon is publicly available for use within the Unitex system.

easier development and use. A local grammar and its corresponding graph that models (and recognizes) definitions of nouns that represent attributes and/or state is given in Figure 1. Below the graph, six definitions are listed that are recognized by the corresponding paths in the graph.



1. **адаптираност** N *стање онога што је прилагођено, прилагођеност, подешеност*. (the state of what is adjusted, the adjustment, the setting)
2. **оповргљивост** N *особина онога што се може оповргнути*; (a property of what can be refuted)
3. **вољкост** N *особина онога што ствара добро расположење*, (an attribute of that which creates a good mood)
4. **паланчанство** N *стање, особина онога који је паланчанин, паланачко порекло, паланачки дух и сл.* (condition, characteristic of the one who is from a small town, a small-town origin, a small-town spirit, etc.)
5. **опречност** N *особина, стање онога што је опречно, супротност, противречност*; (a property, a state of what is opposite, the opposite, the contradiction)
6. **богомољство** N *особина, својство богомољца, претерано ревносна побожност*. (a trait, a property of a worshiper, overzealous piety)

Figure 1: A local grammar that models (and recognizes) definitions of nouns that represent attributes and/or state.

From the point of view of automatic recognition of definitions in the SASA dictionary, definitions can be divided into three main groups. The first group consists of highly schematized definitions that refer to other dictionary entries, e.g. pointing to a common or literary form ("see…"), or to a derived word that retained the meaning of the basic word ("diminutive of …"). These definitions are easy to model. The second group consists of "definitions" of some types of proper names, e.g. names of holidays, saints, monasteries, etc. These definitions are similar to the explanations given in encyclopaedias, they are expressed freely, and are difficult to model with local (shallow) grammars. One example is **Брашанчево** N *католички црквени празник који пада у други четвртак после Духова* (a Catholic church holiday that falls on the second Thursday after Pentecost). The third group consists of proper definitions that are schematized to a certain extent. For instance, feminine gender nouns derived by gender motion have a schematized definition, e.g., **бунтовница** N *жена бунтовник* (woman rebel). However, these definitions are often expanded with additional information, e.g., **верница** N *жена верник, припадница неке религије* (a woman believer, a member of a religion).

We have developed 21 definition models covering definitions from all three groups to some extent: 7 for the definitions from the first group, one for the definitions from the second group and 13 for the definitions from the third group. When developing these models by means of local grammars with FSTs we tend to capture all types of definitions, namely descriptive, reference-based and synonym-based. The coverage of noun definitions in the SASA dictionary by the developed local grammars is represented in Table 1. We have mentioned before that our corpus of definitions consists of 61,213 definitions of nouns. During the development of local grammars, we reduced this set by excluding definitions containing unknown words (not recorded in the e-dictionaries used by local grammars) since local grammars even if appropriate for them could not be successful. We thus obtained a set of 51,729 definitions.

The analysis of definitions has shown that the SASA dictionary contains a large number of schematized definitions (34% of all definitions) and they have mostly been fully recognized (from 91 to 100% depending on type). As expected, definitions of encyclopaedic entries are difficult to capture (only 26% of recognized definitions were fully recognized) and it is possible that quite a number of them were not recognized at all. Definitions schematized to a certain extent were covered with a differing success rate depending on their type, ranging from 38% for nationalities to 87% for "type of…" definitions. Some of the definition types look very specific, for instance, model 19 describing devices, tools etc. This type is specific because it uses a closed set of words for describing artefacts which are further distinguished by prepositional phrases that describe their purpose. Once developed, this type of definition will be used in the models that still have to come to light.

Local grammars that model definitions of nouns (and other types of words) will contribute to the creation of dictionaries and other lexical resources in various ways. For instance, when creating a dictionary, they will be used to check the compliance of definitions with the adopted forms. In addition, they enable automatic linking of dictionary entries in the database, which was illustrated by the examples in Table 1. Moreover, definition models will enable not only automatic linking of the entries in the SASA database, but also enrichment and linking of other lexical resources for Serbian (morphological e-dictionaries and WordNet). Finally, as will be shown in the next section, local definition grammars can also be used to extract definitions from domain corpora.

Graph (group)	Type	No of recognised definitions	To the end	Example
1 (1)	see	10849	9830 (91%)	**акушерка** N *в.бабица* (accoucheuse N v. midwife)
2 (1)	surname	3672	3670 (100%)	**Андрић** N *презиме* (Andrić N surname)
3 (1)	verbal noun	913	912 (100%)	**плакање** N *гл. им. од плакати* (crying N verb. n. from to cry)
4 (3)	type of	1435	1250 (87%)	**бисер** N *врста пенушавог вина* (pearl N a kind of sparkling wine)
5 (1)	diminutive	1629	1575 (97%)	**ветрић** N *дем. и хип. од ветар;* (small wind N dim. and hyp. of wind)
6 (1)	first name or a nickname	914	857 (94%)	**Пачавра** N *лични надимак* (Pačavra N personal nickname)
7 (3)	women	596	483 (81%)	**бедевуша** N *танка, висока незграпна жена.* (bedevuša N a thin, tall clumsy woman)
8 (1)	augmentative	464	454 (98%)	**бубетина** N *аугм. и пеј. од буба.* (big bug N aug. and pej. from bug)
9 (3)	attribute or condition	507	379 (75%)	**плиткоумност** N *особина, својство онога који је плиткоуман,* (shallowness N a trait, a property of one who is shallow-minded)
10 (3)	geographic name	436	280 (64%)	**Абисинија** N *ранији назив за Етиопију.* (Abyssinia N former name of Ethiopia.)
11 (3)	inhabitant	214	178 (83%)	**Папуанка** N *припадница домородачког становништва Папуанаца* (Papuan woman N a member of the indigenous population of the Papuans)
12 (3)	people	21	8 (38%)	**Бугари** N *јужнословенски народ* (Bulgarians N South Slavic people)
13 (2)	proper names (facilities, deities, astral bodies, holidays)	94	24 (26%)	**Вечерњача** N *народни назив за планету Венеру* (Večernjača N folk name of the planet Venus)
14 (1)	collective nouns	103	101 (98%)	**браћа** N *зб. им. од брат* (brothers N col. noun derived from brother)
15 (3)	animals	2115	1564 (74%)	**арнаут** N *назив за помамна, љута коња;* (arnaut N name for a mad, angry horse;)
16 (3)	plants	1371	1005 (73%)	**брусница** N *шумска ниска жбунаста биљка…* (cranberry N low shrubby forest plant…)
17 (3)	plant and animal organs	84	56 (67%)	**папоњак** N *мали, закржљали клип кукуруза.* (paponjak N small, stunted corn cob)
18 (3)	the one who is… that what is…	2221	978 (44%)	**балавац** N *онај који је недорастао, незрео,* (snot-nosed kid N one who is not grown-up, immature **пикантерија** N *оно што је пикантно* (piquancy N what is spicy)
19 (3)	tool, device, machine…	340	168 (49%)	**алтиметар** N *справа за мерење надморске висине* (altimeter N device for measuring altitude)
20 (3)	set, part…	2874	1984 (69%)	**бифтек** N *комад говеђег меса од леђне печенице* (beefsteak N a piece of roast beef from the back
21 (3)	prepositional clause	715	343 (48%)	**бавриљуга** N *у дечјој бројаници* (bavrljuga N in a children's tongue twister)

Total		**31567 (61%)**	**26099 (83%)**	2125 definitions recognised by more than 1 graph
Different		**29442 (57%)**	**24347 (83%)**	

Table 1: The analysis of recognized definitions from the SASA dictionary; the defined word is given in small caps, connected entries are given in bold, while trigger words enabling the connection are given in italic.

4 Corpus Analysis

4.1 Creating a Textbook Corpus

For testing the automatic extraction of definitions from unstructured text by means of local grammars presented in the previous section, we created a corpus of 25 primary and secondary school textbook covering the following domains: biology, history, logic, music, computer science, physics and design and technology (technological education). The biology domain is covered by seven textbooks: one for primary school, four for high school, and two for agricultural professional / vocational school; history domain is represented by four primary school textbooks; computer science is represented by one primary school textbook and tree high school ones, while physics is represented by one primary school textbook. General technique is represented by three textbooks for professional/vocational technical schools relating to the following subjects: power engineering, mechanical engineering for agriculture and tools and mechanization. Logic and philosophy, on the other hand were covered by two high school textbooks focusing on humanities subjects. The domain of music is represented by two music high school textbooks: History of Music and Century of Jazz. The corpus is being developed and 35 textbooks have been used for the purposes of this experiment. However, the current version contains 31 textbooks and it is expected to grow in the near future. The textbooks were scanned, optical character recognition was performed, and recognition errors were manually corrected (though a certain number of OCR errors remained). The corpus consists of 85,628 sentences, 3,4M tokens (165K different) and 1,4M words (165K different).

The corpus was processed by using electronic dictionaries of the Serbian language that recognized approximately 238,000 word forms, 5,000 multi-word units, while 5,700 word forms remained unrecognized. Among the unrecognized words are the remaining optical character recognition errors, foreign names and abbreviations (Facebook, WWW, HTML, SMS, RNA, ATR, HIV), as well as a number of words belonging to domain-specific terminology, such as: biology - *eukariote* (eukaryotes), *citoplazma* (cytoplasm), chemistry - *acetilhlorin* (acetylchlorine), *organele* (organelles), *hloroplas*t (chloroplast), music - *diksilend* (dixieland), etc. The majority of unknown words are the result of OCR errors that remained in the text, but we are working on correcting them.[3]

4.2 Recognition of Candidates in the Textbook Corpus

To determine whether it is possible to recognize definitions of domain-specific terms in the domain corpus text, a subset of local grammars presented in Section 3 was applied to the corpus consisting of 25 textbooks; namely we excluded models for schematized definitions that link entries in the dictionary, since they are not relevant to the unstructured texts. Table 2 shows examples of successful, or partly successful, definition recognition. It should be noted that in some cases only the initial parts of the definition were recognized, for example, full definition for *drama* is **Драма** је *врста књижевног дела које настаје да би се изводило на позорници* (Drama is a type of literary work that is created to be performed on stage) (the underlined sequence was not recognized, see example 4. in Table 2). On the other hand, some long and useful definitions were completely retrieved, e.g. **Научни систем** је *јединствен скуп знања (чињеница, закона, теорија) која су међусобно повезана и сређена на основу извесних принципа* (A scientific system is a unique set of knowledge (facts, laws, theories) that are interconnected and arranged on the basis of certain principles.). Automatic recognition of complete definitions is a complex task, because it requires grammars that perform a complete parse. However, even partial recognitions can be useful for pointing to the sentences that potentially contain definitions. The definitions recognized in the textbook corpus are in some cases similar to the definitions in the SASA dictionary, e.g. definitions for the orchid family (example 1 in Table 3), while in other cases they differ, stressing a different attribute, e.g. definitions of *alcoholic* (example 2 in Table 3). In some cases, there are several definitions for different senses of a lemma, while textbooks offer yet another one, e.g. definitions of *winch* (example 3 in Table 3). On the other hand, three related but different definitions were retrieved from the corpus for the bird family (example 4 in Table 3; note that this lemma has not yet been covered by the SASA dictionary). In some cases, definitions were retrieved for domain specific terms that are not listed in the SASA dictionary, e.g. the definition of *bentonite* (example 5 in Table 3).

[3] The automatically lemmatized and POS tagged corpus of textbooks (Stanković et al. 2020) is available on the corpus search platform of the Jerteh Society for Language Resources and Technologies https://noske.jerteh.rs/#dashboard?corpname=SkolKor. The search system for monolingual and multilingual corpora is based on the NoSketch Engine platform (Rychly 2007), which is open-source software - a simplified version of the commercial Sketch Engine tool. Each textbook is supplied with metadata: title, autor, publisher, year of publishing, subject, school level (primary, secondary) and school class. As a guest, a user can presently search several corpora under NoSkatchEngine more corpora will be available in the near future.

domain	scope	recognized	correct	examples
biology	full	44	40	**Биљне ваши** су *ситни инсекти који се хране биљним соковима.* (Plant lice are tiny insects that feed on plant juices.) (15) **Протеини** су *група органских молекула са разноврсним функцијама...* (Proteins are a group of organic molecules with various functions …) (20)
	partial	97	94	**Бреза** је *брзорастућа врста чије се дрво* (Birch is a fast-growing species whose tree) (4)
history/ logic/ music	full	10	10	**Авари** су *номадско племе турског порекла.* (The Avars are a nomadic tribe of Turkish origin.) (12)
	partial	55	34	**Драма** је *врста књижевног дела које настаје...* (Drama is a type of literary work that is created …) (4)
computer science/ design and technology (technological education).	full	13	12	**Рачунарска мрежа** је *скуп међусобно повезаних рачунара који комуницирају и размењују информације.* (A computer network is a set of interconnected computers that communicate and exchange information.) (20)
	partial	57	45	**Програмски језик** је *скуп правила којима се рачунару…* (A programming language is a set of rules that ... to a computer…) (20)

Table 2. Examples of definitions recognized in the textbook corpus.

comparison	source	examples
1. similar	SASA	**орхидеја** N *у мн.: фамилија вишегодишњих зељастих украсних биљака* (orchid N in pl.: family of perennial herbaceous ornamental plants)
	textbooks	**Орхидеје** су *вишегодишње зељасте биљке,...* (Orchids are perennial herbaceous plants, ...)
2. different	SASA	**алкохоличар** N *човек који се одао пићу*; (alcoholic N a man who indulges in drinking;)
	textbooks	**Алкохоличар** је *особа која показује зависност од алкохола, који штетно делује на организам,...* (An alcoholic is a person who exhibits alcohol addiction, which has a harmful effect on the body, ...)
3. multiple in SASA	SASA	**витло** N *назив за разне направе које се окрећу*; (winch N name for various rotating devices;) **витло** N *обртна направа у виду елисе (у машини).* (winch N rotating device in the form of a propeller (in a machine).)
	textbooks	**Витло** је *уређај који углавном служи за подизање или вучу терета.* A winch is a device that is mainly used for lifting or towing loads.
4. multiple definitions	textbooks	**птице** су *далеко најбројнија група копнених кичмењака,...* (birds are by far the most numerous group of terrestrial vertebrates, ...) **Птице** су *јединa основна група данашњих кичмењака која има…* (Birds are the only basic/core group of today's vertebrates that has…) **Птице** су *најмлађа основна група копнених кичмењака.* (Birds are the youngest basic/core group of terrestrial vertebrates.)
5. not in SASA	textbooks	**Бентонит** је *врста глине (хидратисани алуминијум…* (Bentonite is a type of clay (hydrated aluminum ...)

Table 3. A comparison of definitions in the SASA dictionary and those extracted from the textbook corpus.

The productivity of the developed models is represented in Figure 2. It is noticeable that model 20 (defining sets, groups, parts, etc.) is by far the most successful in retrieving definitions from the textbook corpus. Presently it covers 9% of all noun definitions in the SASA dictionary surpassing all other models, except those applied to schematized definitions.

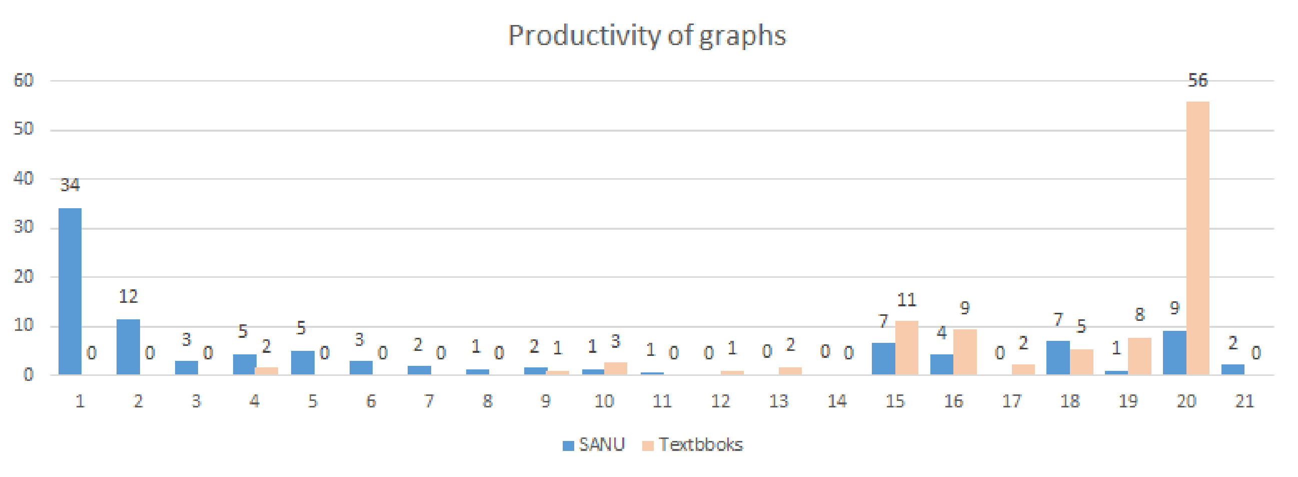


Figure 2. Productivity of models on the SASA dictionary and the corpus of textbooks.

The results obtained by this experiment can be considered preliminary and promising, but far from satisfactory. It is necessary not only to refine local grammars, but also to develop additional models. Besides, more possibilities to connect a word to its definition should be explored; namely, on this occasion we used only one pattern: **word** is/are *definition*. It can be seen in Table 2 that especially partially recognized definitions are not always correct, which is most prominently present in the history, logic, and music domains. This means that more strict conditions (like gender and/or case agreement) should be used for extraction from unstructured texts than are necessary when modelling dictionary definitions.

5 Conclusion

The paper presents preliminary results of the automatic extraction of candidates for dictionary definitions from unstructured texts in the Serbian language, with the aim of accelerating the development of dictionaries. In order to model different types of dictionary definitions, we used a corpus consisting of definitions taken from the five digitized volumes of the SASA dictionary. The analysis has shown that a large number of noun definitions have a structure that can be modelled (57%). Our long-term objective is definition extraction from unstructured texts. To that end, we prepared a corpus consisting of primary and secondary school textbooks to which we applied the models developed for the SASA dictionary.

While the rule-based approach is intuitive and has high precision, it suffers from the low recall issue. In order to overcome it, we plan to investigate other methods to complement this approach. Hill et al. (2016) report the effectiveness of both neural embedding architectures and definition-based training for developing models that understand phrases and sentences, with two applications of these architectures: reverse dictionaries that return the name of a concept given a definition or description and general-knowledge crossword question answers.

Inspired by the work of Noraset et. al (2017), Bosc & Pascal (2018), our model will be improved by learning to compute word embeddings by processing dictionary definitions and trying to reconstruct them. We will use Dict2vec, based on natural language dictionaries that builds new word pairs from dictionary entries, so that semantically related words are moved closer, and negative sampling filters out pairs whose words are unrelated in dictionaries. (Tissier et al. 2017) We will investigate whether definition models can improve standard word embeddings.

6 References

A Style Guide for Dictionary-Making, Belgrade: SASA Institute for Serbo(-Croatian) (manuscript), 1959 and (supplemented) 2017 [Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017]

Bosc, T. & Pascal, V. (2018). Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1522-1532.

Barnbrook, G. (2002). *Defining Language, A local grammar of definition sentences, Studies in Corpus Linguistics*, (Vol. 11). John Benjamins Publishing.

Gortan Premk, D. (1980). O gramatičkoj informaciji i semantičkoj identifikaciji u velikom opisnom rečniku. [On grammatical information and semantic identification in a large descriptive dictionary.], *Naš jezik*, XXIV/3, pp. 107–114.

Gortan Premk, D. (1983). Synonymous array in lexicographical definition. [Sinonimski niz u leksikografskoj definiciji.] In *Naučni sastanak slavista u Vukove dane*, 12/1, pp. 45–50.

Gortan Premk, D. (2014). Definisanje u srpskoj leksikografiji. [Definition in Serbian lexicography]. In *Savremena srpska leksikografija u teoriji i praksi*, R. Dragićević (ed.), Beograd: Filološki fakultet, pp. 131–139.

Hill, F., Cho, K., Korhonrn, A. & Bengio, Y. (2016). *Learning to understand phrases by embedding the dictionary*. Transactions of the Association for Computational Linguistics, 4, 17-30.

Jin, Y., Kan, M. Y., Ng, J. P., & He, X. (2013). Mining scientific terms and their definitions: A study of the ACL anthology.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 780-790.

Kilgarriff, A. & Rychlý, P. (2010). Semi-Automatic Dictionary Drafting, In *A Way with Words: Recent Advances in Lexical Theory and Analysis*. Uganda: Menha Publishers Ltd., pp. 299-312.

Kitanović, O., Stanković, R., Tomašević, A., Škorić, M., Babić, I. & Kolonja, Lj. (2021). A Data Driven Approach for Raw Material Terminology. *Applied Sciences,* 11 (7): 2892, pp. 1-22.

Krstev, C. (2008*). Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade.

Krstev, C., Vitas, D. & Stanković, R. (2015). A Lexical Approach to Acronyms and their Definitions. In *Proceedings of 7th Language & Technology Conference, November 27-29, 2015, Poznań, Poland*, eds. Zygmunt Vetulani & Joseph Mariani, 219–223, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań, pp. 219-223.

Navigli, R. & Velardi, P. (2010). Learning Word-Class Lattices for Definition and Hypernym Extraction. In *Proceedings of the Forty-Eighth Annual Meeting of the Association for Computational Linguistics*. *Uppsala, Sweden*. pp. 1318–1327.

Noraset, T., Liang, C., Birnbaum, L., & Downey, D. (2017). Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence,* Vol. 31, No. 1.

Pollak, S., Vavpetic, A., Kranjc, J., Lavrac, B. & Vintar, Š. (2012). NLP workflow for on-line definition extraction from English and Slovene text corpora. In *Proceedings of KONVENS 2012*, *Vienna, 19 September 2012*, pp. 53–60.

Ristić, S., Konjik Lazić, I. & Ivanović, N. (2018). Metajezik leksikografske definicije u deskriptivnom rečniku (na materijalu rečnika srpskog jezika) [Metalanguage of lexicographic definition in descriptive dictionary (on the material of the Serbian language dictionary)]. *Južnoslovenski filolog* 74/1, 2018, pp. 81–96.

Rundell, M. (2014). Macmillan English Dictionary: The End of Print? In *Slovenščina 2.0: empirical, applied and interdisciplinary research, 2.2*, pp.1–14.

Rychly, P. (2007). Manatee/bonito – a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing, 2007*, pp. 65–70.

Tissier, J., Gravier, C., & Habrard, A. (2017). Dict2vec: Learning Word Embeddings using Lexical Dictionaries. In *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Sep 2017, Copenhague, Denmark*. pp. 254-263.

SASA Dictionary: Речник српскохрватског књижевног и народног језика САНУ, I–XXI [The Dictionary of the Serbo-Croatian Standard and Vernacular Language] (1959–2020). Београд: Институт за српски језик САНУ и САНУ.

Spala, S., Miller, N., Yang, Y., Dernoncourt, F. & Dockhorn, C. (2019). DEFT: A corpus for definition extraction in free-and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop, 2019*, pp. 124–131.

Stanković, R., Stijović, R., Vitas, D., Krstev, C. & Sabo, O. (2018). The Dictionary of the Serbian Academy: from the Text to the Lexical Database. In Čibej, Jaka, Gorjanc, Vojko, Kosem, Iztok & Krek, Simon (eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana. Ljubljana University Press, Faculty of Arts. pp. 941–949.

Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D. & Marković, A. (2019). SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian. In I. Kosem et als. (eds.) Electronic lexicography in the 21st century*,* Proceedings of the eLex 2019 conference, 1–3 October 2019, Sintra, *Portugal*, pp. 248–269.

Stanković, R., Šandrih, B., Krstev, C., Utvić M. & Škorić, M. (2020). Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of the 12th International Conference on Language Resources and Evaluation, LREC* eds. Nicoletta Calzolari et al., pp. 3947–3955.

Stijović, R. & Stanković, R., (2018). Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU. [Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary (in Cyrillic)] In: *Naučni sastanak slavista u Vukove dane* 47/1, Beograd: Međunarodni slavistički centar. Filološki fakultet, 2018, 427–440.

Veyseh, A. P. B., Dernoncourt, F., Dou, D. & Nguyen, T. (2020). A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. In *AAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9098-9105.

Vitas, D. & Krstev, C. (2012). Processing of Corpora of Serbian Using Electronic Dictionaries. In *Prace Filologiczne*, vol. LXIII, Warszawa, 2012, 279–292.

Acknowledgements

This paper is supported by the COST Action CA19102 - LITHME "Language in the Human-Machine Era" https://lithme.eu/. Access to SketchEngine is provided by the ELEXIS project funded by the European Union's Horizon 2020 research and innovation program under grant number 731015.

License to use: ELEXIS survey on licensing lexicographic data and software

Kosem I., Nimb S., Tiberius C., Boelhouwer B., Krek S.

Jožef Stefan Institute, Slovenia, Det Danske Sprog- og Litteraturselskab, Denmark, Dutch Language Institute the Netherlands
iztok.kosem@ijs.si, sn@dsl.dk, Carole.Tiberius@ivdnt.nl, Bob.Boelhouwer@ivdnt.nl

Abstract

Lexicographic resources are extremely valuable, not only for the general public but also for other applications, such as natural language processing, linked open data, etc. As many resources are still not available or are only available under very strict conditions, it is important to understand their owners' or creators' stance towards data sharing. This is particularly relevant for the European Lexicographic Infrastructure (ELEXIS) project, which has as one of its main aims the development of the Dictionary Matrix that will be formed of extensive links between key elements found in different types of dictionaries. This paper reports on a survey on licensing lexicographic data conducted amongst partner and observer institutions in ELEXIS. The results show that there are many differences on how institutions in different countries approach data licensing. Moreover, the differences can be observed at the level of dictionary microstructure, as institutions are more protective towards certain types of lexicographic data. Using a case study, it is demonstrated how a more open approach to sharing data can benefit the community of a particular language, and the ELEXIS community in general.

Keywords: licensing; survey; data; ELEXIS; Dictionary Matrix; lexicographic resources; dictionary; corpus

1 Introduction

The creation of a dictionary of quality requires a large amount of highly skilled labour. Therefore, such a product is of high value to its creators, sponsors, and users. The owners or compilers of dictionaries will want to protect their data, but for different reasons. If the owner is a private organisation, the reason will probably lie in the commercial value of the data. Public organisations, on the other hand, will have other reasons, such as the need to prove their relevance to the funding provider by reporting visits to their dictionary website. Furthermore, certain organisations may not be allowed by their funding provider or governing organisation to hand the data to others. The quality of the data may also be a consideration; some organisations might want to maintain tight control over their data in order to avoid that diluted or deprecated versions of the data undermine its usability and the organisation's reputation.

As a result, general (open) access to lexicographic data is still extremely limited, which prohibits reuse of valuable datasets in other fields, such as natural language processing, linked open data and the Semantic Web, as well as in the context of digital humanities. One of the main objectives of the European Lexicographic Infrastructure (ELEXIS) is to address these issues, i.e. to enable access to lexicographic data and to promote an open access culture in lexicography, in line with the *European Commission Recommendation on access to and preservation of scientific information*. Serious efforts have been made within ELEXIS to address these Intellectual Property Rights issues currently preventing the inclusion of lexicographic data into open access infrastructures. In the context of this work, a survey was conducted among partner and observer institutions in order to get information on, and understanding of, their existing licensing practices.

This paper presents some of the main findings of the survey, both from the perspective of current practices and situations at different types of institutions, and in terms of their future plans and concerns. Also, a case study of licensing practices at one of the ELEXIS partner institutions is presented in more detail. The paper concludes by presenting the licensing options that the ELEXIS consortium prepared for partners/observers in order to facilitate the content sharing process, and to obtain enough content for the Dictionary Matrix.

2 European Lexicographic infrastructure (ELEXIS)

ELEXIS (Krek et al. 2018, 2019; Pedersen et al. 2018; Woldrich et al. 2020) is a Horizon 2020 project dedicated to creating a sustainable infrastructure for lexicography. The main objectives of ELEXIS are to (1) enable efficient access to high quality lexicographic data so that it can also be used by other fields including NLP, AI and digital humanities, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources.

To realise these goals, ELEXIS has an inclusive multi-layered organisation that aims at engaging different user groups with various levels of intensity during the project, shown in Figure 1. The core of the organisational structure consists of 17 consortium partners. The consortium is composed of content-holding institutions and researchers with complementary backgrounds: lexicography, digital humanities, standardisation, language technology, Semantic Web and artificial intelligence. Furthermore, the consortium cooperates with existing infrastructures, i.e. CLARIN and DARIAH.

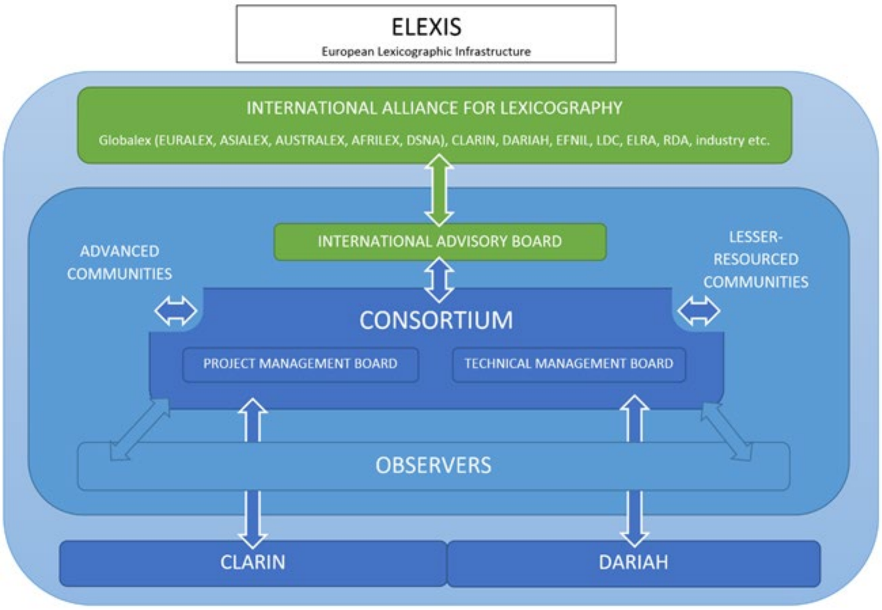

ELEXIS
European Lexicographic Infrastructure
INTERNATIONAL ALLIANCE FOR LEXICOGRAPHY
Globalex (EURALEX, ASIALEX, AUSTRALEX, AFRILEX, DSNA), CLARIN, DARIAH, EFNIL, LDC, ELRA, RDA, industry etc.
INTERNATIONAL ADVISORY BOARD
ADVANCED COMMUNITIES
LESSER-RESOURCED COMMUNITIES
CONSORTIUM
PROJECT MANAGEMENT BOARD
TECHNICAL MANAGEMENT BOARD
OBSERVERS
CLARIN
DARIAH

Figure 1: ELEXIS organisational structure.

Another organisational layer is formed by observer institutions that are directly included in outreach and dissemination activities through various channels. The central group of institutions that fall under the observer category are those producing quality lexicographic data and resources. Typically, but not exclusively, these institutions include (European) national language institutes, large dictionary publishers and other prominent producers of lexicographic data. At the time of writing, ELEXIS had 58 observers.

One of the main data deliverables of ELEXIS will be the Dictionary Matrix that will be formed of extensive links between key elements found in different types of dictionaries - monolingual, multilingual, modern, historical etc. With the Dictionary Matrix, ELEXIS is creating a universal lexicographic metastructure spanning across languages and time. The focus is on (direct or indirect) linking of existing lexicographic resources, minimally on the headword and part of speech level, but where possible also on the level of senses, examples, translations, collocations, and other types of information in lexicographic resources. The Dictionary Matrix will be available as a public service, and the links between dictionary elements will be shared as Linguistic Linked Open Data (LLOD) enabling other fields to exploit the high-quality semantic data from lexicographic resources.

Given that obtaining enough datasets from partner and observer institutions is a key prerequisite for the Dictionary Matrix, and licensing lexicographic content is one of the important topics of discussions between consortium partners, as well as between partners and observers, a survey was conducted among partner and observer institutions in order to gain an insight into their existing licensing practices. The aim of the survey was not only to gain an insight into the licensing situation at different institutions, but also to get an idea of common concerns and problems connected with licensing and sharing lexicographic data.

3 Survey on licensing practices of ELEXIS institutions

The survey was conducted in the final months of 2019. We used the 1ka survey system,[1] which was previously used for several other surveys conducted by members of the team, in ELEXIS as well as in other projects. The survey consisted of 36 questions, 12 of them were multiple-choice and the rest open-ended. Several open-ended questions (many of them optional) were used in order to offer the respondents a possibility to elaborate on their answers. The average time of survey completion was approximately 10 minutes, which was less than expected but probably the consequence of the fact that due to the specific nature of the questions, the respondents had to gather the information in advance, and then enter it into the survey.

The survey was completed by 38 ELEXIS partner and observer institutions from 25 different countries, predominantly from Europe. Almost half of the institutions (18) were public, and 13 of them were universities or university departments. Four responding institutions were non-profit organisations, and two were private companies. Two of the institutions reported to be a mixture of public and private.

3.1 Source of funding

Most of the institutions (28) reported using public funding at the national level for the creation of their lexicographic

[1] https://www.1ka.si/

resources. In most cases, the reported source of funding was the ministry responsible for science, or a research agency/council. Some of the institutions also reported combining national funding with their internal funding or private funding.

Nine institutions reported using public funding at the international level to create their lexicographic resources, either as the only source of funding or in combination with other sources. The reported funding sources included H2020 funding, Marie Curie and ERC grants, European Social Fund, and the European Regional Development Fund. Other types of public funding, used by seven institutions, included specific regional funding, small-scale project funding, or scholarships.

More than a third of the institutions (13) reported using private funding to create their lexicographic resources. The funding sources included sponsorship (by companies, foundations), collaboration with private companies such as publishers, and institution's own funds. One institution reported investing profit from investments into the stock market and real estate into lexicographic resources.

3.2 Intellectual property rights (IPR)

29 institutions (80.5%) reported having a cleared IPR status for their lexicographic data, with 16 institutions having cleared IPR status for all their lexicographic data, and 13 institutions only for some. The reasons for not having the IPR status cleared varied: still trying to negotiate the agreement with the authors of the resources, lack of time, and considering the data as not interesting for external parties. On the other hand, seven institutions reported not having a cleared IPR status for their lexicographic data; many were in the process of sorting it out. It is noteworthy that out of those seven institutions, six receive public funding for their lexicographic resources.

Only nine institutions provided details on the copyright holders of their data. In most cases, the institutions themselves are the copyright holders, with exceptions mainly being limited to particular datasets where the copyright holders are the authors (either (formerly) employed or collaborating externally by contract). Out of ten institutions that commented on how difficult it was to obtain the copyright clearance most said that it was easy; this was related to the fact that they compiled the resources themselves and did not need any external clearance. It should be noted that almost all of these institutions use open licenses for their data. The problems mentioned by some of the institutions were bureaucracy, vague initial contracts with the authors, and the connections between resources (derivatives etc.).

Half of the institutions (N=32) reported having a special person or department dealing with IPR issues. A closer look at further explanations of the answers, however, revealed that none of the institutions had a person or department that specialised solely in IPR. 15 out of 32 institutions had a legal department or a legal expert dealing with all legal issues, and two more institutions reported hiring a legal expert when necessary. At five institutions, a non-legal person - such as deputy director, manager of language resources or head of the centre - deals with IPR issues. One institution reported on the benefits of being in the national CLARIN consortium as the CLARIN department deals with all IPR issues for them.

3.3 Distribution of lexicographic data

63% of the institutions (20 out of 32) reported having a policy on the distribution of lexicographic data. Five of those institutions reported having the policy publicly available, whereas others have an internal policy only. On the other hand, 12 institutions reported not having such a policy (public or internal). It is also noteworthy that seven out of ten universities (70%) that answered this question did not have a policy on the distribution of lexicographic data. The percentage of public institutions without such a policy is significantly lower (30%).

We asked the institutions on how they make their data available, either as content to the language users for consultation purposes, or as datasets. Moreover, we were interested in getting a better understanding of which types of data they are willing or not willing to share. The results are presented in the following subsections.

3.3.1 Current status of data availability

On the question of how they make their lexicographic resources available to the language users, 84% of the institutions reported offering their lexicographic resources to the users online for free, which is in line with the fact that most of the institutions are publicly funded. Only five institutions reported offering (some of) their lexicographic resources online for a fee. Interestingly, many institutions reported publishing paper dictionaries, over half of the respondents in fact (20 out of 38).[2] Six institutions (four public, one non-profit and one a mixture of public and private) reported using a publication model where they publish the paper version first, and provide the online version after some time for free. It is noteworthy that universities or university departments seemed to be more oriented toward (free) online access (Figure 2), whereas public institutions and non-profit organisations exhibited a similar balance of online and paper format.

[2] A similar finding was observed in the survey of user needs (Kallas et al. 2019; Kallas et al. 2020), where it was stated that the reason for still publishing print dictionaries was tradition, i.e. the previous volumes of a dictionaries being also published in print.

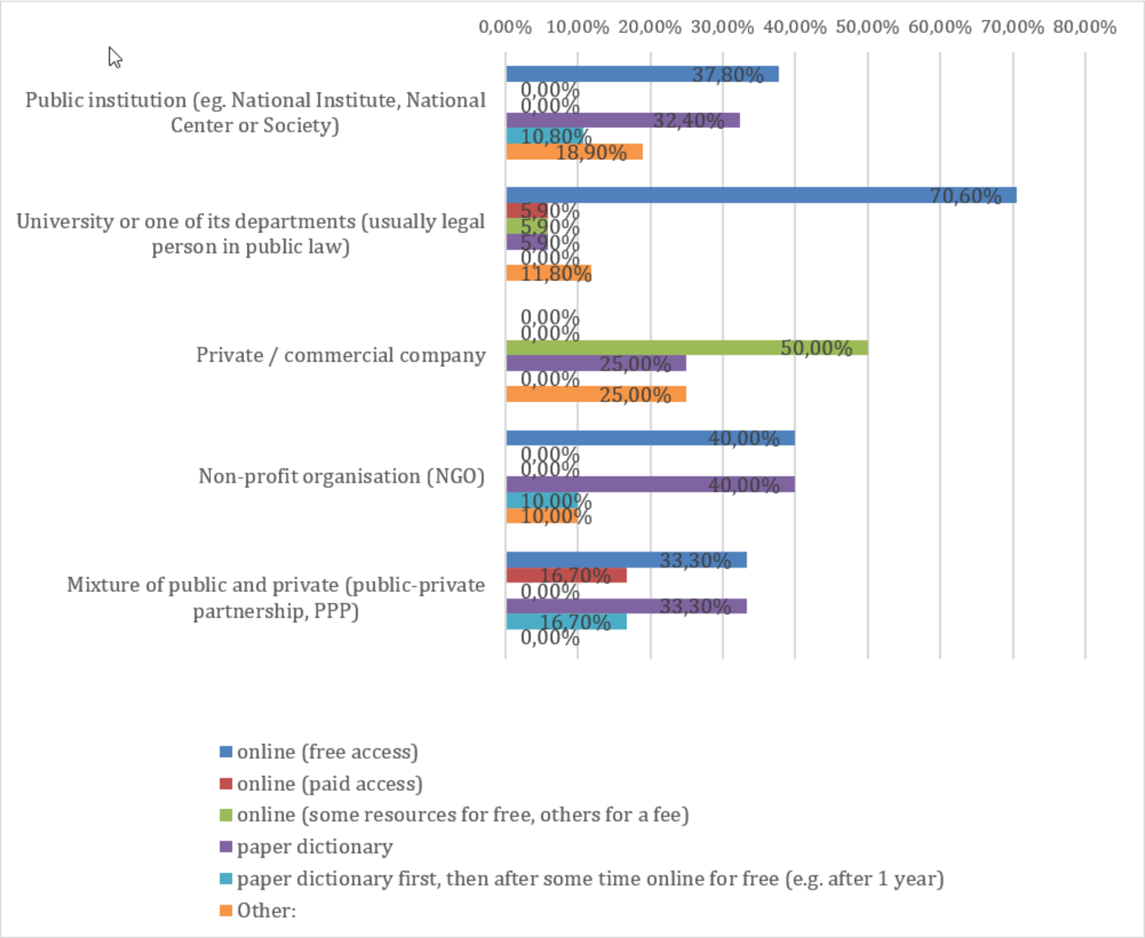


Figure 2: Availability of lexicographic resources by type of institution.

In terms of datasets, a high number of institutions reported making their lexicographic data available for reuse by others, with the majority of them (18) offering free download of the data. Only two institutions reported charging for download of their data. Creative Commons (most often CC-BY or CC-BY-SA) was used as a standard licensing schema for lexicographic data by the majority of institutions (86%; N=22). CLARIN licensing framework and Open Data Commons were used by only a few institutions, five and one respectively. A few institutions mentioned that they choose a licensing schema on a case-to-case basis.

17 institutions reported making their data available via API; 13 institutions were offering free API access and 4 institutions paid API access. A few institutions mentioned they were in the process of setting up API access. It is important to point out that when explaining their answers several institutions reported making only certain parts of their lexicographic resources available to others (e.g. headword lists, lists of typical misspellings), and/or introducing usage limits to number of requests or amounts of data. Paid API access is thus used as an additional service to free access, for example for substantial usage, or for using lexicographic data for commercial purposes.

Customized services seem to be used often, with 13 institutions reporting they offer them. The customers are researchers or companies, and individual cases mentioned ranged from preparing lemma/headword lists with selected information from entries, lists of commonly misspelled words, audio files and images etc.

Only seven institutions reported making their lexicographic data available through brokers; two used ELRA, one META-SHARE, and one ELRC-share. Three institutions reported using a CLARIN repository.

15 out of 31 institutions reported keeping track of the use of their datasets, the other 16 reported they do not. The use of datasets is monitored in one of three ways: by requiring users to register, by asking users to report on how they used the data, or by monitoring the API use.

3.3.2 Willingness to share different types of data

When asked about different types of lexicographic data, most institutions reported to be willing to share, under different licenses, lemma lists (28 institutions). Many institutions would also be willing to share examples (23), synonyms (22), sense structure (22), morphological information (22), definitions (21), collocations (20), fixed expressions (19), frequency information (19), and syntactic information (18). Fewer institutions reported willingness to share etymological information (14), pronunciation information (12), and frequently misspelled word forms of lemmas (9). It must be noted that lower numbers of institutions at certain types of data are also linked to the fact that some institutions do not have such

types of data. Nonetheless, the aforementioned three types of lexicographic data with the lowest number of institutions being willing to share them also exhibit the highest percentages of institutions selecting the "would not share" option.
As shown in Figure 3, the most frequently selected license for nearly all the data types was public (open) data, followed by restricted (non-commercial) license. Taking the ratio between these two licenses into account, the institutions seem to be more protective of frequently misspelled word forms of lemmas, definitions, synonyms, and collocations. Several institutions commented on the problematic or unclear status of corpus examples. Academic license and commercial license were selected by a significantly smaller portion of the institutions, even smaller than the portion under the "would not share" option. Some institutions are willing to share (some) types of their data with other organisations only by using specially prepared contracts between institutions.
It is interesting to note that if non-applicable answers are excluded (as they mean that such types of data are not made or available at the institutions), there were 13 institutions that reported offering all their available types of data under public (open) access.

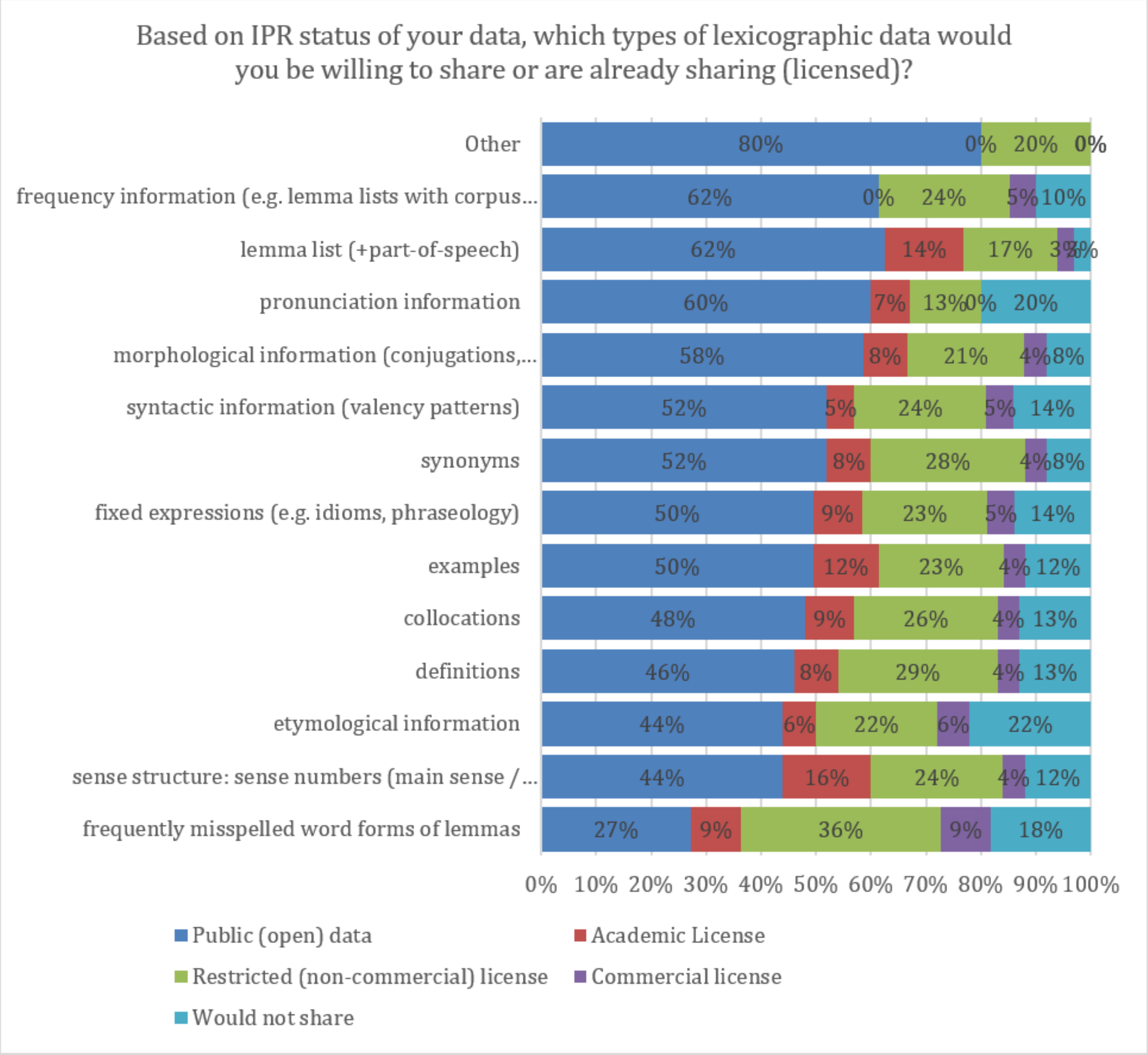


Figure 3: Sharing different types of lexicographic data.

3.3.3 Concerns about data sharing

The main concerns institutions have about sharing their data can be divided into three groups. Firstly, many institutions expressed concerns about how the data might be used by others, in particular they pointed to:

- Commercial use of their data, especially by competitors. The concerns are especially connected with producing low quality products for profit generation only.
- Misuse by others, e.g. use beyond the purposes allowed by the license. Also, misuse may result in breach of contract with data providers, e.g. when making a corpus.
- Fear of someone beating them to analysis or source preparation.

Then, there were concerns about the unclear status of their data because they were obtained from corpora with licensing restrictions. Finally, some institutions pointed to the lack of standardized documentation for sharing lexicographic data.

Despite various concerns about data sharing, the majority of the responding institutions (90%) had never taken legal action related to the use of their lexicographic data, indicating that this seems to be rare in the lexicographic world, especially as far as publicly funded institutions are concerned. One institution reported on reaching a settlement (in court) related to some of their dictionaries, and the other on the fact that they conduct surveillance of possible bad practices or illegal use and take action when necessary. The third institution reported on a case of forensic linguistic analysis of various bilingual dictionaries, which had been conducted to determine their originality, i.e. to assess the possibility of theft of intellectual property.

3.4 Case study - DSL: How to approach licensing of lexicographic content

DSL (the Society for Danish Language and Literature) is one of the two partners in ELEXIS that reported to be a mixture of public and private organisation. When the ELEXIS project was initiated in 2018, DSL did not have cleared IPR status for their lexicographic data. The society has edited and published monolingual Danish dictionaries since 1911, and several of these projects have throughout the years been funded partly by the Danish Ministry of Culture, and partly by private funding, i.e. the Carlsberg Foundation. Since 2009, DSL has published the Danish dictionary DDO (Den Danske Ordbog) freely online, 4 years after the last volume of the printed DDO dictionary was released for sale. Today, the DDO dictionary, which is well known in the Danish society, is also freely available via API, but only for non-commercial purposes and under special (time limited) agreement.

Over the last decade, there have been an increasing number of inquiries regarding the spin-off data from the dictionary compilation process, e.g. lemma lists and frequency lists. These data were often given out for free for non-profit purposes, i.e. research or study projects, but always as stand-alone resources, that is without internal links between different types of data. No specific person at DSL was responsible for dealing with copyright issues concerning the digital lexicographic data. A legal advisor is affiliated with the society, however before the focus had been on copyright issues regarding publications in print. Moreover, the personal views of the DDO dictionary editors represent a variety of different opinions from being rather open, e.g. wanting to share any type of lexicographic data for research purposes, especially to partners in a research project, to the completely opposite position, i.e. wanting to keep the data in-house in order to guarantee the future lexicographic business of DSL. One major reason for not wanting to share the data is fear of abuse, e.g. use beyond the purposes allowed by the license, and that misuse, e.g. of corpus citations, may result in the breach of contracts with the data providers. DSL has quite restricted agreements with several key text providers which have been delivering texts for more than 25 years - texts that are typically only allowed to be used for internal corpus investigations at DSL and as citations in the DDO dictionary. However, it was unclear whether the providers would really have the right to protest if the citations in the dictionary, many of them dating back to the 1990's, were used for other purposes, e.g. for research.

In 2018, at the same time as the ELEXIS project started, the Danish Ministry of Culture set up a language technology committee, which included a DSL representative. The purpose was to clarify the major problems preventing Danish language technologies from being developed in line with the English language technology industry. The Danish datasets and lexical resources that existed at the time were described in a concluding report in 2019, including under which conditions they could be of benefit to the private business community. One of the main conclusions of the report was that a large, open source resource, which would integrate the existing lexical data from major Danish language institutions, was very much needed in order to facilitate the development of language technology products for Danish (Kirchmeier et al. 2020).

Combined with the participation in the survey on licensing practices of ELEXIS institutions described in this paper where many questions could not be answered clearly, it became clear that DSL needed a detailed policy regarding the sharing of more complex and integrated lexicographic resources, both to be able to fulfil the data requirements in the ELEXIS project and to be able to play a role in the development of an open source consolidated resource for Danish. Moreover, with such a policy, DSL would be able to answer the growing number of requests for the computational lexicographic data developed at DSL in a more homogeneous and standardised way.

3.4.1 Applying the ELEXIS survey for internal purposes

In order to produce a more specific policy in line with the opinion of the DDO editors, we carried out an internal survey. We reused most of the questions from the ELEXIS survey reported in this paper. The editors were asked individually about their views on the sharing of different types of data in the DDO dictionary. While answering the questions, they were allowed to consult the general results from the ELEXIS survey in Figure 3. They were for example able to see that two thirds of the lexicographic institutions involved in ELEXIS were willing to share lemma as well as frequency lists as open source data, and that most of the institutions also considered it acceptable to share synonyms, valency information and morphological information. The hypothesis was that the insight into which type of data most lexicographic institutions were in fact willing to share would probably influence the views of the editors.

The results of the internal survey showed that the views of the editors are, to a much higher degree than expected, in line with the views of the majority of the ELEXIS partner and observer institutions; in fact, in some cases the DDO editors were willing to share even more data, e.g. data on misspellings and sense structure. However, definitions and etymologies were considered by the DDO editors to be authored data, which cannot be shared, except for research purposes in projects where DSL is a partner, or against a substantial payment. The editors were also willing to share the manually selected citation examples in DDO, but as already mentioned this is a more complicated matter due to copyright issues.

The results of the internal survey led to several actions at DSL. Firstly, the legal consultant was involved in order to clarify the more complicated case of citation examples. He guaranteed that it would be legal to hand over the example citations to a third party as long as they remained integrated in other lexical data (e.g. in a sense structure); this, however, would not hold true if examples were extracted separately, for example for the purpose of creating a stand-alone text corpus. Secondly, a number of lexicographic datasets, which were formerly given for free by DSL only if the specific (non-commercial) purpose stated by the user was acceptable, were made directly available for download online at korpus.dsl.dk. Today, the only restriction on the data is that it must not be used to publish an independent dictionary that is in competition with DSL's products. The interested parties must accept this condition before making the download. Furthermore, the website for downloads has been improved, and the number of open source lexical resources that can be directly downloaded has increased, now also including lists of synonyms and common misspellings. Also, a new strategy on customized services was confirmed by the DSL board: namely, DSL decided to contribute to the development of technologies for the Danish language by allowing users (including companies) to request special datasets based on DDO as long as they cover DSL's payroll expenses for the preparation of the dataset, and under the condition that the dataset is afterwards made freely available on the DSL website korpus.dsl.dk.

Furthermore, the clarification of the editors' standpoints and the knowledge on which data types they are willing to share has led to a new, nationally funded project. In 2020, it significantly facilitated the work on preparing the project application, and in March 2021, a three-year project with the University of Copenhagen, the Danish Language Council and the Danish Agency for Digitisation was approved. DSL contribution lies in a high amount of already linked lexicographic data consisting of lemmas, morphology, misspellings, synonyms, sense structure, and examples from DDO, as well as semantic links to the Danish WordNet which were developed from (and at the same time linked to) DDO data in the DanNet project 2004-2010 (Pedersen et al. 2009).

Last, but not least, the clarification of the editors' views and the general DSL policy also facilitated DSL's contribution to the collection of lexicographic data in the ELEXIS project. Accordingly, the DDO data that DSL submitted consists of linked lexicographic information, while authored text, e.g. full definitions, was not included. Still, a sample of approximately 5,000 lemmas with full sense descriptions has been provided for research purposes.

4 Conclusions

As the survey among ELEXIS partners and observers has shown, there are many differences across different countries, without any clear patterns of, for example, similar practices by similar types of institutions. While there are many institutions that promote open access for all, or nearly all, of their data, there are on the other hand still several institutions that are very protective of their data. Interestingly, there seem to be different levels of concern for different types of lexicographic data, with definitions, examples, synonyms, collocations, and frequently misspelled forms of lemmas being the most protected.

The majority of the institutions reported using Creative Commons licensing schema, while few others reported (also) using the CLARIN licensing framework or Open Data Commons. In addition, just under 50% of the institutions reported keeping track of their datasets, and 50% of the institutions reported having a special person dealing with intellectual property rights issues. Given the concerns reported and the importance of licensing mentioned by the institutions, this percentage can be considered quite low.

The case study of DSL clearly shows that raising awareness on the benefits of opening the data, and sharing experiences and opinions on data sharing can lead to important changes in the approach to licensing lexicographic content. We can report that other institutions are adopting a similar approach, opening their datasets more as compared to their status in the ELEXIS agreement. To date, 118 different datasets, e.g. general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), and lemma lists have been collected from 32 ELEXIS partner and observer institutions. A sample list of the datasets can be found in the ELEXIS Deliverable 6.3 Intermediate interoperability report (Kosem et al. 2021).

These datasets will be used for linking purposes in creating the Dictionary Matrix. Since some of the data contributors are still very protective of their data, the ELEXIS consortium has come up with a number of flexible and diverse licensing options to encourage the institutions to contribute their data (or parts of it) to the Dictionary Matrix:

1. The owners of background data can decide how much data they want to contribute to the Dictionary Matrix. They have the option to contribute entire dictionaries or only parts of them, e.g. full entries for a certain letter only, or only certain elements for all entries. Minimally, a headword list with part-of-speech information is required to develop the links for the Dictionary Matrix.
2. Owners of a set of dictionary data can choose an (open access) license of their liking.
3. Dictionary content that will be presented to users will be accompanied by the information on the rights that rest on the data, as well as the appropriate attribution if the chosen license for the dictionary requires that.
4. If the owners of dictionary content still hesitate to share their data directly, it will be possible to link from the Dictionary Matrix to external resources. These links, when queried, will direct the users to the data that will be served up at the proprietary websites of the owners.

This licensing scheme is the result of a special task in ELEXIS which was dedicated to providing guidelines and solutions for handling copyright and authorship protection, resulting in a deliverable entitled *Recommendations on legal and IPR issues for lexicography* (Boelhouwer et al. 2020).

The process of making lexical resources more openly available has already started in the lexicographic community, but more promotion and raising awareness is needed. As voiced by one of the respondents of the survey: "A culture of data sharing across institutions is still to come. According to many, now (generational change, end of paper-based publishing) could be the moment for an initiative."

5 References

Boelhouwer, B., Kosem, I., Nimb, S., Jakubíček, M. Tiberius, C., Krek, S. & Rosenmeier, M. (2020). Recommendations on Legal and IPR Issues for Lexicography. Deliverable 6.2 of the European Lexicographic Infrastructure Project. https://elex.is/wp-content/uploads/2020/02/ELEXIS_D6_2_Reccommendations_on_Legal_and_IPR_Issues_for_Lexicography.pdf [15/04/2021]

Kirchmeier, S., Diderichsen, P., Henrichsen, P. J., Nimb, S., Pedersen, B. S. (2020). World Class Language Technology - Developing a Language Technology Strategy for Danish. In *Proceedings of the 12th International Conference on Language Resources and Evaluation*, Marseilles, France.

Kosem, I., Navigli, R., McCrae, J. & Jakubíček, M. (2021). Intermediate Interoperability Report. Deliverable 6.3 of the European Lexicographic Infrastructure Project. https://elex.is/wp-content/uploads/2021/02/ELEXIS_D6_3_Intermediate_interoperability_report.pdf [15/04/2021]

Krek, S., Declerck, T., McCrae, J. P. & Wissik, T. (2019). Towards a Global Lexicographic Infrastructure. Presented at the *Language Technology 4 All* Conference. Zenodo: http://doi.org/10.5281/zenodo.3607274. [05/04/2021

Krek, S., McCrae, J., Kosem, I., Wissik, T., Tiberius, C., Navigli, R. & Pedersen, B. S. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds) *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts (EURALEX 2018)*, Ljubljana, Slovenia, 17-21 July 2018, pp. 881–892. Zenodo. http://doi.org/10.5281/zenodo.2599902. [05/04/2021]

Pedersen, B. S., McCrae, J., Tiberius, C. & Krek, S. (2018). ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In F. Bond, T. Kuribayashi, C. Fellbaum & P. Vossen (eds.) *Proceedings of the 9th Global WordNet Conference (GWC 2018)*, Global Wordnet Association, Singapore, pp. 339-344. Zenodo. http://doi.org/10.5281/zenodo.2599954. [15/04/2021]

Woldrich, A., Goli, T., Kosem, I., Matuška, O. & Wissik, T. (2020). ELEXIS: Technical and social infrastructure for lexicography. In K Lexical News. Available at: https://kln.lexicala.com/kln28/elexis/ [15/04/2021]

Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Verbal multiword expressions: a preliminary study on the fixedness degree, application to Modern Greek and French

Constant M., Fotopoulou A.

Université de Lorraine Nancy France, Institute for Language and Speech Processing Athens Greece
mathieu.constant@univ-lorraine.fr, afotop@athenarc.gr

Abstract
Multiword expressions display multidimensional properties and a varying degree of compositionality. In this paper, we show a preliminary study to systematically characterize multiword expression types using a set of lexical, morphosyntactic and semantic features, in order to identify their fixedness degree. In particular, we built two sample lexical databases of 100 verbal (mainly emotion) multiword expressions for French and for Modern Greek, systematically encoding these features. We then explore the correlation between semantic features and lexical/morphosyntactic features, in order to better understand the link between lexical and morphosyntactic fixedness and semantic compositionality. This pilot study opens an interesting path of lexicographic research that would consist in systematically exploring a larger spectrum of linguistic features and of types of multiword expressions.

Keywords: multiword expressions; degree of fixedness; modelling and encoding MWEs

1 Introduction

Multiword expressions (MWEs) are combinations of several lexical items that display some idiosyncrasy at one or several linguistic levels. They cover a large number of linguistic phenomena (Sag et al. 2001), including nominal and verbal idioms, support verb constructions, phrasal verbs, complex grammatical words, named entities. They have been the focus of a wide body of research both in linguistics and natural language processing for some decades. As an example, we can mention the European research network PARSEME (COST Action 2013 - 2017) that brought together researchers of multiple scientific fields, enabling substantial progress in modelling and processing MWEs.

This paper aims at presenting a preliminary work on the identification of the fixedness degree of verbal multiword expressions. We specifically focus on Modern Greek and French as verbal MWE databases exist with a large coverage of encoded syntactic and semantic properties, that are organized in a lexicon-grammar (M. Gross 1986). Our study is built on the work of M. Gross (1982) for French and the work of Fotopoulou (1993) and Mini (2009) for Modern Greek. Our objective is to build a model representing the different fixedness cases (also for lexicographic purposes), by means of a spectrum of linguistic features. The verbal MWEs in these databases are treated as elementary sentences for which all possible fixed and non-fixed (or variable) arguments (if any) are consistently and uniformly encoded. The MWE structures are represented as part-of-speech sequences. Selectional restrictions over the non-fixed or variable elements of MWEs as well as syntactic phenomena (i.e., clitic and passive alternation, etc) - if any - are also encoded formally. Finally, other grammatical phenomena such as agreement features (in person and number) are accounted for.

The article is structured as follows. First, it provides some background on multiword expressions (section 2). Section 3 describes the linguistic data, the methodology as well as the criteria used for our experiments. Section 4 presents the encoding of the sample lexical databases. And finally, we provide a detailed analysis of the obtained results (section 5).

2 Background

Linguistic studies in the literature often characterize MWEs by their fixedness and non-compositionality. There are these multiword expressions whose idiomatic meaning cannot be deduced from the meaning of their parts (Fraser 1970; Bobrow & Bell. 1973; Swinney & Cutler 1979; Chomsky 1980; Gross 1982; Van der Linden 1992). In this case, one cannot derive the idiomatic meaning of the idiom *bite the dust* (i.e. cease to exist) based on the meanings of the words *bite*, *the*, *dust*. More precisely, there are three criteria that can help identify them. These criteria define the “fixedness” of these expressions: (a) the semantic criterion: the meaning of an expression is holistic or non-compositional. This means that its meaning cannot be derived from the regular combination of its constituents; (2) the lexical criterion: at least one of the expression’s constituents does not have (or nearly so) paradigmatic variation; (3) the morpho-syntactic criterion: at least one of the constituents cannot be treated as it would be in free sentences, because constraints for example on the determiner or on the transformations, e.g. in passive (Lamiroy 2003). These criteria, however, do not always apply uniformly, and the observed variability leads to the notion ‘degree of fixedness’ (Gross 1996).

Nunberg (1979) introduced the notion of idioms as combining expressions and, in these terms, semantic compositionality refers to idiom elements that “carry identifiable parts of the idiomatic meaning” (Nunberg et al. 1994: 496). Based on this notion, while distinguishing semantically decomposable idioms (*spill the beans*) from non-decomposable ones (*kick the* *bucket*), he supports that idioms vary in function of their semantic decomposability. What differentiates them is the interaction between the literal and figurative meanings of their parts. In the example *pop the question*, there is a clear association between *pop* and *question,* and their corresponding parts of the figurative meaning *'propose marriage*' (normally decomposable idioms), whereas in the MWE *meet your maker*, the word *maker* makes a metaphorical reference to a deity (abnormally decomposable idioms). It is, however, to be noted that an expression may be considered fixed or less fixed depending on the theoretical framework adopted and the criteria used.

From a more computational linguistics perspective, the degree of compositionality of MWEs is often characterized by numerical scores: for instance, using scales from 0 to *n* (McCarthy et al. 2003; Reddy et al. 2011; Roller et al. 2013; Ramisch et al. 2016) with several individual judgements by MWE type, or based on binary judgments (Farahmand et al. 2015). The encoding can also be categorial like in (Gurrutxaga & Alegria 2013) where the expressions are classified in three categories (idiomatic, collocational and free combination).

Concerning the European research network PARSEME (COST Action 2013 - 2017), various surveys have been produced on either their representation, their grammatical modeling, and their processing (Sailer & Markantonatou 2018; Parmentier & Waszczuk 2019). One of the greatest outcomes is a multilingual corpus annotated in verbal MWEs for 20 languages relying on unique guidelines based on decision diagrams integrating precise linguistic tests (Savary et al. 2017; Ramisch et al. 2018).

3 Linguistic data - methodology - criteria

For our study, we constructed a small set of 65 verbal multiword expressions for both languages. This list of modern Greek and French MWEs that mainly pertain to the semantic field of *emotions* was manually compiled from data listed in existing lexicon-grammar tables for Greek and French. The extracted expressions have different syntactic structures in order to test the fixedness degree with respect to a series of tests. These tests are those which, in principle, enable to define a fixed expression like more and more studies show such as in M. Gross (1982), G. Gross (1996), Lamiroy (2003), Vincze (2011), Sailer & Wintner (2014), Stone (2015):

1. **Lexical criteria:** Fixedness can be identified by testing whether there exists a paradigmatic break on each lexical elements, e.g. while having *Max casse (le jouet+le verre) à Marie* with the compositional meaning *Max breaks Marie's (toy+glass*) (lit. *Max breaks the (toy+glass) to Marie*}, we get *Max a cassé (les pieds +*le jouet+*le verre) à Marie* with the meaning *'Max gets on Marie's nerves'* (lit. *Max broke the (feet +*toy+*glass) to Marie*); These criteria enable the evaluation of the exclusive co-occurrence of the expression components.
2. **Morphosyntactic criteria**: non-regular restrictions apply on the determiner distribution or on the morphological variants (e.g. number), as well as on some transformations like passivation or pronominalization. For instance, the verbal expression *Μου κόπηκαν* ***τα*** *ήπατα* (lit. *I have the livers cut)* with the meaning *΄I was very frightened*' do not allow any modification over the fixed constituents: **Μου κόπηκε* ***το*** *ήπαρ* (lit. *I have the liver cut*).
3. **Semantic criteria**: traditionally, the main criterion is the following: the meaning of the expression is non-compositional, i.e. it is not predictable from the meaning of its components. But generally speaking, the meaning of a verbal expression can emerge from different combinatorics. Consequently, to detect whether an expression is semantically compositional or not, we have examined independently (and in correlation) the verb and the nominal arguments of the expression similarly to Mini et al. (2011), with tests like:

i. the element keeps its literal meaning,
ii. the element has a metaphorical meaning or is an extension of the literal meaning.
iii. the element or the meaning of the whole sequence has nothing to do with the literal meaning

Mini et al. (2011) examined in each expression the type of relationship between the verb and the nominal arguments. They studied whether the configuration verb + nominal complements in a given expression is unique and specific to derive the global meaning, or the given elements, on the contrary, always keep the same meaning by combining with other elements. In the example *Luc nage dans le bonheur* [FR] (lit. *Luc swims in happiness*), meaning '*Luc is happy*' the word *nage* (swims) is not related to its literal meaning but it is a metaphorical usage (case ii) and the word *bonheur* (happiness) has its literal meaning (case i). On the other hand, *lui casser les pieds* [FR] (lit. h*im (Ppv) to-break the feet*), meaning '*get on his nerves*' the words *casser* (break) and *pieds* (feet) are not related to their literal meaning (case iii.). On the contrary, in the example μου *ράγισε την καρδιά* [GR] (lit. *he broke my heart*) the verb retains the core meaning of '*cracking without being cut into pieces*' that it has in the sentence *ράγισε το ποτήρι* ("the glass cracked") by a semantic extension to a non-tangible/abstract level (case ii). On the other hand, the noun *καρδιά* ("heart") is a semantically autonomous constituent of the sentence since it refers to the inner emotional world *η καρδιά μου 'my heart cracked*' (case ii). In example *Marie monte sur ses grands chevaux* [FR] (lit. *Marie (climbs/rises/gets) on her big horses)* the meaning of the verb *monter* and its complement *ses grands chevaux* has nothing to do with the meaning *Marie gets on her big/high horse (*case iii).[1]

We also used standard tests to identify support verb constructions (Gross 1988; Fotopoulou 1992). They are specific multiword expressions as the nominal element keeps its literal meaning, while the verb has a neutral value. We examined three criteria: reduction in a noun phrase with deletion of the support verb, nominalization of the expression *prendre une décision* (lit. *take a decision*, with the meaning *make a decision*) = *prise de décision* (meaning *decision making*), unique relation of the nominal element to the subject. We considered that encoding such tests would contribute to modeling the fixedness degree.

4 Encoding

The expressions are encoded in a table: a row corresponds to a lexical entry (a multiword expression), a column corresponds to a lexical, syntactic or semantic feature like in M. Gross (1986). More precisely, the features encoded are:

- the source lexicon-grammar table in French and modern Greek from which each individual MWE has been extracted,
- the lexical values of the components of the MWEs at every syntactic position,
- valency properties of the verbal expressions, including its arguments and their distributional properties; for instance, the French MWE *casser les pieds* (lit. *break the feet*, '*get on one's nerve*') has two arguments: a non-restricted subject and a human complement introduced by the preposition *à* (to),
- properties on the lexicalisation of the MWE components: non-substitution by a semantic neighbour (i.e. paradigmatic break), non-deletion,
- Syntactic study of the whole sentence: pronominalisation, passive/ergative transformation, support-verb construction,
- Semantic study of the individual lexical elements of the expressions: literal, metaphoric, or meaning extension,
- Global meaning of the sentence: non-compositional- metaphoric/extension

A sample of these tables are provided in Table 1 for French and in Table 2 for Modern Greek. The description of feature labels and their possible values are defined in table 3. Note that, for clarity, the features given in this table constitute a subset of all encoded properties:

- the list of lexical components at various syntactic positions: V (verb), P (preposition), D (determiner), A (adjective), N (noun),
- some lexical, morphosyntactic and semantic features regarding the lexical MWE components (cf. table 3) ,
- the glosses and translations in English of the MWEs

We do not show, for instance, the selected prepositions, the list of lexical elements (if any) that can substitute the MWE elements when there is no paradigmatic break using the lexical criteria, and other syntactic properties (ex. pronominalization).

MWE					LEX		MORPHOSYNT				SEM			GLOSS	translation
V	P	D	A	N	L1	L2	M1	M2	M3	M4	S1	S2	S3		
essuyer		une		insulte	+	+	+	+	+	+	1	1	1	endure an insult	receive an insult
trembler	de			peur	+	+	-	+	+	/	0	1	0	shake of fear	be terrified
remuer		l'		âme	+	+	+	+	+	+	0	1	0	stir the soul	touch
briser		le		coeur	-	+	+	+	-	+	0	0	0	breat the heart	break one's heart

[1] According Mini (2009) and Mini et al. (2011), the combinatorics of the above criteria led to two general groups: non compositional/typical (*τα φόρτωσε στο κόκορα* (literally *He loaded them on the rooster*) meaning '*I did not act at all, as I was feeling lazy*' and compositional/non typical expressions that can be divided into quasi-typical expressions such as *ράγισε η καρδιά μου*, (literally *my heart cracks*) meaning '*I am in deep grief*' or '*it broke my heart*' and the conventional expressions such as *μου άνοιξαν νέοι ορίζοντες* (literally *they opened new horizons to me).*

monter	à	la		tête	-	+	+	-	-	/	-1	0	0	climb to the head	got to one's head
monter	sur	Poss	grands	chevaux	-	-	+	-	-	/	-1	-1	-1	climb on Poss big horses	get on Poss high horses
casser		les		pieds	-	-	-	-	-	-	-1	-1	-1	break the feet	get on one's nerves

Table 1: Sample of the French MWE table.

MWE					LEX		MORPHOSYNT				SEM			GLOSS	translation
V	P	D	A	N	L1	L2	M1	M2	M3	M4	S1	S2	S3		
ανοίγω				δρόμο (για)	-	-	+	-	-	-	1	0	0	The road is open for...	There are no obstacles left for
χάνω		τα		λόγια Poss-0	+	-	-	-	-	+	0	1	0	N loses his words	be speechless (with emotion)
πάγωσε		το		αίμα Poss-0	-	-	+	-	-	/	0	0	0	My blood freezes	I was terrified
άναψαν		τα		λαμπάκια Poss-0	-	-	-	-	-	/	-1	-1	-1	They lit the lamps	I was very angry
δεν μου καίγεται				καρφί	-	-	-	-	-	/	-1	-1	-1	Not care a scrap	'I couldn't care less'

Table 2: Sample of the Greek MWE table

CAT	Label	Description	Value
LEX	L1	possible substitution of the noun	true (+) or false (-)
	L2	possible substitution of the verb	true (+) or false (-)
MORPHO-SYN	M1	the determiner is flexible	true (+), false (-) or a class label (ex: Poss = possessive determiner)
	M2	possible ellipsis on the verb	true (+) or false (-)
	M3	possible ellipsis on the noun	true (+) or false (-)
	M4	possible passivation	true (+), false (-) or N/A (/)
SEM	S1	semantic meaning of the verb	literal (1), metaphorical (0) or no relationship (-1)
	S2	semantic meaning of the noun	literal (1), metaphorical (0) or no relationship (-1)
	S3	semantic meaning of the MWE	literal (1), metaphorical (0) or no relationship (-1)

Table 3: Sample of feature label descriptions

5 Results and discussions

Results show a large variety of behaviors. In general, we can observe a correlation between the different types of criteria in the extreme cases of the continuum between entirely non-compositional and almost free expressions. We also observe a grey zone with some unexpected behaviors. A traditional assumption considers that the more an expression does not accept lexical substitutions, the more it does not accept syntactic transformations like passivation, the more it has a non-compositional meaning. However, in some cases, this is not true. For instance, in French, *casser du sucre sur le dos* (lit. *break some sugar on the back*), meaning '*talk about someone behind her/his back'* accepts passivation, while it does not accept any substitution for its lexical components.

In order to better visualize the correlation between lexical/morphosyntactic features and semantic ones, we decided to provide a numerical score for each of the two types of features. To do so, we first replaced each non-numeric value in the tables by numeric ones using the following rule: a + symbol is assigned a +1 value, a - symbol is assigned a -1 value, other symbols are given a 0 value, except when it is not appropriate (/ symbol in the tables). Note that for the semantic features, a given lexical element of a given expression is associated with the value 1 when it has its literal meaning, 0 when it has its metaphoric meaning, -1 when it has no relation with the literal meaning. The global score for a type of feature for each individual MWE is the average of the scores of all encoded features of this type.

For instance, the expression *briser le coeur* in Table 1 encodes two - values (-1) and four + values (+1) regarding lexical/morphosyntactic features. Therefore, by averaging the corresponding numerical values, it reaches a score of 0.33 (2/6). Regarding semantic features, the average of the corresponding values (i.e. 0 for the three semantic tests) is 0. Thus, the expression *briser le coeur* is associated with the pair of values (0.33,0). It is the same for Greek expressions like *πάγωσε το αίμα μου* (litt. *My blood freezes*) '*I was terrified*'. In this case, lexical and morphosyntactic features are encoded with four - values and one + value, leading to the score -0.6 (-3/5). The semantic encoding includes three 0 values, leading to a 0 score. The expression then corresponds to the pair (-0.6,0).

Such an approach means that we consider that all features have the same weight. Thereafter, we consider that the global score for the lexical/morphosyntactic features corresponds to the degree of lexical/morphosyntactic fixedness, and the global score for the semantic features corresponds to the degree of semantic non-compositionality. For the two examples above, *briser le coeur* with a positive value (0.33) tends to be flexible from a lexical and morphosyntactic point of view. The greek expression *πάγωσε το αίμα μου* with a negative value (-0.6) tends to show lexical and morphological fixedness.

Results on French are given in the bubble chart in Figure 1. The horizontal axis (resp. vertical axis) corresponds to the global score of the lexical/morphosyntactic features (resp. semantic features). The bubble size depends on the number of multiword expression entries having the corresponding pair of scores: the more numerous the bigger.

The score pairs corresponding to entirely frozen and non-compositional expressions are positioned in the lower-left corner such as *casser les pieds* with only - values (-1) for lexical/morphosyntactic tests and -1 values for semantic tests. It is the same for Greek expressions like *δεν μου καίγεται καρφί* (lit. *care a scrap*) '*I couldn't care les*s', which only encodes - values both for semantic and lexical/morphosyntactic features, showing it is semantically non-compositional. The score pairs corresponding to almost free and compositional expressions are positioned in the upper-right corner such as *essuyer une insulte 'receive an insult'*. The results seem to confirm the existence of a grey zone in between the extreme cases, as one can observe a continuum between the extreme cases, i.e. the part of the graph in between the extreme cases is fully filled with bubbles.

Furthermore, the chart overall shows a rough linear correlation between the degree of lexical/morphosyntactic fixedness and the degree of semantic non-compositionality. Bubbles tend to be in the lower-left and upper-right parts of the chart, tending to show that the more the expression is lexically and morpho-syntactically flexible the more it is semantically compositional.

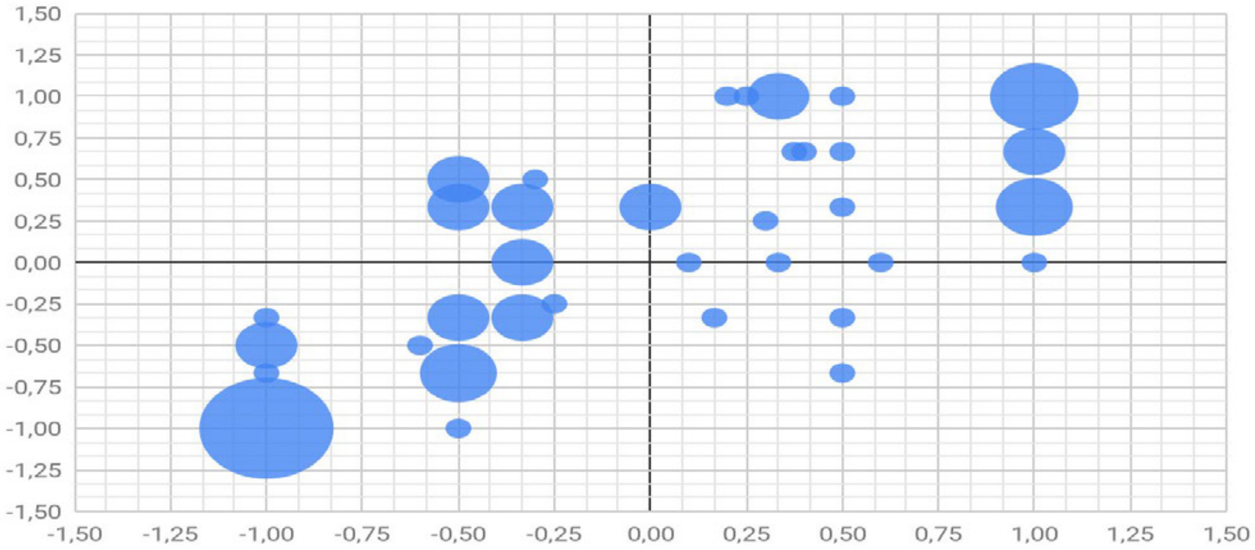


Figure 1: Bubble chart showing the correlation between lexical/morphosyntactic features and semantic ones for French. The horizontal axis (resp. the vertical -axis) corresponds to the global score of the lexical/morphosyntactic features (resp. semantic features). The bubble size depends on the number of multiword expression entries having a given pair of scores.

6 Conclusion

This paper presented a methodology to encode the fixedness degree of verbal multiword expressions that is characterized by a set of lexical, morphosyntactic and semantic features. This pilot study with a specific focus on two languages and a limited set of multiword expressions has shown that formal lexical and morphosyntactic properties tend to approximate semantic compositionality degree though this correlation is somewhat unclear in a 'grey zone'.

Future work should consist in extending this encoding to a larger set of multiword expressions for both languages. This methodology can be applied to other languages, in the same way as it has been applied in the lexicon-grammar methodology for multiple languages. We also plan to compare such an approach with statistical and distributional methods.

7 References

Bobrow, A., & Bell, S.S.M. (1973). On catching on to idiomatic expressions. *Memory and Cognition*,1, pp. 343-346.

Chomsky, N. (1980). Rules and Representations. Columbia Classics in Philosophy. Columbia University Press.

Fotopoulou, A. (1992). Dictionnaires électroniques des phrases figées. Traitement d'un cas particulier: phrases figées/phrases à *Vsup*. In *COMPLEX. Papers in Computational Lexicography.* Hungarian Academy of Sciences, pp. 147–161.

Fotopoulou, A. (1993). *Une classification des phrases à compléments figés en grec moderne - étude morphosyntaxique des phrases figées*. Ph.D. thesis, Université Paris VIII - St. Denis.

Fraser, B. (1970). Idioms within a Transformational Grammar. *Foundations of language*, 6(1), pp. 22–42.

Farahmand, M., Smith, A. & Nivre, J. (2015). A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pp. 29–33

Gross, M. (1982). Une classification des phrases "figées" du français. *Revue Québécoise de Linguistique*, 11(2), pp. 151–185.

Gross, M. (1986). Lexicon-Grammar. The Representation of Compound Words. In *Proceedings of the 11th conference on Computational linguistics* (COLING '86).

Gross, M. (1988). Les limites de la phrase figée. *Langages*, 90, pp. 7–22.

Gross, G. (1996). *Les expressions figées en français. Noms composés et autres locutions*. Ophrys.

Gurrutxaga, A. & Alegria, I. (2013). Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions,* pp. 116–125, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Lamiroy, B. (2003). Les notions linguistiques de figement et de contrainte. *Lingvisticae Investigationes*, 26(1), pp. 53–66.

McCarthy, D., Keller, B. & Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL* 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 18, pp. 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mini, M., Diakogiorgi, K. & Fotopoulou, A. (2011). What can children tell us about idiomatic phrases' fixedness: the psycholinguistic relevance of a linguistic model". *DISCOURS 9 (Revue de linguistique, psycholinguistique et informatique).*

Mini, M. (2009). *Linguistic and Psycholinguistic Study of Fixed Verbal Expressions with Fixed Subject in Modern Greek: A Morphosyntactic Analysis, Lexicosemantic Gradation and Processing by Elementary School Children.* Unpublished doctoral dissertation. Ph.D. thesis, University of Patras.

Nunberg, G., Sag, I. A. & Wasow T. (1994). Idioms. In Stephen Everson, editor, *Language,* pp. 491–538. Cambridge University Press.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3. pp. 43–184.

Parmentier, Y., & Waszczuk, J. (2019). Representation and Parsing of Multiword Expressions: Current trends. Language Science Press, Phraseology and Multiword Expressions.

Ramisch, C., Cordeiro, S.R., Zilio, L., Idiart, M., Villavicencio, A., Wilkens, R. (2016). "How Naked is the Naked Truth? A Multilingual Lexicon of Nominal Compound Compositionality", *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany.

Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Nurrieta, U., Kovalevskait ̇e, J., Krek, S., Lichte, T., Liebeskind, Ch., Monti, J., Parra Es-cart ́ın, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A. & Walsh. A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multi-word Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG 2018)*, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Reddy, S., McCarthy, D., & Manandhar. S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Roller, S., Schulte S. im Walde. & Scheible, S. (2013). The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pp 32–41, Atl Santa, Georgia, USA, June. Association for Computational Linguistics.

Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2001). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing-2002, pp. 1–15).

Sailer, M. & Wintner S. (2014). Multiword expressions: linguistic properties and lexical representation. *Slides from the PARSEME Prague Training School*.

Sailer, M. & Markantonatou S. (2018). Multiword expressions: insights from a multi-lingual perspective. *Phraseology and Multiword Expressions series*. Language Science Press.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshopon Multiword Expressions (MWE 2017)*, pp. 31–47, Valencia, Spain, April. Association for Computational Linguistics.

Stone, M. S. (2015). Systematic flexibility in Verb-Object Idioms. In *Proceedings of the 8th Brussels Conference on Generative Linguistics*.

Swinney, D. & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18. pp. 523–534.

Van der Linden, E.J. (1992). Incremental processing and the hierarchical lexicon. *Computational Linguistics*,18, pp. 219–238.

Vincze, V. (2011). *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. Ph.D. thesis, University of Szeged, August.

Acknowledgments

This work has been partially funded by the French Agence Nationale pour la Recherche, through the PARSEME-FR project (ANR-14-CERA-0001).

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Papers

Lexicography and Semantic Theory

Building a Paralympic, Frame-based Dictionary – Towards an Inclusive Design for *Dicionário Paraolímpico* (Unisinos/Brazil)

Chishman R.[1], da Silva B.[1,2], Nardes dos Santos A.[1], Vianna A.L.T.[1], de Oliveira S.[1], Martins M.L.[1], de Schryver G-M.[2];[3]

[1] Applied Linguistics Graduation Program, Unisinos University, São Leopoldo, Brazil
[2] Department of Languages and Cultures, Ghent University, Ghent, Belgium
[3] Department of African Languages, University of Pretoria, Pretoria, South Africa

rove@unisinos.br, bruna.dasilva@UGent.be, aline.nardes@gmail.com, alvianna@edu.unisinos.br, sandra_san05@hotmail.com, mikaelalm@edu.unisinos.br, gillesmaurice.deschryver@UGent.be

Abstract

This paper presents some theoretical and methodological issues emanating from the building of *Dicionário Paraolímpico* (Paralympic Dictionary), an online lexicographical resource that will describe the lexicon of Paralympic sports in Portuguese and English, structured according to the notion of semantic frame. It follows the lead of previous works published by the SemanTec research group (Unisinos/Brazil), such as *Dicionário Olímpico* (Olympic Dictionary, 2016). For the current project, some features from the previous works were kept, such as the basic microstructure of scenarios and the megastructure. There are, however, significant changes to be introduced in the Paralympic Dictionary. Some of them are the result of the Olympic Dictionary's revision, and address issues such as content multiplicity of the sport and scenario definitions, and the absence of relevant information in the microstructure of lexical units. In addition, some of the changes concern the features that distinguish the Paralympic Dictionary from the Olympic Dictionary, since Paralympic sports have specific frames. Another important issue to be addressed refers to the accessibility of the dictionary itself by people with disabilities. After discussing these issues, the paper concludes by outlining future plans, including further developments for the Paralympic Dictionary and its broader implications in the context of the SemanTec research group.

Keywords: Paralympic Dictionary; Frame Semantics; inclusive lexicography

1 Introduction

Language and culture always went hand in hand within the context of Frame Semantics. Since the first versions of this framework, Charles Fillmore (1976: 26) has stated that the meaning of a word can be described in terms of the activation of a frame, that is, "an associated cognitive schema current in the speech community, which this word activates". Moreover, considering the role of dictionaries as cultural institutions that reflect social conceptualizations, the challenging task of compiling a domain-specific dictionary includes the in-depth exploration and description of the specific domain. This is one of the reasons Frame Semantics has contributed to practical lexicography, since a word or lexical unit that evokes a domain-specific context, in a frame-based dictionary, "will be linked with the cognitive structures (or 'frames'), knowledge of which is presupposed for the concepts encoded by the words" (Fillmore & Atkins 1992: 75).

The results of previous studies conducted by the SemanTec (Semantics & Technology) research group have shown the effective convergence between Frame Semantics and lexicography for the purpose of describing the lexicon of sports (Chishman et al. 2015, 2017, 2019). More recently, the group has been working on the compilation of the *Dicionário Paraolímpico* (Paralympic Dictionary), a dictionary of the Paralympic sports that is currently under development. This project has brought new challenges not only related to the specific domain and the resulting particularity of its frames, but also concerning the design of a dictionary that, more than describing Paralympic sports, aims to be accessible to the ones who are part of such domain – e.g., Paralympic athletes, whose special needs have to be considered in order to propose an inclusive design for the dictionary. Considering such challenges, this paper presents some theoretical and methodological issues brought about by the building of the Paralympic Dictionary.

Our paper is structured as follows. Section 2 provides some background to the development of the *Dicionário Olímpico* (Olympic Dictionary), whose main structure will serve as a base for the compilation of the Paralympic Dictionary. Section 3 describes some of the challenges faced by the editors in the ongoing lexicographic description of the Paralympic sports, including domain-specific frames that have emerged from the corpus. Section 4 discusses some issues regarding the accessibility of the dictionary itself. Finally, Section 5 offers some conclusions and plans for future research and development of the Paralympic Dictionary.

2 Frame Semantics and its Lexicographic Application within the Context of the SemanTec Research Group

The SemanTec research group has specialized in compiling frame-based dictionaries of sports since the launching of the Field – Football Expressions Dictionary (http://dicionariofield.com.br), a trilingual resource (in English, Spanish, and Portuguese), and Olympic Dictionary (http://www.dicionarioolimpico.com.br/), a Portuguese-English dictionary of Olympic sports developed within the context of the 2016 Summer Olympic Games. For the current project, the Paralympic Dictionary, some features from the previous works were kept, such as the basic microstructure of scenarios used in Field; and the megastructure employed in the Olympic Dictionary. In this regard, since the Paralympic Games follow the general structure of the Olympic Games, which results in some convergences between the description of Olympic Dictionary's and Paralympic Dictionary's frames and lexical units, the main features of Olympic Dictionary are first presented.

2.1 Background to Olympic Dictionary's Lexicographic Structure

The Olympic Dictionary is a lexicographic resource which describes the lexicon of the 40 Summer Olympic sports. On the Olympic Dictionary homepage (see Figure 1), each sport has a corresponding icon through which users can access the first level of information: the superframe. On this page (see Figure 2), there is a general written description of the sport (called supergloss), a conceptual map, the list of words and scenarios, a corresponding image, and a trivia section. The next level (see Figure 3) corresponds to the description of each frame – called 'scenario' in the dictionary. In addition to a description of the frame (i.e., a gloss), users have access to an image of the respective frame, a list of words corresponding to the frame-evokers, a conceptual map that illustrates frame relations with other scenarios and words, and a list of related scenarios. Finally, on the last level (see Figure 4), information concerning each lexical unit is provided: grammatical category, link to the corresponding frame, synonyms or variants, English translation equivalent, one example (in English), and a list of words that evoke the same frame.

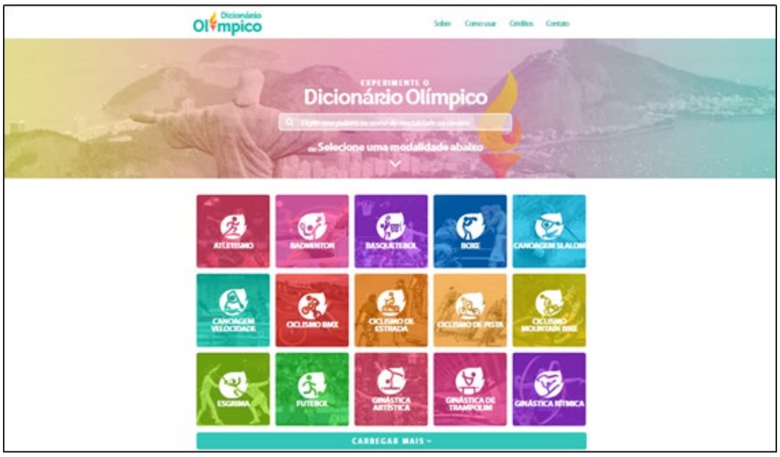
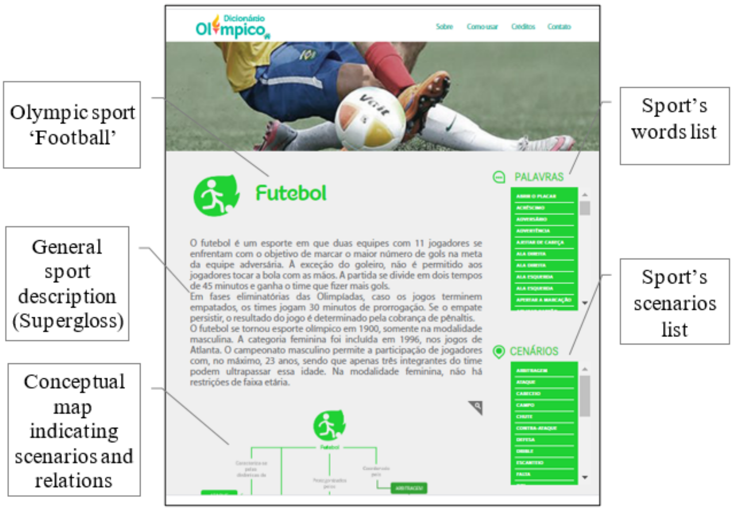


Figure 1: Olympic Dictionary's homepage.

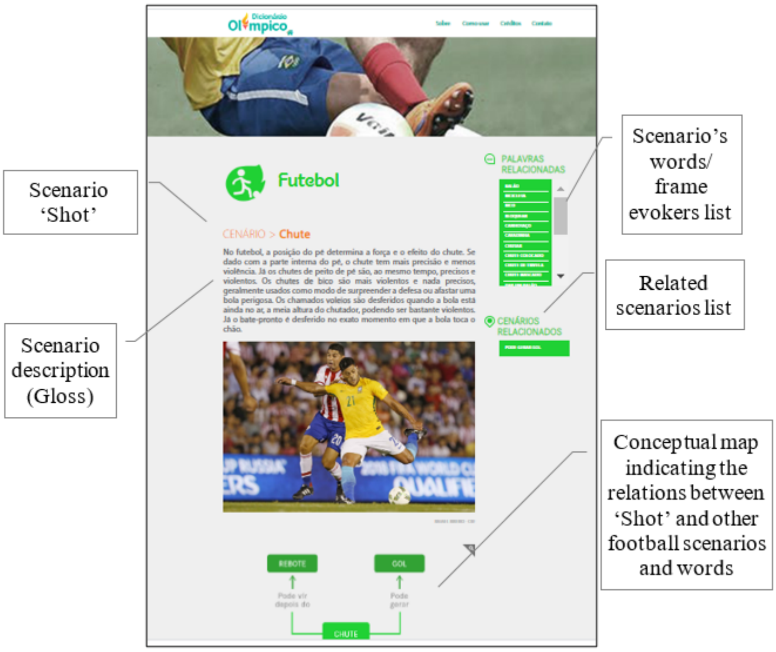


O futebol é um esporte em que duas equipes com 11 jogadores se enfrentam com o objetivo de marcar o maior número de gols na meta da equipe adversária. À exceção do goleiro, não é permitido aos jogadores tocar a bola com as mãos. A partida se divide em dois tempos de 45 minutos e ganha o time que fizer mais gols.

Em fases eliminatórias das Olimpíadas, caso os jogos terminem empatados, os times jogam 30 minutos de prorrogação. Se o empate persistir, o resultado do jogo é determinado pela cobrança de pênaltis.

O futebol se tornou esporte olímpico em 1900, somente na modalidade masculina. A categoria feminina foi incluída em 1996, nos jogos de Atlanta. O campeonato masculino permite a participação de jogadores com, no máximo, 23 anos, sendo que apenas três integrantes do time podem ultrapassar essa idade. Na modalidade feminina, não há restrições de faixa etária.

Figure 2: Superframe/Sport level.

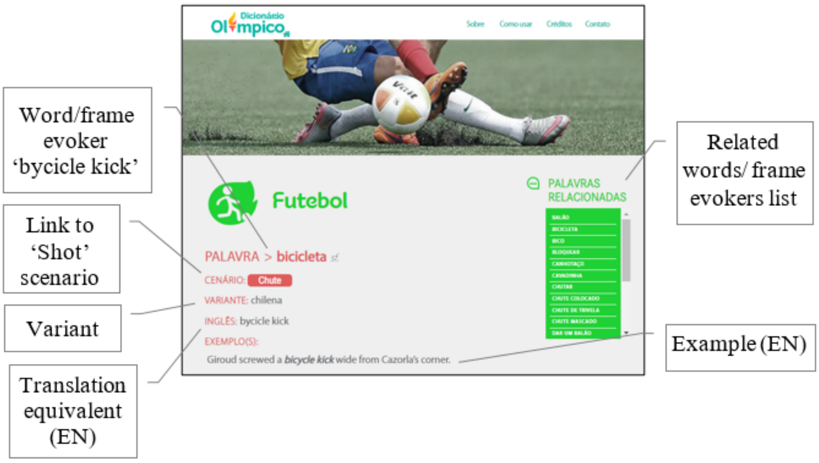


Futebol

CENÁRIO > Chute

No futebol, a posição do pé determina a força e o efeito do chute. Se dado com a parte interna do pé, o chute tem mais precisão e menos violência. Já os chutes de peito de pé são, ao mesmo tempo, precisos e violentos. Os chutes de bico são mais violentos e nada precisos, geralmente usados como modo de surpreender a defesa ou afastar uma bola perigosa. Os chamados voleios são desferidos quando a bola está ainda no ar, a meia altura do chutador, podendo ser bastante violentos. Já o bate-pronto é desferido no exato momento em que a bola toca o chão.

Figure 3: Frame/Scenario level.



Futebol

PALAVRA > bicicleta *sf.*

CENÁRIO: Chute

VARIANTE: chilena

INGLÊS: bycicle kick

EXEMPLO(S):

Giroud screwed a ***bicycle kick*** wide from Cazorla's corner.

Figure 4: Lexical unit/Word level.

2.2 Olympic Dictionary's Revision and its Impact on Paralympic Dictionary

Recently, a process of revising the Olympic Dictionary was initiated in order to reflect on the enhancement of the dictionary. This work has provided the initial basis for planning the Paralympic Dictionary, since it can be considered an extension of the Olympic Dictionary. Therefore, we shall now address the main issues identified during this review, in order to determine the starting point for the development of the Paralympic Dictionary. Among these issues are: (i) aspects related to the elements of sports description and (ii) aspects related to the tool's design.

With regard to the set of elements in Portuguese which fulfil the function of describing sports, the need to include examples and word definitions was identified. Such inclusions concern the purpose for which the dictionary is used (cf. Atkins & Rundell 2008: 25) and would lead to a different lexicographic design of the dictionary at the word level. Currently, it cannot be said that the microstructure of the word provides the necessary elements for decoding (since the absence of word definitions compromises the understanding of the meaning) nor for encoding (since, to be able to use the word, the user would additionally need to have access to examples). In the planning of the Paralympic Dictionary, the inclusion of these elements is now foreseen. Regarding the elements in English available only in the microstructure of the word (translation equivalents and examples), some inconsistencies stand out. Firstly, it must be stated that, in a frame-based dictionary, words and scenarios share a similar status of meaning description, for it is from their combination that the definition is established (double-decker definitions). Therefore, there is an interdependent relationship between these elements (cf. Fillmore 2003). A different treatment for words, in the sense of presenting or omitting elements, is not justified or seems inconsistent with the proposal for a frame-based dictionary. If the words are matched with translation equivalents, the scenarios (and, consequently, the sports) should as well. Secondly, the role that the elements in English play in the dictionary is not well defined when analysing such elements in isolation from the elements in Portuguese, since the equivalents and examples are not sufficient for the user to be able to understand or use the word in English. Finally, there is also an inconsistency related to the definition of the target audience. Currently, the dictionary can be classified as a unidirectional (Atkins & Rundell 2008: 24) or monodirectional (Welker 2008: 23) bilingual, since the access to the tool's content is only possible from Portuguese and, therefore, the dictionary only suits the Portuguese speaking/learning public. One way to address all of these inconsistencies would be to develop a bidirectional bilingual lexicographic structure for the dictionary. The planning of the Paralympic Dictionary already foresees this structure.

With regard to the elements of sports description in the Olympic Dictionary, the revision demonstrated that these elements assumed very different identities in terms of form and content throughout the dictionary. In other words, elements such as scenario description and sport description did not follow general guidelines that determined, for example, criteria for the selection/composition of the content (information type, level of detail, etc.) and criteria for content presentation (description size, paragraph size, use of links and bold type, etc.). The diversity displayed by these elements resulted from the methodology adopted in the dictionary's development: each editor was responsible for the study of a set of sports (randomly distributed) and, as a result, they were given autonomy to make decisions regarding the description of the sports under their responsibility. However, it can be said that, in the midst of this heterogeneity of content and form, there are similarities which did not result from the adopted methodology, but from the similarities among sports. For the Paralympic Dictionary, therefore, this is the starting point: the similarity among sports is the base of the work methodology, less centred on the individual autonomy of the editors, but mainly focused on the standardization of the dictionary elements and, consequently, on the definition of an identity for the tool. So far, it can be said that this methodology has already demonstrated success in enabling the standardization of the dictionary for it has proved itself efficient for scenario descriptions, especially those which exist in all sports, such as 'equipment', 'competition venue', 'competition officials' and 'athletes'. The experiment with the scenarios determines the success of the methodology because the scenarios fulfil a central function in defining the other elements, such as conceptual maps, word definitions, and the proposition and naming of the scenarios.

With regard to the tool's design, Olympic Dictionary's revision was based on the dictionary's use experiences by the members of the research group and also based on discussions with the programming specialist responsible for developing the Paralympic Dictionary's interface. On this basis, it was possible to identify a list of obstacles that hinder the tool's usability. To exemplify, it is possible to mention issues related to (i) the layout of the homepage, (ii) the search box, (iii) the use of hyperlinks, (iv) map reading, and (v) the development of an interface for smartphones. Below, we describe each of these issues, exhibiting how we intend to address them in the Paralympic Dictionary.

i. The layout of the homepage is programmed to show only one third of the sport icons for each click on "Load more". Loading can be automatic, requiring only the use of mouse scroll.

ii. The search box allows the user to search for sports, scenarios or words. Regarding visual aspects, the search box displays white letters on a transparent white background which makes it difficult to read the instruction ('Type a word or scenario or sport') and view the information typed by the user. The search option can appear through the image of a magnifying glass positioned at the top right of the page, as in most sites, and can expand with a click. When presenting the results of a search, the words are accompanied by the initials of the level they represent – P for word (*palavra*), C for scenario (*cenário*) and M for sport (*modalidade*). It is questionable, however, if the user is able to decode this information. The complete word can replace the use of initials.

iii. The use of links has been little explored. Throughout the descriptions of scenarios and sports there is no use of links, which does not encourage navigation through the site. Links can be incorporated into the descriptions of sports, scenarios and words, in order to facilitate access to related information. Concerning maps, there is no indication of links' presence, which does not allow the user to realize they even exist. Here too, the use of links plays an important role in encouraging navigation throughout the dictionary.

iv. The option to zoom in was used as a means to enhance the visualization of the information displayed on the conceptual maps. However, such a strategy has not proven itself to be so effective for the purpose, since, by enlarging some specific aspect, users can lose the overview of the whole, which is necessary for comprehension. On maps with a greater flow of information, this problem is even more evident. The alternative for a better visualization of the maps is to elaborate navigable/adaptable maps, built from and centred around the user's interaction.

v. The Olympic Dictionary does not have a version developed specifically for smartphones. As a result, viewing content from these mobile devices presents several problems. Addressing this issue involves developing an interface based on the characteristics and experiences resulting from smartphone-use experience.

3 Specific Challenges with the Paralympic Dictionary

In the present section, the specific aspects of the Paralympic context that brought, and still bring, implications for the design of the Paralympic Dictionary will be addressed. These specificities are mainly related to the nature of Paralympic sports, the description of which involves adopting a perspective to address 'disability'. The decision to develop the Paralympic Dictionary was taken after the launch of the Olympic Dictionary, and was driven by the fact that the research group already had a frame-based sports description methodology and a semi-ready interface that could be modified depending on the number and types of Paralympic sports. However, a first exploration of the Paralympic domain demonstrated substantial differences with direct implications for the planning of the tool. If, in the Olympic context, the development work was limited to the study and description of sports, in the Paralympic context, it was observed that these activities also involve an understanding of 'disability'. The consequences of this finding for the planning and development of the Paralympic Dictionary will be discussed below.

3.1 Disability Implications for the Dictionary Planning

From the study of the Paralympic context, it was observed that there is an overlap between the notions of 'athlete' and 'person with disability'. As a result, differently from what happened in the development of the Olympic Dictionary, there was a need to explore, in addition to the sport, the context of disability with the aim to verify how notions from broader contexts are mapped onto the Paralympic context.

As part of the research group's effort to understand the Paralympic universe, de Oliveira (2019) carried out a study on the conceptualization of 'Paralympic athlete', drawing on the theoretical framework of Frame Semantics. In this study, the different perspectives from which the notion of 'person with disability' can be understood were discussed, taking into account internal and external issues to the Paralympic context, in order to identify the elements that constitute the concept of 'Paralympic athlete'. The investigation revealed that the 'person with disability' frame can be understood from four different perspectives: the charitable; the medical; the social; and the rights-based one.[1] From the charitable perspective, the person with disability is seen as pitiable and a victim of their own disability; from the medical perspective, disability is seen as an organic problem and, therefore, the person with disability needs to be cured; from the social perspective, the person with disability is seen as dependent on or hostage to the social environment, which tends to promote exclusion, inaccessibility and prejudice; finally, from the rights-based perspective, people with disabilities are understood from their rights to equal opportunities and social participation with a focus on their empowerment and responsibility (de Oliveira 2019: 46-49). The perspectives identified by de Oliveira (2019) are related to the historical evolution of the discussion about disability and have different statuses with regard to their legitimacy within the community of people with disabilities, whose purpose it is to disseminate an image with no prejudices in relation to people with disabilities.

Regarding the notion of 'Paralympic athlete', the study revealed that there is a frame that unites elements from the broader notion of 'athlete' to elements from one perspective of 'person with disability'. On the one hand, the Paralympic athlete frame is based on features, such as 'athlete's principles' (equality, inspiration, courage, determination, fair play) and 'athlete's attitudes' (balanced training, sporting excellence, adequate nutrition), resulting from a general conception of 'athlete'. On the other hand, the same frame is also based on features that are presented in the form of commitments or responsibilities, such as 'changing perceptions', 'redefining limits of what is possible', 'stimulating the world', 'contributing to a more inclusive society', which are related to the condition of an athlete with disability (IPC 2015). Furthermore, the image of the Paralympic athlete is based on the concept of the Paralympic sport as a high-performance sport.

Another important aspect of the discussion about the 'Paralympic athlete' approach addressed by de Oliveira (2019) concerns the relationship between the lexicon and the different conceptualizations of 'person/athlete with disability'. According to de Oliveira, the community of people with disabilities favours the use of certain words or terms that contribute to the recognition and appreciation of a person with disability and rejects others that would contribute to minimize an athlete's abilities (poor, crippled, sick – charitable or medical perspective) or to overestimate their actions and achievements (superhero, example of overcoming). In both situations, there would be an attempt (even if unconscious) to remove the person with disability from their 'person' condition and to emphasize the disability.[2] In this regard, language policies play a crucial role, since, by offering guidelines for the use of appropriate words to refer to people with disabilities, they contribute to disseminating an image of people with disabilities that is consistent with that advocated by the community of people with disabilities.

[1] According to the International Classification of Impairments, Disabilities, and Handicaps (ICIDH), the medical and social models are a synthesis of the various models proposed to describe 'disability', describing it from the biological, individual, and social perspectives (ICIDH: 18).

[2] Although some authors defend the replacement of 'disabled' by 'person with disability' (e.g. Sassaki 2005), it is important to highlight that both the World Health Organization and documents such as the ICIDH (cf. note 1) do not take a rigid position in relation to these terms, considering that when it comes to the dimension of the individual, there is no consensus (ICIDH: 188).

The investigation that started with the work of de Oliveira (2019) and that continues to be carried out by the research group has enabled reflections on the identity of the Paralympic athlete and has guided decisions about how the dictionary will represent it. In practical terms, this discussion affects the definition of the structure of the Paralympic Dictionary in many ways. First, it points to the importance of the lexical choices that will be made in the different elements of the dictionary to refer to the Paralympic athlete. Such choices should reflect the commitment to valuing people with disabilities and building an inclusive society, and the task of discouraging discriminatory practices.[3] Second, it is foreseen to include a tab that has the function of describing the notion of Paralympic athlete that prevails in the Paralympic context, following a format similar to the description of sports (scenarios and words). A third point refers to the authority argument. Considering that disability is an unknown reality among the members of the research group, the discussion served as an alert for the need of the participation of people capable of guiding the group on making knowledgeable and informed decisions (cf. Bergenholtz & Kaufmann 1997: 93). The development of the Olympic Dictionary benefited from the contribution of external collaborators (athletes, coaches, teachers, etc.) whose function it was to review and validate the contents to be published. In the Paralympic Dictionary, in addition to the contributions related to the elaboration of sports content, there is a need to have collaborators who can play a broader role and who help to outline the identity of the dictionary as a whole with regard to the promotion of inclusion and dissemination of the Paralympic athlete's image. Finally, the discussion about disability also influences the definition of the target audience of the dictionary, since by assuming the commitment to disseminate a view of the athlete with disability legitimized by the community of people with disabilities, it was also assuming, indirectly, the commitment to promote digital accessibility in the context of the Paralympic Dictionary. In this sense, the discussion on defining the target audience takes place at a broader level, in order to ensure that people with disabilities have access to the dictionary.

Therefore, it can be said that the investigation – in progress – of the disability context has given rise to indispensable reflections for the planning of the Paralympic Dictionary, the most relevant being the finding that the use of certain terms implies the adoption of a specific perspective on 'Paralympic athlete' and 'disability'.[4] The social commitment to disseminate the vision of a Paralympic athlete as a high performance athlete is thus a central premise for the development of the Paralympic Dictionary. In general, it is intended to be in line with what the Paralympic athlete and the official bodies/institutions want to emphasize in relation to the performance and achievements of the athlete in sport and also in relation to the identity of people with disabilities.

3.2 Disability Implications for the Dictionary Development

In this section we discuss the study of the Paralympic sports and the repercussions of the disability in the proposal and elaboration of the elements of sports description. Based on the experience of working with Olympic sports, the exploration of Paralympic sports has revealed, so far, among the transversal (or ontological[5]) frames, the existence of a new frame and the need for restructuring one of the frames already identified in the development of the Olympic Dictionary. Two of the central characteristics of Paralympic sports refer to eligibility – deficiencies allowed for the practice of a particular sport, and functional classification – a system that aims to minimize the impact of impairments on athletes' performance and ensures fair competition.

As a result, in the context of the Paralympic Dictionary, the 'functional classification' frame was identified, designed to account for information regarding the types of impairments that are accepted in the practice of a particular sport (eligibility) and the ways of classifying these impairments (functional classification/functional classes). In this regard, in para powerlifting, for example, physical impairments in the legs and hips are eligible, and the sport has one functional class; for athletics, physical, intellectual or visual impairments are eligible, and these are grouped into 32 functional classes. The 'functional classification' frame, therefore, covers information of this nature. The fact that the more specific information about the disability would appear, above all, in the 'functional classification' frame, raised a discussion about the place of the disability in the dictionary. With regard to the description of sports, it was observed that disability could be used as a starting point. However, such an approach seems to give more emphasis to the impairment than to the high-performance feature. As a result, a proposal is being evaluated that treats disability as one of the elements of Paralympic sport and not

[3] It is important to highlight that, in the case of the Paralympic Dictionary, these decisions will not be aimed at meeting individual preferences, but at representing the choices of, above all, athletes and other people with disabilities involved in Paralympic sports. For this reason, the dialogue with them will be central to guaranteeing representativeness.

[4] Regarding the perspectives from which 'disability' can be seen, one can mention the example of the term 'impairment', widely used in Paralympics' official documents, which represents a medical perspective to categorize disabilities and athletes from an organic point of view. Paralympic Dictionary is intended to address social aspects of disability in addition to the organic ones, and as such it is likely that this will result in the use of different terms according to what one wishes to highlight. In general, with regard to individuals, the term 'person with a disability' will be preferred. To refer to body parts or body functions, the terms 'disability' and 'impairment' will be used, as defined by the ICIDH (cf. note 1): disability (umbrella term for impairments, p. 158) and impairment (loss or abnormality of a body part (i.e. structure) or body function, p. 165). Observe that in Portuguese, except in very specific cases, both 'disability' and 'impairment' can be translated as '*deficiência*' (disability).

[5] Frames that refer to knowledge about objects, participants and other static entities (cf. de Souza 2015). Such frames are recurrent in many if not all sports.

as the lens through which Paralympic athletes and sports are seen. In this decision-making process, the participation of collaborators will play a fundamental role in giving legitimacy to the dictionary proposal.

Another particular aspect of the Paralympic Dictionary concerns the restructuring of the 'technical team' frame, identified in the development of the Olympic Dictionary and which, there, comprised information related to the members of the teams/delegations, such as coaches, technicians, assistants, masseurs, etc., that is, professionals who, indirectly (through technical-tactical guidance and medical-therapeutic care), assist athletes during competitions. Although it maintains its original coverage derived from the methodology used in the Olympic Dictionary, in the Paralympic Dictionary, this frame also covers other participants. Indeed, the study of Paralympic sports revealed that there are professionals who perform the function of assisting athletes, but in a more direct way, within the actions/situations that constitute the competitions. In track cycling, for example, in the BC3 class (athletes with visual impairment), cyclists compete together with a sighted rider, who occupies the front seat of the bicycle (tandem) and makes tactical decisions. Another example is bocce ball, a sport in which athletes from classes BC1[6] and BC3[7] and some from BC4[8] can be supported by an assistant. In the BC3 class, for example, the assistant's function is to receive instructions from the athlete, adjust the height and position of the ramp and put the ball in position for the athlete to push. Bearing in mind that these professionals are not considered athletes, in the Paralympic Dictionary, the descriptions related to their functions would be part of the 'technical team' frame. Again, it is important to note that such decisions will be validated by a general specialist and sports experts. Such adjustments (inclusion and reformulation of frames) reflect the current stage of the Paralympic Dictionary development and do not end the discussion on the elaboration of the dictionary elements.

4 Digital Accessibility and Inclusion: From Olympic Dictionary to Paralympic Dictionary

During the early stages of the Paralympic Dictionary's development process, it was noted that in addition to the content issues already mentioned, technical adjustments aimed at promoting digital accessibility and inclusion would also be necessary.

Bearing in mind that the Paralympic Dictionary's interface was designed based on Olympic Dictionary's interface, it was concluded that it would be pertinent and productive to evaluate Olympic Dictionary in terms of accessibility, both for the Olympic Dictionary's second edition and for the Paralympic Dictionary itself. These are the reasons behind the addition of a new layer to the revision of the Olympic Dictionary, in order to assess this aspect. As a brief case study, we now present an initial digital accessibility analysis aimed at assessing how Olympic Dictionary meets the needs of the blind and visually impaired.

To perform the analysis exercise, we used the NVDA[9] (NonVisual Desktop Access) screen reading software. NVDA is a free assistive technology, available only for Windows systems, developed to help blind people and people with vision impairment in using the computer. It is worth noting that there are two possibilities for reading the screen with such software: the user can use the mouse cursor, or the TAB key on the keyboard. By using a mouse, the user manages access to information by hovering the cursor over the content of the website pages. Regarding the textual content of the Olympic Dictionary, this reading worked satisfactorily – all strictly textual content was identified by the software;[10] with regard to the order in which the information is read, it is the user who chooses the content to be read and it is up to them to pass the cursor over the text in order to guarantee the full reading of the contents.[11] When using a keyboard, the user manages access to information via the TAB key – whenever the user presses this key, the reading of a hyperlink displayed on the page begins. In this reading, all the text elements that do not have a hyperlink attached are not readable by the software. With regard to the order in which the information is read, the reading of the conceptual maps via keyboard access does not present the contents in a logical sequence; the display order of the information is defined from the position which the textual elements occupy on the page, starting from top to bottom and from left to right. In both forms of access (mouse and keyboard), the software did not read the sports icons, images and the textual content without hyperlinking presented by the conceptual maps, since the Olympic Dictionary does not provide a textual description for these items. Thus, it can be said that the main result obtained from the experiment with the NVDA software was the absence of a textual description for the Olympic Dictionary icons, images and maps.

[6] Athletes in the BC1 sports class have severe activity limitations that affect their legs, arms and torso due to coordination deficiencies and generally rely on a motorized wheelchair (Tokyo 2020).

[7] Athletes competing in the BC3 sport class have significantly limited functions in their arms and legs, and little or no torso control due to brain or non-brain origins (Tokyo 2020).

[8] The BC4 sports class comprises athletes with disabilities without a cerebral origin. Possible health conditions include progressive weakness and loss of muscle mass (muscular dystrophy), spinal cord injuries or amputations that affect the four limbs (Tokyo 2020).

[9] Software available at: https://www.nvaccess.org/.

[10] The textual content presented in the conceptual maps, however, is not readable by the software since it was added to the dictionary interface as an image.

[11] Whenever the formatting changes (bold, italics, …), the reading done by the software is stopped and the user needs to position the cursor over the word to restart the reading.

From an initial bibliographic exploration on digital accessibility approaches for visual elements, a survey of possible actions was taken in the context of the Olympic and Paralympic dictionaries. Below, we list some of these actions, indicating the target audiences.

i. Considering blind people, it is important to add the description of the images on the site and develop conceptual maps readable by assistive technologies;
ii. For people with mild or moderate vision impairment, it is important to enable the options 'high contrast' and 'increase/decrease the font size', as well as to provide a textual structure with shorter paragraphs to facilitate the reading made by assistive technology;
iii. For colour-blind people and people with mild or moderate vision impairment, it is important to rethink the colour scheme of the page;
iv. For people with mild, moderate, and severe vision impairment and blindness, it is important to make different media available for textual illustration, such as audio (and video).
v. It is worth mentioning that the actions listed here are the result of an initial exercise to evaluate digital accessibility in the Olympic Dictionary; complementary analysis could confirm the problems identified, as well as indicate the need for additional adjustments, aiming to serve other audiences with disabilities not covered so far (such as those for deaf people).

5 Conclusion

The main focus of this study was to address the challenges which emerge when dealing with the Paralympic context in lexicography and how these challenges impact the development of the Paralympic Dictionary. By doing so this exercise made us to reflect upon, to some extent, the role of dictionaries as cultural institutions and the challenges which often appear when compiling domain-specific dictionaries, in our case those based on the theory of Frame Semantics (Fillmore 1982, 1985).

The reflections brought about by the ongoing development of the Paralympic Dictionary have a significant influence on the work being carried out by the SemanTec research group, not only in the sense of contemplating the specifics of the new tool, but also to open the way for deeper reflections on a metalexicographical level with regard to "completed" projects as well as future ones. On the one hand, the revision of the Olympic Dictionary has guided practical decisions regarding the activity of describing sports, such as the writing of encyclopaedic definitions, and even the frame recognition/identification methodology itself; on the other hand, the demands related to disabilities and digital accessibility raised questions about the user profile, which have direct implications for the design of the new tool. Together, these findings and their implications represent an important milestone in the group's work as they force the group to re-evaluate decisions taken under different circumstances and to update the positions adopted by the group in a manner that is more consistent with the realities described.

Regarding the limitations of this study, it should be noted that the considerations on digital accessibility presented here still represent a very early stage of the investigation: in addition to deepening the study of the different types of disabilities and the needs related to them, it will also be necessary to verify how and to what extent the dictionary will be able to meet the different demands of people with disabilities. In this sense, it will be of great importance to be aware of the different assistive technologies available and how well they meet the needs of their users. In addition, this work also reinforces the centrality and urgency of establishing a dialogue with entities and organizations that represent athletes with disabilities, as well as the athletes themselves, so that the dictionary can also be an empowerment tool for all these people.

6 References

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York, NY: Oxford University Press.

Bergenholtz, H. & Kaufmann, U. (1997). Terminography and lexicography. A critical survey of dictionaries from a single specialised field. *Hermes, Journal of Linguistics*, 18, pp. 91-125.

Chishman, R. (2014). *Field – Dicionário de Expressões do Futebol. 2.ed.* [Field – Football Expressions Dictionary. 2nd ed.]. São Leopoldo: Unisinos. Available from http://dicionariofield.com.br/.

Chishman, R. et al. (2015). The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary. *Revista de Estudos da Linguagem*, 23, pp. 769-796.

Chishman, R. et al. (2016). *Dicionário Olímpico* [Olympic Dictionary]. São Leopoldo: Unisinos. Available from http://www.dicionarioolimpico.com.br/.

Chishman, R. et al. (2017) Dicionário Olímpico: a semântica de frames encontra a lexicografia eletrônica. In M.J.B. Finatto et al. (eds) *Linguística de Corpus: Perspectivas*. Porto Alegre: Instituto de Letras, UFRGS, pp. 265-298.

Chishman, R. et al. (2019). Challenges and difficulties in the development of *Dicionário Olímpico* (2016). In I. Kosem et al. (eds) *Electronic Lexicography in the 21st Century (eLex 2019): Smart lexicography. Conference proceedings. Sintra, Portugal, 1-3 October 2019*. Brno: Lexical Computing, pp. 622-641.

de Oliveira, S. (2019). *O atleta com deficiência no contexto paraolímpico: uma análise dos frames que entram no jogo* (MA dissertation). São Leopoldo: Unisinos. Available from http://www.repositorio.jesuita.org.br/handle/UNISINOS/7799.

de Souza, D.S. (2015). *Jogada de letra: um estudo sobre colocações à luz da Semântica de Frames* (MA dissertation). São Leopoldo: Unisinos. Available from http://www.repositorio.jesuita.org.br/handle/UNISINOS/3924.

Fillmore, C.J. (1976). Frame semantics and the nature of language. *Proceedings of the Conference on the Origin and Development of Language and Speech, New York, 1976*. New York, NY: New York Academy of Sciences, pp. 20-32.

Fillmore, C.J. (1982). Frame Semantics. In The Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., pp. 111-137.

Fillmore, C.J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2), pp. 222-225.

Fillmore, C.J. (2003). Double-decker definitions: The role of frames in meaning explanations. *Sign Language Studies*, 3(3), pp. 263-295.

Fillmore, C.J. & Atkins, B.T.S. (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbors. In A. Lehrer & E. Kittay (eds) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale, NJ: Erlbaum, pp. 75-102.

IPC – International Paralympic Committee. (2015). *Strategic Plan 2015 to 2018*. Available from https://www.paralympic.org/sites/default/files/document/150619133600866_2015_06+IPC+Strategic+Plan+2015-2018_Digital.pdf.

NVDA. (2021). *NonVisual Desktop Access Screen Reader* (Screen Reader used by blind and vision impaired people to use the computer). Available from https://www.nvaccess.org/.

Sassaki, R.K. (2005). Atualizações semânticas na inclusão de pessoas: Deficiência mental ou intelectual? Doença ou transtorno mental? *Revista Nacional de Reabilitação*, 9(43), pp. 9-10.

Tokyo – The Tokyo Organising Committee of the Olympic and Paralympic Games. (2020). *Paralympic Games Classification*. Available from https://olympics.com/tokyo-2020/en/paralympics/games/classification/.

Welker, H.A. (2008). Lexicografia Pedagógica: Definições, história, peculiaridades. In C. Xatara, C. Bevilacqua & P. Humblé (eds) *Lexicografia Pedagógica: Pesquisas e Perspectivas*. Florianópolis: Núcleo de Tradução, Universidade Federal de Santa Catarina, pp. 9-45.

Acknowledgements

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001; the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); and the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS).

EURALEX XIX
Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Pedagogical Lexicography

The design of an explicit and integrated intervention program for pupils aged 10-12 with the aim to promote dictionary culture and strategies

Gavriilidou Z.[1], **Konstantinidou E.**[2]

[1] *Democritus University of Thrace, Greece*
[2] *Democritus University of Thrace, Greece*
zoegab@otenet.gr, evi1990@hotmail.gr

Abstract

The purpose of this paper is to elaborate on the theoretical principles of an intervention program created for promoting dictionary culture through the adoption of dictionary use strategies for pupils aged 10-12 attending Greek schools and also to describe, one by one, the steps of its implementation and content. We also aim to present the pedagogical priorities, the instructional choices, in terms of materials, topics, tasks, assignments and projects, and the ways to assess the impact of the program on pupils' dictionary use. The program is integrated in the course of Greek language teaching in mainstream public schools in Greece and it follows the principles of a strategy-based, differentiated and explicit instruction.

Keywords: dictionary use strategies, reference skills, dictionary culture, strategy-based learning, explicit teaching, pedagogical lexicography.

1. Introduction

Previous research (Chatzipapa 2018; Mavrommatidou 2018, Mavrommatidou et al 2019) has shown that dictionary users do not possess dictionary culture and that they have not adopted dictionary use strategies. Dictionary culture has been defined as "the critical awareness of the value and limitations of dictionaries and other reference works in a particular community" (Hartmann and James 1998: 41). Gouws (2013) maintains that dictionary culture refers to the familiarity with dictionary using skills and knowledge of when to use a specific dictionary or other tool. On the other hand, Gavriilidou (2013), elaborating on the idea of language learning strategy (Oxford 1990), refers to strategic dictionary use and defines dictionary use strategies as techniques used by the effective dictionary user in order to decide whether to use or not an appropriate type of dictionary and make a quick and successful search in it. The development of dictionary culture and strategic dictionary use can be achieved when dictionary pedagogy and dictionary routines are introduced in everyday classroom activities in a systematic way in order to bridge the gap between lexicographers and users. Strategic dictionary instruction should be an integral part of language education (first, second, foreign or heritage), since it helps students acquire dictionary culture, gain greater proficiency and confidence in dictionary use, and become self-aware about when and how we chose to use a dictionary in an autonomous way. This training empowers dictionary users by directing them to employ effective dictionary use strategies, developing their appropriate reference skills, and allowing them to take absolute control of the process of word look ups (quickly finding the right entry and choosing the correct meaning successfully). Subsequently, successful dictionary consultations enhance users' motivation to use dictionaries more often, which may be strongly discouraged by unsuccessful dictionary use.

The purpose of this paper is to elaborate on the theoretical principles of an intervention program created for promoting dictionary culture through the adoption of dictionary use strategies for pupils aged 10-12 attending Greek schools, and also to describe, one by one, the steps of its implementation and content. We also aim to present the pedagogical priorities, the instructional choices, in terms of materials, topics, tasks, assignments and projects, and the ways to assess the impact of the program on pupils' dictionary use.

In the first part of the paper, we offer a literature review justifying the theoretical background which underlies the intervention program: the notions of dictionary culture, dictionary use strategies, strategy-based instruction, explicit and integrated teaching are briefly presented in this part. In the second part we present the learning outcomes of the program, the list of activities, the adapted activities for people with various degrees of knowledge of Greek, the material used and the criteria for the assessment of each learning outcome. The presentation ends with the conclusions and limitations of the study.

2. Dictionary Culture and Dictionary Use Strategies (DUS)

The educational value of dictionaries is not always acknowledged by teachers, pupils, or students. This happens because modern pedagogy and national policy makers have not highlighted the importance of including dictionary training in classroom nor have they tried to establish a dictionary culture, resulting in pupils and students considering dictionaries as boring inaccessible books that have nothing to offer. Students are left unaware of how a dictionary can help acquire new knowledge or help in problem solving situations arising inside and outside class, even though it is acknowledged that the way a dictionary is included "in an educational system may affect the development of dictionary skills" (Campoy-Cubillo 2015: 120). Referring to dictionary culture, Nkomo (2015: 74) suggests that "a valid distinction is made between a societal or collective dictionary culture and an individual or idiolectal dictionary culture" implying that individual dictionary culture should be backed up by the society and policy makers. In other words, when an educational system incorporates dictionary use in curricula of language teaching (L1, L2, etc.) and provides constant in-service training to teachers for gaining expertise in dictionary use training, then dictionary use is valued.

Dictionary use strategies (Gavriilidou 2013) seem to be the appropriate learning tool for achieving dictionary culture. In that paper, the author connected the descriptive notion of 'reference skills' with the theoretical construct of 'language learning strategies' in an attempt to establish a strong linkage between theory (of learning) and practice (how to train dictionary use). Based on the results of a factor analysis, she classified DUS for paper dictionaries in four categories: 1) Dictionary awareness strategies which refer to the critical awareness of the value and shortcomings of the dictionary that lead to the decision to use a dictionary in order to resolve a specific problem encountered during learning inside or outside the classroom , 2) Dictionary selection strategies which allow the choice of an appropriate dictionary depending on the problem to be solved and guarantee the familiarity with one's own dictionary, 3) Lemmatization strategies, which help dictionary users find the citation form of inflected forms found in the text by relying on morphological indices (stems, prefixes, suffixes, inflectional morphemes) of the unknown word they come across in the/a text in order to make hypotheses about the look-up form of that word. Lemmatization strategies also include skills in alphabetical sequencing, otherwise lemmatization is not possible, and 4) Look-up strategies, which control and facilitate the localization of the correct section of the entry where different meanings of the same polysemous word form are included. These four types of strategies are summarized in Tables 1 and 2. It should be mentioned here that given the variety of electronic dictionary types (De Schryver 2003), "novel ways of accessing lexicographic data" are required (Lew 2013:16); In this perspective electronic dictionary use strategies may overlap with digital literacy (Lew & De Schryver 2014) or differ than strategies employed during paper dictionary look-ups. For instance, users require navigation strategies or look-up strategies in the new electronic environment. Mavromatidou et al. (2019) offer a detailed list of DUS for electronic dictionary use.

Depending on the type of processing involved, these strategies can be further classified into metacognitive, cognitive, memory and compensation. Metacognitive DUS such as self-management, self-monitoring, self-reflection, decision making, planning, etc. (see tables 1 and 2) "are higher order executive skills" (O' Malley and Chamot 1990: 44) that can be applied in receptive or productive dictionary use for conflict resolution or evaluating dictionary use success. Furthermore, they make dictionary users aware of what they are doing and help them setting look up goals and deploying alternative plans when the goals are not met. Cognitive DUS such as inferencing or alphabetization, on the other hand, "operate directly on incoming information" (O' Malley and Chamot 1990: 44) processing it in ways that lead to successful look-ups. Memory DUS such as use of mnemonics to remember the word to be searched are used to help users remember information that facilitates look-ups. Finally, compensation DUS, such as paying attention to headwords, signposts or example sentences, enable dictionary users to better navigate the dictionary and are intended to make up for inadequate information or skills.

Like Language Learning Strategies, DUS are problem-oriented; they are used because there is a problem to solve (e.g., the need to search the meaning of a word which obscures reading comprehension), a task to accomplish (e.g., a synonym exercise in the textbook), an objective to meet (e.g., a successful look-up), or a goal to attain (e.g., new vocabulary acquisition, participation in oral communication, etc.). They are also action-based, since users have to accomplish specific actions to ensure successful word look-ups. These actions depend on users' characteristics. Some of them contribute directly to successful look-ups (e.g., alphabetizing), while others contribute indirectly but efficiently (e.g., decision making or self-monitoring). They are not always observable and students are often unaware of using them. Furthermore, they are flexible in the sense that users chose and combine them in a quite individual manner that does not allow to identify specific sequences or patterns (Oxford 1990). Their choice depends on variables such as gender, motivation, learning style, educational and proficiency level, school type, purpose of the task to be accomplished, career orientation and general reference skills (Campoy-Cubillo 2015; Chadjipapa et al 2020; Gavriiidou et al 2020). "The dictionary skills of a language learner depend upon dictionary look-up strategies and the language learners' ability to use the best possible strategy in a particular context and for a specific purpose" (Campoy-Cubillo 2015: 120). Finally, they are teachable through strategy training which aims to make students aware of why, how and when they should be used inside and outside the classroom (see 3 below).

Strategy Group	Representative Strategy		Definition
Dictionary awareness strategies	Dictionary use to find semantic information	Dictionary use to find the meaning	Deciding to use the dictionary to look up an unknown word
		Dictionary use to find synonyms	Deciding to use a dictionary to look up a synonym of a word you need to complete a vocabulary exercise or find an appropriate synonym while writing
		Dictionary use to find antonyms	Deciding to use a dictionary to look up an antonym of a word you need to complete a vocabulary exercise or find an appropriate antonym while writing
		Dictionary use to find word families	Grouping and classifying words according to their semantic attributes
		Dictionary use to find the meaning of phraseology	Deciding to use the dictionary to look up word phraseologies
	Dictionary use to find grammatical information	Dictionary use to find the syntax of a word	Deciding to use the dictionary during productive dictionary use to look up how a word is used in a sentence
		Dictionary use to find inflection/the derivatives of a word	Deciding to use the dictionary to find derivatives of a word or verify how a word is inflected
		Dictionary use to find the spelling of a word	Deciding to use the dictionary to find the spelling of a word
	Dictionary use to find the etymology of a word		Deciding to use a dictionary to find the etymology of a word
	Receptive dictionary use		Deciding to use the dictionary during text comprehension tasks
	Productive dictionary use		Deciding to use the dictionary during text production tasks
	Dictionary use for pragmatic reasons (register)		Deciding to use dictionary labels for accessing pragmatic information
	Dictionary use at home		Deciding to use the dictionary at home for receptive or productive language skills
	Dictionary use for translation		Deciding to use the dictionary during translation tasks
Dictionary selection strategies	Recognizing different types of dictionaries and the type of information they include	Selecting to use an etymological dictionary	Being aware of the content and form of an etymological dictionary and identifying the tasks that require the use of it
		Selecting to use a general/learners' dictionary	Be aware of the content and form of a general or learner's dictionary and identifying the tasks that require its use
		Selecting to use a bilingual dictionary	Being aware of the content and form of a bilingual dictionary and identifying the tasks that require its use
		Selecting to use a dictionary of technical terms	Being aware of the content and form of a dictionary of technical terms and identifying the tasks that require its use
	Self-reflection on one’s needs		Deciding in advance which are the basic learning needs a dictionary can satisfy
	Key dictionary purchasing criteria	Decision to purchase considering the macrostructure and microstructure	Using key purchasing criteria to make a dictionary selection
		Decision to purchase considering the type of information included	Using content criteria to make a dictionary selection

Table 1 Dictionary Use Strategy classification and definitions I

Strategy Group	Representative Strategy		Definition
Strategies for lemmatization and acquaintance with dictionary conventions	Inferencing	Inferencing of the citation form	Using available information to predict the citation form
		Inferencing of the word spelling	Using available information to guess the spelling of a word
	Self-monitoring		Monitoring the success of a word look-up and readjusting it
			Monitoring the success of a proverb look-up and readjusting it
	Functional planning	Acquaintance with the Introduction of the dictionary	Planning to use the information from a dictionary introduction in order to find out how the entries are arranged
		Abbreviation awareness	Planning to get acquainted with the abbreviation list in order to learn what the abbreviations stand for
		Label awareness	Planning for getting acquainted with the labels used to better navigate in the entries
Look-up strategies	Alphabetization		Using previous knowledge on alphabetical order to locate a word in dictionary
	Memorization	Memorization of the word to look up	Retrieving a word from memory during word searches
		Memorization of the initial letter of the word to look up	Use mnemonics to remember the initial letter of the word to be looked up so that to effectuate a successful alphabetizing
		Selecting the appropriate meaning of a word assisted by the example sentences	Using example sentences as clues for selecting the appropriate meaning of a polysemous word
	Self-evaluation during receptive use	Using the context to evaluate how successful was the look-up	Checking the outcomes of the look-up by returning to the text to confirm that the word matches the context
	Self-management during productive use	Selection of the appropriate grammatical form	Understanding the importance of grammatical information for the successful use of the word

Table 2 Dictionary Use Strategy classification and definitions

3. Strategy-based Instruction in Dictionary Skill Training

Even though a dictionary is a valuable learning tool it requires special skills. Herbst and Stein (1987), Walz (1990) and Bishop (2000) are among the very few researchers who designed learning activities for training students how to use a dictionary. These first attempts to teach learners when and how to use a dictionary lacked systematicity and an underlying theoretical background that would maximize their effect.

Taking into consideration previous research which demonstrates that strategy use leads to skill-specific improvement (Chen 2007; Cohen, Weaver & Li 1998; Macaro 2001), can be taught (Cohen & Macaro 2007; O'Malley & Chamot, 1990) and, as a result, helps learners to become more efficient and self-regulating in their learning (Chen 2007; Hassan et al, 2005; O'Malley & Chamot 1990, Oxford 2011), it was decided to adopt a strategy-based instruction (SBI) model in training dictionary skills.

SBI is a learner-centered approach which refers to "any intervention focusing on strategies to be adopted and used autonomously by learners in order to improve their L2 learning and performance" (Vrettou 2015). It helps learners to take control of their learning, become autonomous and aware of their needs, strengths or weaknesses; in other words, it encourages them to 'learn how to learn'. In this approach, teachers describe and model useful strategies, elicit examples of student's experience, help learners reflect on their own strategy use, encourage them to experiment with strategy use and integrate strategies inside and outside the classroom (Cohen 2000).

The two most prominent strategy-based approaches to date are the Styles- and Strategies-Based Instruction (SSBI) model (Cohen 1998, 2000) and the Cognitive Academic Language Learning Approach (CALLA) (Chamot 2018; Chamot & El Dinary 1999, Chamot & O'Malley 1994, 1996). We opted to adopt the principles of the CALLA, considering previous literature that documented the positive effect of that approach on raising learners' autonomy (Chamot 2007; Gu 2007; Nguyen & Gu 2013).

CALLA is based on cognitive theory and integrates grade appropriate content, academic language development based on content and direct strategy instruction (see explicitness in section 4). Four types of tasks are used: a) easy and supported by the context, b) difficult but supported by the context, c) easy without the support of the context, d) difficult without any context support. Content, language and strategies are taught in a five-stage cycle (see section 5 below). This five-stage model is flexible and aims at raising learners' metacognitive skills and a gradual shift from the teacher to learner autonomy (O'Malley & Chamot 1990).

The DUS instruction program we developed follows CALLA's principles. In what follows, we set out the principles, directions and focuses that underly the compilation of the program.

4. Explicitness of Purpose and Integration in Language Course

Two crucial questions have to be taken into consideration when designing a syllabus or an intervention program for training learners in dictionary use strategies: the explicitness of purpose while teaching and the effectiveness of integrating strategy instruction into a language class. Previous research (Andersson 2002; Chamot 2005; Sarafianou & Gavriilidou 2015; Wenden 1986) has stressed that explicit instruction, i.e., instruction where teachers raise students' awareness by modelling strategy use, naming different strategies and creating opportunities in the classroom for strategy practice and self-evaluation of the effectiveness of strategy use, is more effective because it cultivates students' metacognition by helping them reflect on their own learning and thinking. This happens because, in explicit teaching, learners are informed about the importance of particular strategies and how to perform them successfully in specific classroom activities for facilitating attainment of learning goals. Thus, students connect specific strategies with specific learning tasks and are given feedback about their performance so that they can self-monitor their strategy use and transfer it to new situations.

Talking about dictionary use, explicit teaching of DUS results in appropriate knowledge and skill development to successfully use a dictionary, raises the independence and confidence of students as dictionary users, increases their motivation to use a dictionary, which may be negatively affected by unsuccessful look-ups, and develops their awareness of the positive strategies to be adopted while navigating dictionary entries. This is the reason why explicitness in dictionary use strategy teaching was adopted in our program.

Previous research (O'Malley & Chamot 1990; Oxford, 1993; Oxford et al. 1990; Walters 2006) has also investigated whether strategy instruction should be embedded into the language course or constitute a separate component, independent of the language course, in 'learning to learn' courses and training programs. It was demonstrated that learning in context is more effective because it is tied to specific tasks and learning goals. Furthermore, the learner realizes the usefulness of the strategies used in connection with specific activities, which facilitates retention.

In the same vein, dictionary use strategy teaching should be embedded in a language course, since research has shown that students maintain DUS when they can use them in situations similar to the ones in which they learned that specific strategy; and a language course, like everyday communication, offers multiple opportunities to look up words. More specifically, DUS training "should be tied to specific course objectives and fully integrated with other course content (Carduner 2003: 74). This means that in the frame of a language course, teachers may select DUS to teach, based on typical language tasks to be performed in the classroom (word definitions, synonym or antonym finding, etc.) and help their students see the applications of DUS in specific problem-solving situations. Thus, dictionary use strategy teaching embedded in the language course offers opportunities for contextualized dictionary practice. Additionally, during look-ups, students have the opportunity to collaborate in the class with their classmates, learn from their peers' performance and share with them successful DUS.

5. The Theoretical Principles of the intervention Program

In this section, we offer a detailed description of the theoretical choices adopted in our program. First of all, the program adopts the principles of *strategy-based instruction*. As already discussed in section 2, SBI enables learners to take an active role in the learning process by helping them to monitor and evaluate the way they learn (Cohen & Macaro 2007). The Cognitive Academic Language Learning Approach (CALLA) (Chamot 2007) has provided us with a useful framework for teaching dictionary use strategies. A five-phase recursive cycle for introducing, teaching, practicing, evaluating, and applying dictionary use strategies was implemented. This cycle was complemented with highly explicit instruction in applying DUS to learning tasks which gradually fades so that the pupils become more autonomous in selecting and applying their own preferred DUS. The five phases of the intervention program are the following:

a) *Preparation,* where students identify DUS they are already using and develop metacognitive awareness about the relation between DUS and successful look-ups. Activities in the preparation stage include class discussions, interviews, or think-aloud sessions about DUS recently used for specific learning tasks. More specifically, in this program learners are asked about their dictionary use habits, for instance, how often they look up words, if it takes them a long time to find the words they need, if there are some symbols in the lemmas that they do not understand, how they select the appropriate meaning, etc. Furthermore, pupils are informed about what a user can find in a dictionary entry in addition to its meaning, namely, its pronunciation, what part of speech it is, synonyms, the role of lexicographic examples and other information depending on the type of dictionary. The above are illustrated with examples from the school dictionary «Το λεξικό μας» (Our dictionary). This discussion is important because many dictionary users just look up the meaning of a word disregarding all other information.
b) *Presentation,* where the teacher models every DUS and explains, by using specific DUS names, how they are used, their characteristics, their effectiveness, their field of application. It is in this phase that the teacher presents in detail dictionary awareness strategies, dictionary selection strategies, lemmatization strategies (like finding the citation form of inflected forms included in the text by relying on morphological indices such as stems, prefixes, suffixes, inflectional morphemes of the unknown word they come across in the/a text in order to make hypotheses about the look-up form) and, finally look-up strategies (like alphabetizing, memorization), etc.
c) *Practice/Scaffolding,* where pupils are asked to practice all the above mentioned DUS in authentic learning situations such as reading comprehension, writing, explaining unfamiliar words, etc.
d) *Self-evaluation,* where pupils evaluate their success in look-ups, discuss the results of DUS practice, argue the usefulness of different DUS, talk about their favorite DUS, etc. This phase empowers pupils' metacognitive knowledge and experience, which constitutes a prerequisite for the following phase, that of expansion.
e) *Expansion,* where the pupils apply their preferred DUS to new contexts, different courses and outside the classroom.

Another crucial characteristic of the program is that the teaching is *explicit*, meaning that teachers overtly mention specific DUS and learners are informed about how, why and when to adopt dictionary use strategies and how to evaluate them and transfer them to new tasks. Dictionary users are given the opportunity to realize the benefits of strategic dictionary use, acquire a dictionary culture, evaluate the effectiveness of their dictionary look-ups, and expand dictionary use during various linguistic tasks.

Furthermore, the program is integrated in the language course activities of upper elementary school and follows the school textbook, because practicing dictionary use on authentic language tasks enables learners to perceive the relevance of a task, enhances comprehension and retention (Chamot & O'Malley 1987), while it can also help users maintain or enhance their motivation to use dictionaries.

Finally, the program adopts differentiated learning where teachers tailor their teaching approach to match their students' learning styles and needs. This can include choice of activities with different degree of difficulty for practicing the same DUS and offer every pupil multiple learning paths. The program also proposes adapted activities in order to respond to the needs of users with disabilities (learning difficulties, impaired vision, etc.).

6. The Content and Learning Outcomes of the Intervention Program

The intervention program includes 12 units of targeted paper dictionary use strategy instruction for pupils attending the two classes of upper elementary schools in Greece. Each unit corresponds to and is closely connected to a different chapter of the school textbook for teaching Greek as L1. The program may be conducted over a minimum of a 4-week period. However, the duration may be extended depending on the classroom needs, level and interest. The units focus on raising students' awareness about all four types of DUS: dictionary awareness strategies, dictionary selection strategies, lemmatization strategies and look-up strategies. Table 3 presents the number of strategies included in the program by strategy category.

Educational Level	Dictionary awareness strategies	Dictionary selection strategies	Lemmatization strategies	Look-up strategies
5th grade	36	6	14	18
6th grade	32	6	19	21
Total	68	12	33	39

Table 3: Nr of DUS types in syllabus by educational level

The material for the 12 units was created by the researchers to complement exercises and tasks from the school textbook with teaching resources that best cater students' dictionary use skills. So, there was a shift from a textbook-based and content-based mode to a more interactive skill-based approach. The choice of the content of instruction was based on items included in the *Strategy Inventory for Dictionary Use* (Gavriilidou 2013). Table 4 provides detailed information of the content of the program.

Syllabus Unit	Textbook chapter	Focus	Types of tasks	DUS Types
introduction		Recognizing different types of dictionaries and the type of information they include	selecting to use different types of dictionaries, decision to purchase considering the type of information included	selection strategies
1	1	Finding semantic information	find synonyms find the etymology of a word find the meaning	dictionary awareness /selection strategies
2	2	Inferencing of the reference form Finding semantic information	using available information to predict the reference form checking the outcome of a proverb look-up and start a new one in case it was unsuccessful find word families find synonyms and antonyms	lemmatization strategies, dictionary awareness strategies, dictionary selection strategies
3	4	Inferencing of the reference form Self-monitoring Finding grammatical and semantic information	using available information to predict the reference form checking the outcome of a proverb look-up and start a new one in case it was unsuccessful find the derivatives of a word find synonyms find the meaning of phraseology	lemmatization strategies, dictionary awareness strategies,
4	6	Alphabetization Selecting the appropriate meaning of a word assisted by the example sentences Using the context to evaluate how successful was the look-up Finding semantic information Finding grammatical information	making assumptions about the correct section of the dictionary to look-up the word by using previous knowledge on word order using example sentences as clues for selecting the appropriate meaning of a polysemous word checking the outcomes of the look-up by returning to the text to confirm that the word matches the context find synonyms find the derivatives of a word	look-up strategies, dictionary awareness strategies
5	7	Finding grammatical /semantic information Inferencing of the word spelling Inferencing of the reference form	find the spelling of a word/ word families/antonyms using available information to guess the spelling of a word using available information to predict the reference form making assumptions about the correct section of the dictionary to look-up the word by using previous	dictionary awareness strategies, lemmatization strategies look-up strategies

		Alphabetization	knowledge on word order	
6	9	Finding semantic information Inferencing of the reference form Selecting the appropriate meaning of a word assisted by the example sentences Using the context to evaluate how successful was the look-up	find word families find synonyms find the meaning using available information to predict the reference form using example sentences as clues for selecting the appropriate meaning of a polysemous word checking the outcomes of the look-up by returning to the text to confirm that the word matches the context	dictionary awareness strategies lemmatization strategies look-up strategies
7	10	Finding semantic information Memorization of the word to look up	find synonyms find antonyms use mnemonics to remember the word to be looked up during word searches	dictionary awareness strategies, look-up strategies
8	11	Memorization of the word to look up Alphabetization Inferencing of the reference form Label awareness Finding semantic information Finding grammatical information Inferencing of the word spelling Self-monitoring Recognizing different types of dictionaries and the type of information they include	use mnemonics to remember the word to be looked up during word searches making assumptions about the correct section of the dictionary to look-up the word by using previous knowledge on word order using available information to predict the reference form planning for getting acquainted with the labels used to better navigate in the entries find the meaning find antonyms find the spelling of a word using available information to guess the spelling of a word checking the outcome of a proverb look-up and start a new one in case it was unsuccessful selecting to use different types of dictionaries decision to purchase considering the type of information included	look-up strategies, lemmatization strategies, dictionary awareness strategies, selection strategies
9	13	Memorization of the word to look up Alphabetization Inferencing of the word spelling Inferencing of the reference form Finding semantic/grammatical information	use mnemonics to remember the word to be looked up during word searches making assumptions about the correct section of the dictionary to look-up the word by using previous knowledge on word order using available information to guess the spelling of a word using available information to predict the reference form find synonyms, antonyms, word families, spelling of a word	look-up strategies, lemmatization strategies, dictionary awareness strategies,
10	15	Selecting the appropriate meaning of a word assisted by the example sentences Memorization of the word to look up Memorization of the initial	using example sentences as clues for selecting the appropriate meaning of a polysemous word use mnemonics to remember the word to be looked up during word searches	look-up strategies, lemmatization strategies, dictionary awareness strategies

		letter of the word to look up Alphabetization Using the context to evaluate how successful was the look-up Inferencing of the reference form Finding grammatical information	use mnemonics to remember the initial letter of the word to be looked up so that to effectuate a successful alphabetizing making assumptions about the correct section of the dictionary to look-up the word by using previous knowledge on word order checking the outcomes of the look-up by returning to the text to confirm that the word matches the context using available information to guess the spelling of a word find the spelling of a word find the syntax of a word	
11	16	Selecting the appropriate meaning of a word assisted by the example sentences Inferencing of the word spelling Finding grammatical information Finding semantic information	using example sentences as clues for selecting the appropriate meaning of a polysemous word using available information to guess the spelling of a word find the spelling of a word find the meaning	look-up strategies, lemmatization strategies, dictionary awareness strategies
12	17	Inferencing of the reference form Finding grammatical information Finding semantic information Alphabetization	using available information to predict the reference form find the derivatives of a word find word families making assumptions about the correct section of the dictionary to look-up the word by using previous knowledge on word order	lemmatization strategies, dictionary awareness strategies look-up strategies

Table 4: DUS instruction units

Finally, in order to be able to assess the effectiveness of the tasks included in the program for cultivating different DUS and measure achievement, we set the following learning outcomes that users will demonstrate upon successful completion of the program:

- Use dictionaries effectively and be aware of the importance of using them as tools in writing and reading
- Demonstrate awareness of when and how to use a dictionary
- Demonstrate awareness of different types of dictionaries
- Be able to select the appropriate type of dictionary according to the task to be accomplished
- Use dictionaries to find definitions, syllabication, spelling, parts of speech, synonyms, and antonyms
- Be able to alphabetize
- Be able to lemmatize using compensation or inferencing
- Know how to use headwords, example sentences, examples
- Be able to locate synonyms and antonyms
- Be able to select the appropriate meaning of a polysemous word
- Understand etymological information
- Navigate entries to find information about phraseology
- Be able to find and interpret pragmatic information
- Implement the outcomes of look-ups in situations inside or outside class
- Identify the outcome of look-up and start a new one in case it was unsuccessful
- Understand function and working of cross references

7. Concluding Remarks

This paper attempted to illustrate how an instructional model for enhancing dictionary use skills can originate in (cognitive) theory and research and promote classroom activities that are understandable for teachers and dictionary users. It is hoped to lead to further refinements that will expand the proposed intervention program. This is a flexible program that, with the appropriate adjustments, can be adapted for teaching different proficiency levels and in various socio-cultural frameworks. Future research should incorporate into this program activities and tasks for electronic dictionaries, focus on the implementation of the program and its effect in developing dictionary use strategies of younger or older dictionary users and train teachers to incorporate DUS in their teaching. Finally, the learning outcomes presented in section 6 should be matched to different proficiency levels in order to provide a coherent syllabus.

8. References

Anderson, N. J. (2002). *The role of metacognition in second language teaching and learning* (Vol. 4646). Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

Bishop, G. (2000). Developing learner strategies in the use of dictionaries as a productive language learning tool. In *Language Learning Journal,* 22, pp. 58-62

Campoy-Cubillo, M. C. (2015). Assessing dictionary skills. In *Lexicography Asialex*, *2*(1), pp. 119-141.

Carduner, J. (2003). Productive dictionary skills training: what do language learners find useful?. In *Language Learning Journal*, *28*(1), pp. 70-76.

Chadjipapa, E. (2018). *Investigating dictionary use strategies adopted by upper elementary and lower secondary students attending Greek Schools*. Ph.D. Thesis, Democritus University of Thrace. [IN GREEK].

Chadjipapa, E., Gavriilidou, Z., Markos, A., & Mylonopoulos, A. (2020). The effect of gender and educational level on dictionary use strategies adopted by upper-elementary and lower-secondary students attending Greek Schools1. In *International Journal of Lexicography*, *33*(4), pp. 443-462.

Chamot, A. U. (2005). Language Learning Strategy Instruction: Current Issues and Research. In *Annual Review of Applied Linguistics,* 25, pp. 112–130.

Chamot, A. U. (2007). Accelerating academic achievement of English language learners. In J. Cummins & C. Davison (eds.) *International Handbook of English Language Teaching* (pp. 317–31). New York, NY: Springer US.

Chamot, A. U. (2018). Developing self-regulated learning in the language classroom, In I. Walker, D. Kwang Guan Chan, M. Nagami and C. Bourguignon (eds.) *New Perspectives on the Development of Communicative and Related Competence in Foreign Language Education,* Trends in Applied Linguistics, 28, pp. 41-51.

Chamot, A. U., O' Malley, J. M. (1987). The cognitive academic language learning approach: A bridge to the mainstream. In *TESOL quarterly*, *21*(2), pp. 227-249.

Chamot, A. U., O'Malley, J. M. (1994). *The CALLA handbook: Implementing Cognitive Academic Language Learning Approach*. New York: Addison-Wesley Publishing Company.

Chamot, A. U., O'Malley, J. M. (1996). The cognitive academic language learning approach: A model for linguistically diverse classrooms. In *Elementary School Journal,* 96(3), pp. 259-273.

Chamot, A., El-Dinary, P. B. (1999). Children's learning strategies in language immersion classrooms. *The Modern Language Journal*, *83*(3), 319-338.http://dx.doi.org/10.1086/461827

Chen, M. (2007). Learning to learn: the impact of strategy training. In *ELT Journal*, *61*(1), pp. 20-29.

Cohen, A.D. (1998). *Strategies in Learning and Using a Second Language*. London: Longman.

Cohen, A. D. (2000). *Strategies in learning and using a second language*. Beijing: Foreign Language Teaching and Research Press.

Cohen, A. D., Macaro, E. (2007). *Language learner strategies.* Oxford: Oxford University Press.

Cohen, A. D., Weaver, S. J. & Li, T. Y. (1998). The impact of strategies-based instruction on speaking a foreign language. In A. D. Cohen (Ed.), *Strategies in learning and using a second language* (pp. 107– 156). London: Longman.

De Schryver, G. M. (2003). Lexicographers' dreams in the electronic-dictionary age. In *International Journal of Lexicography*, *16*(2), pp. 143-199.

Gavriilidou, Z. (2013). Development and validation of the Strategy Inventory for Dictionary Use (S.I.D.U.). In *International Journal of Lexicography, 22*(2), pp. 135-154.

Gavriilidou, Z., Mavrommatidou, S., & Markos, A. (2020). The effect of gender, age and career orientation on digital dictionary use strategies. In *International Journal of Research*, *9*(6), pp. 63-76.

Gouws, R.H. (2013). Establishing and developing a dictionary culture for specialised lexicography. In *Specialised Lexicography*, (ed.) V. Jesenšek. Berlin: De Gruyter, pp. 51–62.

Gu, Y. (2007). Strategy-based instruction. In *Proceedings of the international symposium on English education in Japan: exploring new frontiers* (pp. 21-38). Osaka: Yubunsha.

Hartmann, R.R.K., James. G. (1998). *Dictionary of lexicography*. London: Routledge.

Hassan, X., Macaro, E., Mason, D., Nye, G., Smith, P., & Vanderplank, R. (2005). Strategy training in language learning: A systematic review of available research. *Research Evidence in Education Library.* Downloaded from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=296.

Herbst, T., G. Stein (1987). Dictionary-using skills: A plea for a new orientation in language teaching, in Cowie, A.P. (ed.) the dictionary and the language learner. In *EURALEX Seminar*, pp. 115-127.

Lew, R. (2013). From paper to electronic dictionaries: Evolving dictionary skills. In D.A. Kwary, N. Wulan & L. Musyahda (eds.), Lexicography and dictionaries in the information age. *Selected papers from the 8th ASIALEX international conference*. Surabaya: Airlangga University Press, pp. 79- 84.

Lew, R., De Schryver, G. M. (2014). Dictionary users in the digital revolution. In *International Journal of Lexicography*, *27*(4), pp. 341-359.

Macaro, E. (2001). *Learning strategies in foreign and second language classrooms: The role of learner strategies.* A&C Black.

Mavrommatidou, S. (2018). Exploring the Frequency and the Type of Users' Digital Skills Using S.I.E.D.U. In *18th EURALEX Proceedings*, Ljubljana, pp. 909-914.

Mavrommatidou, S., Gavriilidou, Z., & Markos, A. (2019). Development and validation of the strategy inventory for electronic dictionary use (SIEDU). In *International Journal of Lexicography*, pp. 1-18. https://doi.org/10.1093/ijl/ecz015

Nguyen, L. T. C., Gu, Y. (2013). Strategy-based instruction: A learner-focused approach to developing learner autonomy. In *Language Teaching Research*, *17*(1), pp. 9-30.

Nkomo, D. (2015). Developing a dictionary culture through integrated dictionary pedagogy in the outer texts of South African school dictionaries: the case of Oxford Bilingual School Dictionary: IsiXhosa and English. In *Lexicography Asialex* 2, pp. 71-99.

O'Malley, J. M., Chamot, A. U. (1990). *Learning strategies in second language acquisition.* Cambridge: Cambridge University Press.

Oxford, R.L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House.

Oxford, R. L. (1993b). Instructional Implications of Gender Differences in Second/Foreign Language (L2) Learning Styles and Strategies. In *Applied language learning*, *4*, pp. 65-94.

Oxford, R. L. (2011). Strategies for learning a second or foreign language. In *Language Teaching*, *44*(2), pp. 167.

Oxford, R. L., Tallbot, V. & Halleck, G. (1990). Language learning strategies, attitudes, motivation, and self-image of students in a university intensive ESL program. In *the annual meeting of International Teachers of English to Speakers of Other Languages*. San Francisco, CA.

Sarafianou, A., Gavriilidou, Z. (2015). The Effect of Strategy-Based Instruction on Strategy Use by Upper-Secondary Greek Students of EFL, *Electronic Journal of Foreign Language Teaching*, Vol. 12, No. 1, pp. 21–34. Available online at http://e-flt.nus.edu.sg/v12n12015/sarafianou.pdf .

Vrettou, A. (2015). Language Learning Strategy Instruction. In Psaltou-Joycey, A. & Z. Gavriilidou (Ed.), *Foreign Language Learning Strategy Instruction: A Teacher's Guide*. Saita publications, Kavala, pp. 32- 48.

Walters, J. (2006). Methods of teaching inferring meaning from context. *RELC Journal*, *37*(2), pp. 176-190.

Walz, J. (1990). The dictionary as a Secondary Source in Language Learning. *The French review,* 64 (1), pp. 79-94.

Wenden, A. L. (1986). Incorporate learner training in the classroom. In *System*, *14*(3), pp. 315-325.

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Papers

Phraseology and Collocation

Οι Φρασεολογισμοί-Κατασκευές της Νέας Ελληνικής Γλώσσας: Μια Λεξικογραφική Προσέγγιση

Onufrieva E.

Κρατικό Πανεπιστήμιο Λομονόσοφ της Μόσχας, Μόσχα, Ρωσία
lisa.onufrieva@gmail.com

Abstract

The paper explores the existing forms of lexicographic description of Modern Greek constructional phrasemes, which are productive phraseological patterns with one or more variable components (empty slots). The analysis of strategies chosen by monolingual Modern Greek dictionaries to describe constructional phrasemes shows the lack of a common approach to presenting phraseological units of this type, as well as the existence of some problems related to idiomatic syntactic constructions in general. One of these problems is the inclusion of lexically variable idiomatic constructions in dictionaries as totally fixed expressions, with their slots filled by some lexemes, and without any reference to their variability. The corpus data suggest that in some cases such way to describe constructional phrasemes does not reflect their real potential and use. The findings of the study indicate the need for developing new approaches towards the description of phraseological units of this type, as well as the need for preparing a special separate dictionary for them.

Keywords: phraseology, constructional phrasemes, phraseological patterns, lexicography

1 Εισαγωγή

Η εμφάνιση νέων θεωρητικών και πρακτικών κατευθύνσεων στη γλωσσολογία τις τελευταίες δεκαετίες άλλαξε τη θέση της φρασεολογίας ως επιστημονικού κλάδου. Όπως επισημαίνει ο Dmitry Dobrovol'skij, στη σύγχρονη γλωσσολογία η φρασεολογία μετατέθηκε «από μια περιθωριακή θέση, στην οποία βρισκόταν για μεγάλο χρονικό διάστημα, στο επίκεντρο του θεωρητικού ενδιαφέροντος» (Dobrovol'skij 2016: 19).

Η αύξηση του ενδιαφέροντος για την φρασεολογία έχει άμεση σχέση με την εμφάνιση μεγάλων σωμάτων κειμένων και τη δημιουργία συστημάτων μηχανικής μετάφρασης, καθώς και με τη συνεχιζόμενη στη γλωσσολογία ανάπτυξη διακλαδικών προσεγγίσεων. Κατά τη δημιουργία των πρώτων συστημάτων μηχανικής μετάφρασης, που βασίζονταν στους γενικούς γραμματικούς κανόνες της γλώσσας, οι ερευνητές αντιμετώπισαν το πρόβλημα της ικανοποιητικής μετάφρασης των συντακτικών κατασκευών, πολλές από τις οποίες παρουσίαζαν, σε κάποιο βαθμό, απόκλιση από τους γραμματικούς κανόνες. Η εξερεύνηση των σωμάτων κειμένων επηρέασε επίσης την εξέλιξη της φρασεολογίας, αφού έδωσε στους ερευνητές τη δυνατότητα επαλήθευσης πολλών καθιερωμένων αντιλήψεων που μέχρι τότε βασίζονταν μόνο στη διαίσθηση των ερευνητών. Έτσι, έγινε εμφανές ότι κάποια είδη φρασεολογικών μονάδων που παραδοσιακά θεωρούνταν «πυρήνας» της φρασεολογίας στην πραγματικότητα έχουν χαμηλή συχνότητα εμφάνισης στα σώματα κειμένων, ενώ η συχνότητα εμφάνισης άλλων γλωσσικών φαινομένων που μέχρι τότε θεωρούνταν «περιθωριακά» είναι πολύ υψηλή (βλ. π.χ. Granger & Paquot 2008: 29, Dobrovol'skij 2016: 12).

Ο Dobrovol'skij αναφέρει χαρακτηριστικά:

> «[Ό]ταν οι γλωσσολόγοι άρχισαν να χρησιμοποιούν ενεργά τα σώματα κειμένων, αποδείχθηκε πως πολλά από τα λεγόμενα *περιθωριακά* γλωσσικά φαινόμενα έχουν πολύ υψηλή συχνότητα εμφάνισης και έτσι, ο σκοπός της περιγραφής τους από περιθωριακός γίνεται βασικός. Αυτό αφορά σε όλη τη φρασεολογία, καθώς και στην πληθώρα ποικίλων ανώμαλων – εν όλω ή εν μέρει – γλωσσικών φαινομένων». (Dobrovol'skij 2016: 12)[1]

Εκτός από τη δυνατότητα της επαλήθευσης θεωρητικών υποθέσεων, τα δεδομένα των σωμάτων κειμένων έδωσαν στους ερευνητές μια πιο πλήρη εικόνα για το βαθμό και τους τύπους της μεταβλητότητας στερεότυπων εκφράσεων (βλ. π.χ. Philip 2008: 95, Fellbaum 2016: 412) και αποτέλεσαν μια παραστατική επιβεβαίωση του ισχυρισμού του John Sinclair ότι πολλές από «τις λεγόμενες "παγιωμένες" φράσεις» στην πραγματικότητα είναι κάθε άλλο παρά παγιωμένες» (Sinclair 2004: 30).

Σημαντικό βήμα για τη σύγχρονη θεωρία της φρασεολογίας υπήρξε και η συνειδητοποίηση του γεγονότος ότι οι περισσότερες σταθερές εκφράσεις αποτελούν επί μέρους πραγματώσεις άλλων, πιο γενικών δομών – πλαισίων (patterns) (Steyer 2015, 2020). Αυτή η ανακάλυψη έθεσε τη φρασεολογία και επομένως και τη φρασεογραφία απέναντι σε μια αλλαγή παραδείγματος.

Όπως υπογραμμίζει η Kathrin Steyer, «η παραδοσιακή επικέντρωση σε πολύ λεξικοποιημένες και συνήθως ιδιωματικές πολυλεκτικές εκφράσεις οδήγησε στην υπερεκτίμηση του μοναδικού τους καθεστώτος στο διανοητικό λεξικό» (Steyer 2015: 295–296). Η ίδια η Steyer υποστηρίζει ότι στην πραγματικότητα η πλειοψηφία των λεγόμενων «παγιωμένων» εκφράσεων αποτελεί διαφορετικές πραγματώσεις ίδιων συντακτικών πλαισίων:

[1] Η μετάφραση δική μας.

«[Ο]ι περισσότερες πολυλεκτικές εκφράσεις αποτελούν τυπικές λεξικές πραγματώσεις συγκεκριμένων πλαισίων (‘πλαισίων πολυλεκτικών εκφράσεων’), τα οποία προέκυψαν ως αποτέλεσμα της επανειλημμένης χρήσης και μπορούν να συμπληρωθούν με συνεχώς εναλλασσόμενα γλωσσικά στοιχεία». (Steyer 2015: 279)[2]

Το λεξικό συμπλήρωμα τέτοιων πλαισίων, κατά την Steyer, μπορεί να είναι τόσο ιδιωματικό, όσο και μη ιδιωματικό (Steyer 2015: 281).

Πλαίσια πολυλεκτικών εκφράσεων αποτελούν και οι λεγόμενοι φρασεολογισμοί-κατασκευές (ΦΚ), εν μέρει λεξικοποιημένα ιδιωματικά συντακτικά σχήματα, τα οποία έχουν λεξική σημασία και ανήκουν στο λεξικό ως λεκτικές μονάδες. Ο όρος «φρασεολογισμοί-κατασκευές»[3], που προτάθηκε από τον Dobrovol’skij 2011: 114), αντικατοπτρίζει μια νέα, διακλαδική προσέγγιση στη μελέτη και την περιγραφή των φρασεολογικών μονάδων.

Η παρούσα εργασία έχει ως στόχο να εξετάσει τους τύπους της λεξικογραφικής περιγραφής των φρασεολογισμών-κατασκευών στα λεξικά της νεοελληνικής γλώσσας και να αναδείξει, χρησιμοποιώντας ως παράδειγμα την περίπτωση του νεοελληνικού φρασεολογισμού *Ούτε Χ να Υ*, ότι πίσω από τη λεξικογραφική περιγραφή κάποιων εκφράσεων ως εντελώς παγιωμένων φρασεολογικών μονάδων στην πραγματικότητα μπορεί να κρύβονται παραγωγικά φρασεολογικά μοντέλα.

2 Η Έννοια των Φρασεολογισμών-Κατασκευών και η Θέση τους στη Σύγχρονη Ελληνική Λεξικογραφία

Οι φρασεολογισμοί-κατασκευές συμπεριλαμβάνονται ως ξεχωριστή τάξη φρασεολογισμών στην κατάταξη των φρασεολογικών μονάδων, προτεινόμενη από τους Anatoly Baranov και Dmitry Dobrovol’skij. Σύμφωνα με τον ορισμό των δύο γλωσσολόγων οι φρασεολογισμοί-κατασκευές αποτελούν «αυτόνομες συντακτικά εκφράσεις σταθερού περιεχομένου, από τις οποίες λείπουν κάποια στοιχεία (ορίσματα – απλά (X / Y) ή προτασιακά (P))» (Baranov & Dobrovol’skij 2014: 88). Υπογραμμίζουν δε ότι τα σταθερά λεξικά στοιχεία των φρασεολογισμών-κατασκευών μαζί με τη σύνταξή τους χαρακτηρίζονται από μια ενιαία σημασία, η οποία είναι «σχεδόν λεξική» (Baranov & Dobrovol’skij 2014: 88).

Βασικό δομικό γνώρισμα των ΦΚ είναι ότι αποτελούνται από ένα σταθερό συστατικό στοιχείο και ένα μεταβλητό συστατικό στοιχείο (κενό slot), το οποίο συμπληρώνεται ανάλογα με το περικείμενο και την πρόθεση του ομιλητή (π.χ. *Αυτό κι αν (δεν) είναι Χ! – Αυτό κι αν (δεν) είναι είδηση / έκπληξη / τύχη!* ή *Δεν πα’ να Ρ – Δεν πα’ να κοιτάν / τον τράβαγε από το μανίκι* / *φώναζε*). Ενώ τα λεξήματα στο σταθερό συστατικό στοιχείο ενός ΦΚ στις περισσότερες περιπτώσεις υφίστανται απώλεια σημασιακού περιεχομένου και χάνουν εν μέρει ή εν όλω τα χαρακτηριστικά του μέρους του λόγου στο οποίο ανήκουν, οι λέξεις που αποτελούν το μεταβλητό συστατικό στοιχείο συνδυάζονται μεταξύ τους σύμφωνα με τους γενικούς γραμματικούς κανόνες της γλώσσας και χρησιμοποιούνται στην κυριολεκτική τους σημασία. Έτσι, οι φρασεολογισμοί αυτού του τύπου εν μέρει αναπαράγονται ως έτοιμες φρασεολογικές μονάδες και εν μέρει παράγονται ως συντακτικές κατασκευές. Λόγω αυτής της ιδιότητας ανήκουν στην ενδιάμεση περιοχή ανάμεσα στη γραμματική και το λεξικό.

Οι ΦΚ αντιστοιχούν σε κάποιο βαθμό στα γλωσσικά φαινόμενα που είναι γνωστά στη διεθνή επιστημονική βιβλιογραφία με τις ονομασίες *formal / lexically open idioms* (Fillmore et al. 1988: 505), *Phraseoschablonen* (Fleischer 1982: 135), *φρασεολογήματα* (Σετάτος 1994: 177–178), *φραστικά μοντέλα* (Συμεωνίδης 2000: 83) κ.α. Ο όρος «φρασεολογισμοί-κατασκευές» εμπνέεται, χωρίς αμφιβολία, από την γενικότερη φιλοσοφία της Γραμματικής Κατασκευών (Construction Grammar) του Charles Fillmore και της Adele Goldberg, αλλά προϋποθέτει και κάποιους επιπλέον περιορισμούς, τους οποίους θέτει αναγκαστικά ο χαρακτηρισμός ενός γλωσσικού φαινομένου ως «φρασεολογισμού» (βλ. π.χ. τους περιορισμούς σε ό,τι αφορά τον βαθμό της μεταβλητότητας και της ιδιωματικότητας, τον αριθμό των σταθερών λεξικών στοιχείων που περιέχει μια φρασεολογική μονάδα, κ.α.).

Η σταθερότητα της δόμησης και της χρήσης των ΦΚ, καθώς και η ιδιωματικότητά τους, μας επιτρέπει να κατατάξουμε τις δομές αυτές στο πεδίο της φρασεολογίας ως μία από τις πολλές κατηγορίες εκφράσεων που χαρακτηρίζονται ως «πολυλεκτικές» (Anastassiadis-Symeonidis et al. 2020). Σημαντικό γνώρισμα των ΦΚ είναι επίσης η εκφραστικότητά τους, αφού χρησιμοποιούνται στον λόγο ως έτοιμα εκφραστικά σχήματα με κάποια συγκεκριμένη πραγματολογική σημασία, καθώς και η συντακτική τους αυτονομία, η οποία μας επιτρέπει να τους διακρίνουμε από ένα άλλο είδος φρασεολογισμών, τους λεγόμενους «γραμματικούς» φρασεολογισμούς (π.χ. *μια και, παρόλο που* κ.α.).

Κάθε φρασεολογισμός-κατασκευή αποτελεί ιδιαιτερότητα μιας συγκεκριμένης γλώσσας και απαιτεί ξεχωριστή ανάλυση. Συνήθως υπάρχουν γραμματικοί και σημασιολογικοί περιορισμοί ως προς τη σύνδεση του σταθερού και του μεταβλητού συστατικού στοιχείου, γι’ αυτό η επιλογή των λέξεων που μπορούν να ενταχθούν στο κενό slot είναι περιορισμένη. Επίσης, οι λέξεις που χρησιμοποιούνται στη θέση του μεταβλητού συστατικού στοιχείου μπορεί να μην έχουν την ίδια συχνότητα: πολλές φορές ο φρασεολογισμός εκδηλώνει σαφείς προτιμήσεις όσον αφορά τις λέξεις με τις οποίες θα μπορούσε να συνεμφανιστεί. Τέτοιες πληροφορίες παρέχει η ανάλυση των δεδομένων από τα σώματα κειμένων.

Ενώ οι φυσικοί ομιλητές της γλώσσας χρησιμοποιούν ενστικτωδώς στο λόγο τους έτοιμα ιδιωματικά συντακτικά μοντέλα, τα οποία σε μερικές περιπτώσεις παρουσιάζουν απόκλιση από τους γενικούς κανόνες της γραμματικής, για τους μη φυσικοί ομιλητές υπάρχει πρόβλημα εντοπισμού των ΦΚ στο λόγο και της σωστής ερμηνείας τους.

Στην παρούσα μελέτη εξετάσαμε το πώς περιγράφονται οι ΦΚ στα εξής μονόγλωσσα λεξικά της νεοελληνικής γλώσσας:

- *Χρηστικό λεξικό της νεοελληνικής γλώσσας* της Ακαδημίας Αθηνών (ΧΛΝΓ 2014)
- *Λεξικό της κοινής νεοελληνικής* του Ινστιτούτου Νεοελληνικών Σπουδών του Αριστοτελείου Πανεπιστημίου

[2] Η μετάφραση δική μας.

[3] Αγγλ. phraseme-constructions ή constructional phrasemes, γερμ. Phrasem-Konstruktionen, ρωσ. frazeologizmy-konstruktsii.

Θεσσαλονίκης (ΛΚΝ 1998)
- *Λεξικό της νέας ελληνικής γλώσσας* του Γεωργίου Μπαμπινιώτη (Μπαμπινιώτης 2005)
- *Λεξικό της λαϊκής και της περιθωριακής μας γλώσσας* του Γεωργίου Κάτου (Κάτος 2016)

Η ανάλυση δείχνει ότι στα λεξικά δεν εντάσσονται οι ΦΚ που εμφανίστηκαν στη νεοελληνική γλώσσα τα τελευταία 10–20 χρόνια, παρόλο που πολλοί από αυτούς χρησιμοποιούνται ευρέως στον προφορικό λόγο, τα μέσα κοινωνικής δικτύωσης και τα μέσα μαζικής ενημέρωσης. Σε ορισμένες περιπτώσεις δεν εντάσσονται στα λεξικά και οι ΦΚ που αποτελούνται μόνο από λειτουργικές λέξεις ή αντωνυμίες, ίσως, εξαιτίας του «μη λεξικού» τους χαρακτήρα.

Κάποιες φορές ένας ΦΚ δεν εντάσσεται στο λεξικό ως σταθερή έκφραση με το σύμβολο ΦΡ., αλλά η σημασία που έχει αυτός ο ΦΚ συμπεριλαμβάνεται στο λεξικό ως μία από τις σημασίες μιας λέξης που συμπεριλαμβάνεται στο σταθερό συστατικό στοιχείο του. Για παράδειγμα, ο ΦΚ *Χ να σου (πε)τύχει* δεν παρουσιάζεται στο λεξικό Μπαμπινιώτη ως σταθερή έκφραση με το σύμβολο ΦΡ., αλλά στο λήμμα του ρήματος 'πετυχαίνω' αναφέρεται ότι αυτό το ρήμα μπορεί να χρησιμοποιείται για δήλωση δυσαρέσκειας ή ειρωνείας. Την ίδια στιγμή, τα παραδείγματα που συνοδεύουν αυτή τη σημασία του ρήματος 'πετυχαίνω' στο λεξικό Μπαμπινιώτη αποτελούν διαφορετικές πραγματώσεις του παραγωγικού μοντέλου *Χ να σου (πε)τύχει (γαμπρός / θέση να σου πετύχει!)* (Μπαμπινιώτης 2005: 1396). Σημαντικό μειονέκτημα αυτού του τρόπου περιγραφής των ΦΚ είναι η έλλειψη κάποιας ένδειξης ότι η σημασία που έχει ένα συγκεκριμένο λέξημα πραγματώνεται από το λέξημα αυτό μόνο μέσα σε ένα συγκεκριμένο φρασεολογικό πλαίσιο. Όμως, σύμφωνα με τις Μαριάννα Μίνη και Αγγελική Φωτοπούλου, στα λεξικά της Νέας Ελληνικής «δεν φαίνεται να συνάγεται αυστηρή τυπολογία […] ως προς το τι είναι φράση παγιωμένη και τι είναι χρήση του ρήματος» (Μίνη & Φωτοπούλου 2009: 500).

Σε περιπτώσεις που οι ΦΚ εντάσσονται στα λεξικά της νεοελληνικής γλώσσας ως παγιωμένες εκφράσεις με το σύμβολο *ΦΡ.* ή *έκφρ.*, οι τρόποι της λεξικογραφικής περιγραφής τους διαφέρουν. Οι ΦΚ μπορεί να συμπεριλαμβάνονται στα νεοελληνική λεξικά στο λήμμα κάποιου λεξήματος που αποτελεί μέρος του σταθερού συστατικού στοιχείου τους, ενώ στη θέση του κενού slot εμφανίζονται αποσιωπητικά. Για παράδειγμα, ο ΦΚ *Τι Χ, τι Υ* εντάσσεται στο «Χρηστικό λεξικό της νεοελληνικής γλώσσας» στο λήμμα της αντωνυμίας '*τι'* με αποσιωπητικά στη θέση των μεταβλητών συστατικών Χ και Υ *(τι… τι…)* και συνοδεύεται από μερικά παραδείγματα, στα οποία τα σύμβολα Χ και Υ αντικαθίστανται από κάποιο συγκεκριμένο λεξικό συμπλήρωμα *(τι σήμερα, τι αύριο / τι πρώτος, τι δεύτερος)* (ΧΛΝΓ 2013: 1591). Κατά τη γνώμη μας, τέτοιος τρόπος λεξικογραφικής περιγραφής των ΦΚ επιτρέπει στον χρήστη του λεξικού να αντιληφθεί ότι η συγκεκριμένη έκφραση αποτελεί παραγωγικό φρασεολογικό μοντέλο με μεταβλητό συστατικό στοιχείο.

Σε άλλες περιπτώσεις οι ΦΚ αναγράφονται στο λεξικό χωρίς αποσιωπητικά στη θέση του μεταβλητού συστατικού στοιχείου τους. Για παράδειγμα, ο ΦΚ *Μη σώσει(ς) και Ρ* συμπεριλαμβάνεται στο Λεξικό Μπαμπινιώτη με τη μορφή του *μη σώσεις*. Το γεγονός ότι αυτός ο ΦΚ είναι στην πραγματικότητα ένα παραγωγικό φρασεολογικό μοντέλο με σημασιακά δεσμευμένο μεταβλητό συστατικό στοιχείο γίνεται αντιληπτό μόνο από το παράδειγμα που έπεται – *μη σώσεις κι έρθεις* (Μπαμπινιώτης 2005: 1731).

Επίσης, οι ΦΚ μπορεί να συμπεριλαμβάνονται στο λεξικό με το μεταβλητό συστατικό στοιχείο τους συμπληρωμένο με κάποια συγκεκριμένη λέξη ή κάποιο συγκεκριμένο συνδυασμό λέξεων. Σ' αυτή την περίπτωση οι ΦΚ μπορεί να εμφανίζονται στο λεξικό είτε στο λήμμα ενός λεξήματος που αποτελεί μέρος του σταθερού συστατικού τους, είτε στο λήμμα μιας λέξης που συμπληρώνει το κενό slot. Είναι φανερό, όμως, ότι αυτός ο τρόπος λεξικογραφικής περιγραφής των ΦΚ δεν επιτρέπει στον χρήστη του λεξικού να καταλάβει ότι πίσω από τις εκφράσεις που παρουσιάζονται στο λεξικό ως παγιωμένες στην πραγματικότητα κρύβονται παραγωγικά φρασεολογικά μοντέλα που χαρακτηρίζονται από υψηλή μεταβλητότητα και μπορούν να συμπληρωθούν με ένα ευρύ φάσμα λέξεων ή συνδυασμών λέξεων.

Έτσι, η ανάλυση των τρόπων με τους οποίους οι νεοελληνικοί ΦΚ παρουσιάζονται στα λεξικά αναδεικνύει την απουσία στην ελληνική λεξικογραφία μιας ενιαίας προσέγγισης στην περιγραφή αυτού του τύπου φρασεολογισμών και μαρτυρά την ύπαρξη ορισμένων προβλημάτων που συνδέονται με τη λεξικογραφική περιγραφή αυτών των παραγωγικών φρασεολογικών μοντέλων. Ένα από αυτά τα προβλήματα είναι η ένταξη των ΦΚ στα λεξικά σε συμπληρωμένη μορφή ως εντελώς παγιωμένων εκφράσεων.

Στη συνέχεια θα εξετάσουμε λεπτομερέστερα την περίπτωση του νεοελληνικού ΦΚ *Ούτε Χ να Υ*, η οποία αποτελεί χαρακτηριστικό παράδειγμα των δυσκολιών που σχετίζονται με τη λεξικογραφική περιγραφή των φρασεολογικών μονάδων αυτού του τύπου.

3 Η Φρασεολογική Κατασκευή *Ούτε Χ να Υ* και οι Υπάρχουσες Λεξικογραφικές Περιγραφές της

Ο νεοελληνικός ΦΚ *Ούτε Χ να Υ* αποτελείται από δύο συστατικά μέρη – ένα σταθερό *(Ούτε / μήτε να)* και ένα μεταβλητό (Χ, Υ), που συμπληρώνεται ανάλογα με το περικείμενο και την πρόθεση του ομιλητή:

(1) *Ούτε ψυχολόγος να ήταν, ούτε ψυχίατρος. Χαρτί και καλαμάρι με ήξερε* (Μουρσελάς Κ. «Βαμμένα κόκκινα μαλλιά»).
(2) *Λέγε, ρε Μιστόκλη, τέτοια πολυλογία, αδερφάκι μου, ούτε βουλευτής να 'σουνα* (Χ. Μίσσιος, «Καλά, εσύ σκοτώθηκες νωρίς»).

Η φρασεολογική δομή αυτή, απ' ό,τι φαίνεται, είναι μια ελλειπτική εκδοχή της συντακτικά πλήρους κατασκευής *Ούτε / Μήτε Χ να Υ, (δεν) θα Ρ*. Με βάση αυτό, τα παραδείγματα (1) και (2) μπορούν να αποκατασταθούν υποθετικά ως εξής:

(1α) *Ούτε ψυχολόγος να ήταν, δεν θα με καταλάβαινε τόσο καλά.*
(2α) *Ούτε βουλευτής να 'σουνα, δε θα μιλούσες τόσο πολύ.*

Έτσι, η ιδιωματικότητα του ΦΚ *Ούτε Χ να Υ* έχει να κάνει με την αφαίρεση κάποιων δομικών στοιχείων από μια πλήρη και εύκολα κατανοητή έκφραση. Γι' αυτόν τον λόγο η αποκωδικοποίηση της σημασίας αυτού του ΦΚ απαιτεί

περισσότερη προσπάθεια εκ μέρους του δέκτη του μηνύματος.
Παρόλο που η κύρια σημασία του συνδέσμου ‘ούτε’ είναι αρνητική, ο ΦΚ *Ούτε X να Y* στην ουσία δεν δηλώνει άρνηση, αλλά, αντίθετα, παρομοιάζει μια κατάσταση, στην οποία αναφέρεται, με κάποια άλλη κατάσταση που περιγράφεται από το όρισμα P.
Παρά την φαινομενικά απλή του δομή, αυτός ο ΦΚ παρουσιάζει δυσκολία για αυτούς που μαθαίνουν την ελληνική γλώσσα ως ξένη. Το πρόβλημα αυτό γίνεται αμέσως φανερό, αν δούμε κάποιες μεταφράσεις του ελληνικού ΦΚ σε άλλες γλώσσες, όπως στο εξής παράδειγμα, όπου η σημασία αυτού του ΦΚ αποδίδεται στα αγγλικά με τη διατήρηση της αρνητικής σημασίας του συνδέσμου ‘ούτε’:

(3) *Ούτε ψυχολόγος να ήταν, ούτε ψυχίατρος. Χαρτί και καλαμάρι με ήξερε* (Μουρσελάς Κ. «Βαμμένα κόκκινα μαλλιά»). *Maybe he wasn't a psychologist or a psychiatrist, but he sure as hell knew me inside out* (Mourselas K. “Red Dyed Hair”, μτφρ.: F. A. Reed).

Η αναζήτηση του ΦΚ *Ούτε X να Y* στα μονόγλωσσα ελληνικά λεξικά μας έδωσε τα εξής συμπεράσματα. Στο «Χρηστικό λεξικό της νεοελληνικής γλώσσας» αυτή η φρασεολογική δομή συμπεριλαμβάνεται στο λήμμα του ουσιαστικού *παραγγελία* με τη μορφή του *Ούτε παραγγελία (να το έκανα / είχα κάνει)* και την ερμηνεία ‘για κάτι που έγινε συμπτωματικά, τυχαία όπως ακριβώς ήθελε ο ομιλητής’ (ΧΛΝΓ 2014: 1219). Στο «Λεξικό της κοινής νεοελληνικής» αυτή η φρασεολογική δομή συμπεριλαμβάνεται επίσης στο λήμμα του ουσιαστικού ‘παραγγελία’ με τη μορφή του *Ούτε παραγγελία να το ’χαμε* και την ερμηνεία ‘για κτ. που συνέβη συμπτωματικά, όπως ακριβώς το επιθυμούσαμε’ (ΛΚΝ). Στο Λεξικό Μπαμπινιώτη η παρούσα φρασεολογική δομή δεν συμπεριλαμβάνεται, ενώ στο «Λεξικό της λαϊκής και της περιθωριακής μας γλώσσας» Γεωργίου Κάτου συμπεριλαμβάνονται τέσσερις διαφορετικές πραγματώσεις της:

- *Ούτε γαμπρός να ντυνόσουν / στολιζόσουν!* στο λήμμα του ουσιαστικού *γαμπρός* με τον ορισμό ‘έκφραση αγανάκτησης σε άντρα που καθυστερεί πολύ, προσέχοντας το ντύσιμό του, και μας αναγκάζει να τον περιμένουμε για μεγάλο χρονικό διάστημα’,
- *Ούτε νύφη να ντυνόσουν / στολιζόσουν!* στο λήμμα του ουσιαστικού *νύφη* με τον ορισμό ‘έκφραση αγανάκτησης προς γυναίκα, που, καθώς προσέχει πάρα πολύ το ντύσιμό της κατά τη διάρκεια που ντύνεται, μας αναγκάζει να την περιμένουμε πάρα πολύ’,
- *Ούτε παραγγελία να το ’χαμε!* στο λήμμα του ουσιαστικού *παραγγελία* με τον ορισμό ‘λέγεται για κάτι που από καθαρή σύμπτωση συνέβη ακριβώς έτσι όπως το επιθυμούσαμε’,
- *Ούτε συνεννοημένοι να ’μασταν!* στο λήμμα του ρήματος *συνεννοούμαι* με τον ορισμό ‘έκφραση θαυμασμού στην περίπτωση που δυο άτομα ενεργούν ταυτόχρονα με τον ίδιο τρόπο, χωρίς προηγουμένως να έχουν συνεννοηθεί’ (Κάτος 2016).

Έτσι, η ανάλυση της λεξικογραφικής περιγραφής του ΦΚ *Ούτε X να Y* δείχνει ότι αυτό το παραγωγικό φρασεολογικό μοντέλο συμπεριλαμβάνεται στα νεοελληνικά λεξικά σε πέντε συμπληρωμένες μορφές χωρίς κάποια αναφορά στη μεταβλητότητά του. Σε όλες τις πέντε περιπτώσεις ο ΦΚ εμφανίζεται στο λεξικό στο λήμμα ενός λεξήματος που αποτελεί μέρος του μεταβλητού συστατικού στοιχείου του (*παραγγελία*, *γαμπρός*, *νύφη*, *συνεννοούμαι*), ενώ οι ορισμοί που δίνονται στα λεξικά για αυτές τις πέντε σταθερές εκφράσεις αποτελούν ερμηνείες επί μέρους πραγματώσεων ενός κοινού για αυτές παραγωγικού φρασεολογικού μοντέλου.

4 Η Φρασεολογική Κατασκευή *Ούτε X να Y* στο Σώμα Κειμένων elTenTen14

Για να αξιολογήσουμε την αποτελεσματικότητα των υπαρχουσών λεξικογραφικών περιγραφών του ΦΚ *Ούτε X να Y* και να καταλάβουμε το βαθμό της μεταβλητότητάς του και τη συχνότητα συγκεκριμένων πραγματώσεών του, χρησιμοποιήσαμε το σώμα κειμένων elTenTen14 (Greek Web Corpus), που περιέχει πάνω από 1,6 δισ. λέξεις και την παρούσα στιγμή αποτελεί το μεγαλύτερο σώμα κειμένων της νεοελληνικής γλώσσας από όλα τα διαθέσιμα.
Συνολικά εντοπίσαμε στο παρόν σώμα κειμένων 1 390 περιπτώσεις χρήσης του ΦΚ *Ούτε X να Y*. Το φάσμα των ρημάτων που μπορούν να χρησιμοποιηθούν ως πυρήνας του προτασιακού ορίσματος αυτού του ΦΚ είναι αρκετά ευρύ, αλλά τη μεγαλύτερη συχνότητα χρήσης έχει το ρήμα ‘είμαι’ (64,82%) και πολύ μικρότερη τα ρήματα ‘έχω’ (8,55%) και ‘κάνω’ (7,20%). Σπανιότερα χρησιμοποιούνται και άλλα ρήματα (‘γράφω’ – 1,73%, ‘παίζω’ – 1,58%, ‘πηγαίνω’ – 1,28%, ‘δίνω’ – 1,20%, ‘ξέρω’ – 1,20%, ‘παίρνω’ – 1,05%, ‘γίνομαι’ – 0,98%, κ.α.).
Στον Πίνακα 1 εμφανίζεται η συχνότητα λεξικών συνδυασμών που σχηματίζουν τα ρήματα του μεταβλητού συστατικού στοιχείου του ΦΚ *Ούτε X να Y*:

Συνδυασμός	**Εμφανίσεις**	**Ποσοστό**
Ούτε συνεννοημένοι να ήμασταν	43	3,09%
Ούτε παραγγελία να (το) έκανα	33	2,37%
Ούτε παραγγελία να (το) είχα	31	2,23%
Ούτε επίτηδες να το έκανα	25	1,80%
Ούτε βαλτός να ήμουν	19	1,37%
Ούτε προφήτης να ήμουν	14	1,01%
Ούτε να το ήξερα	14	1,01%

Table 1: Τα πιο σταθερά ορίσματα του ΦΚ *Ούτε X να Y* στο σώμα κειμένων.

Όπως φαίνεται από τον Πίνακα 1, η συνολική συχνότητα των πραγματώσεων *Ούτε παραγγελία να (το) έκανα* και *Ούτε παραγγελία να (το) είχα*, οι οποίες εμφανίζονται στο «Χρηστικό λεξικό της νεοελληνικής γλώσσας» και το «Λεξικό της κοινής νεοελληνικής», είναι μόνο 4,6% όλων των περιπτώσεων χρήσης του ΦΚ *Ούτε X να Y* στο σώμα κειμένων elTenTen14. Η συχνότητα της πραγμάτωσης *Ούτε συνεννοημένοι να ήμασταν*, η οποία συμπεριλαμβάνεται στο «Λεξικό της λαϊκής και της περιθωριακής μας γλώσσας» του Γ. Κάτου, είναι μόνο 3,09%. Οι εκφράσεις *Ούτε γαμπρός να ντυνόσουν / στολιζόσουν!* και *Ούτε νύφη να ντυνόσουν / στολιζόσουν!*, που εντάσσονται στο λεξικό Γ. Κάτου, δεν εμφανίζονται σ' αυτό το σώμα κειμένων ούτε μία φορά.

Την ίδια στιγμή συναντάμε στο σώμα κειμένων μια πληθώρα άλλων πραγματώσεων αυτού του φρασεολογικού μοντέλου, μερικά παραδείγματα των οποίων δίνονται παρακάτω:

(4) *Νέοι, γέροι, παιδιά στους δρόμους, με μπαλόνια στα χέρια. Ούτε απελευθέρωση να γιορτάζαμε.*
(5) *Βρε Δημήτρη αυτή η κατσαρόλα σου όποτε ερχόμαστε είναι γεμάτη, ούτε ο Ερυθρός Σταυρός να ήσουν.*
(6) […] *τον ήχο από το κορνάρισμα του μικρού μου ανιψιού κάθε φορά που φτάνει σπίτι (ούτε τη νύφη να έφερνε!).*

Σε όλες τις περιπτώσεις χρήσης του στο σώμα κειμένων elTenTen14 ο ΦΚ *Ούτε X να Y* αποτελεί συντακτικά αυτόνομο εκφώνημα, η πραγματική λειτουργία του οποίου είναι να δώσει εμφατική αξιολόγηση μιας κατάστασης που περιγράφεται στο περικείμενο. Ενώ το σταθερό συστατικό αυτού του ΦΚ έχει τη λεξική σημασία της παρομοίωσης ή της υπερβολής, το κενό slot συμπληρώνεται με μια πρόταση που δηλώνει το αξιολογικό πρότυπο, με το οποίο συγκρίνεται η κατάσταση που περιγράφεται στο περικείμενο.

Από την ανάλυση προκύπτει ότι η υπάρχουσα λεξικογραφική περιγραφή του παραγωγικού φρασεολογικού μοντέλου *Ούτε X να Y* καλύπτει μόνο το 8% περίπου του συνόλου των εμφανίσεων αυτού του φρασεολογικού μοντέλου στο σώμα κειμένων elTenTen14. Για το υπόλοιπο 92% είναι αδύνατον να βρει κανείς πληροφορίες στα λεξικά, επειδή, ακόμα και σε συμπληρωμένη μορφή, ο ΦΚ *Ούτε X να Y* εντάσσεται στα λεξικά στα λήμματα τεσσάρων διαφορετικών λεξημάτων που αποτελούν μέρους του μεταβλητού συστατικού στοιχείου του. Θα ήταν προτιμότερη η ένταξη του ΦΚ *Ούτε X να Y* στα λεξικά στο λήμμα του συνδέσμου 'ούτε' (καθώς και 'μήτε'), ο οποίος αποτελεί τη βασική λεξική «άγκυρα» του σταθερού συστατικού στοιχείου αυτού του ΦΚ και είναι και ο πρώτος κατά σειρά. Επίσης, θα ήταν αναγκαίο να υπογραμμιστεί με κάθε δυνατό μέσο η παρουσία ενός μεταβλητού συστατικού στοιχείου στη δομή αυτού του ΦΚ, με τη βοήθεια είτε των αποσιωπητικών *(Ούτε να...)*, είτε της υπογράμμισης *(Ούτε να__)*, είτε ενός συμβόλου *(Ούτε X να Y)*. Επειδή μια λεπτομερής περιγραφή στο λεξικό όλων των λεξημάτων που μπορούν να χρησιμοποιούνται στη θέση του μεταβλητού συστατικού στοιχείου του ΦΚ *Ούτε X να Y* δεν είναι δυνατή, θα ήταν καλό να γίνει ένας σύντομος γενικός χαρακτηρισμός των τύπων του λεξικού συμπληρώματος, με το οποίο μπορεί να χρησιμοποιηθεί αυτός ο ΦΚ. Θα έπρεπε επίσης να αναφερθεί αν υπάρχουν κάποιοι σημασιολογικοί περιορισμοί ως προς τη σύνδεση του σταθερού και του μεταβλητού συστατικού στοιχείου ή έστω κάποιες προτιμήσεις σε ό,τι αφορά τις λέξεις ή τους συνδυασμούς λέξεων με τους οποίους θα μπορούσε να συνεμφανιστεί το συγκεκριμένο ΦΚ. Ως ερμηνεία θα έπρεπε να δοθεί η τυπική σημασία του ΦΚ *Ούτε X να Y*, η οποία καθορίζεται μόνο από το σταθερό συστατικό του, δεν εξαρτάται από το περικείμενο και παραμένει ίδια σε όλες τις περιπτώσεις χρήσης αυτού του παραγωγικού φρασεολογικού μοντέλου.

5 Συμπεράσματα

Η παρούσα εργασία αποτελεί μια προσπάθεια περιγραφής και ανάλυσης των βασικών προβλημάτων που σχετίζονται με τη λεξικογραφική περιγραφή των φρασεολογισμών-κατασκευών (ΦΚ) – παραγωγικών φρασεολογικών μοντέλων με ένα ή περισσότερα μεταβλητά συστατικά στοιχεία (κενά slot).

Η ανάλυση των νεοελληνικών μονόγλωσσων λεξικών υπέδειξε την έλλειψη στη νεοελληνική λεξικογραφία μιας ενιαίας προσέγγισης στην περιγραφή των φρασεολογικών μονάδων αυτού του τύπου. Ένα από τα σημαντικότερα προβλήματα που σχετίζονται με τη λεξικογραφική περιγραφή των νεοελληνικών ΦΚ αποτελεί η ένταξή τους στο λεξικό ως εντελώς σταθερών, «παγιωμένων» εκφράσεων με το κενό τους slot συμπληρωμένο από κάποια λέξη ή κάποιο συνδυασμό λέξεων και χωρίς καμία αναφορά στην ύπαρξη ενός μεταβλητού συστατικού στοιχείου. Τέτοια λεξικογραφική περιγραφή των ΦΚ δεν ανταποκρίνεται στη γλωσσική πραγματικότητα και δεν επιτρέπει στο χρήστη του λεξικού να καταλάβει ότι οι εκφράσεις που παρουσιάζονται στο λεξικό ως παγιωμένες χαρακτηρίζονται από υψηλή μεταβλητότητα και κατέχουν ένα ή περισσότερα slots, τα οποία μπορούν να συμπληρωθούν από ένα ευρύ φάσμα λέξεων ή συνδυασμών λέξεων.

Η ανάλυση της χρήσης του ΦΚ *Ούτε Χ να Υ*, ο οποίος συμπεριλαμβάνεται στα νεοελληνικά μονόγλωσσα λεξικά σε συμπληρωμένη μορφή σε πέντε διαφορετικές εκδοχές, έδειξε ότι η συχνότητα χρήσης των συγκεκριμένων πραγματώσεων αυτού του παραγωγικού φρασεολογικού μοντέλου που συμπεριλαμβάνονται στα λεξικά στην πραγματική γλωσσική επικοινωνία δεν είναι υψηλή, ενώ μερικές από αυτές τις πραγματώσεις δεν εμφανίζονται στο σώμα κειμένων ούτε μία φορά. Το 92% όλων των εμφανίσεων αυτού του φρασεολογικού μοντέλου στο σώμα κειμένων, αποτελούν άλλες πραγματώσεις του, οι οποίες, όμως, δεν μπορούν να ερμηνευτούν με τη βοήθεια των υπαρχουσών λεξικών περιγραφών, επειδή στα λεξικά αυτό το φρασεολογικό μοντέλο συμπεριλαμβάνεται σε συμπληρωμένη μορφή στα λήμματα τεσσάρων διαφορετικών λεξημάτων, τα οποία δεν αποτελούν μέρους του σταθερού συστατικού στοιχείου τους.

Ο σκοπός της έρευνάς μας όμως ήταν ευρύτερος και συμπεριλάμβανε πλήρη θεωρητική και πρακτική μελέτη φρασεολογισμών-κατασκευών της νεοελληνικής γλώσσας, κάτι που ακόμα δεν έχει περιγραφεί από τους γλωσσολόγους. Κεντρικό στοιχείο έρευνας στο πλαίσιο της θεωρητικής φρασεολογίας παραδοσιακά αποτελούσε η περιγραφή και η ταξινόμηση των «κλασικών» φρασεολογικών μονάδων – ιδιωτισμών, συνάψεων, παροιμιών και ρητών, γλωσσικών κλισέ. Πολύ λιγότερη προσοχή έχει δοθεί μέχρι πρόσφατα στα φαινόμενα ιδιωματικότητας στη σύνταξη. Η διεύρυνση των ερευνητικών στόχων της φρασεολογίας και η μετατόπιση στο επίκεντρο των θεωρητικών ενδιαφερόντων της πολυλεκτικών φρασεολογικών πλαισίων αναπόφευκτα θέτουν νέους στόχους στη μελλοντική έρευνα τόσο στην ίδια τη φρασεολογία, όσο και στη φρασεογραφία και απαιτούν την ανάπτυξη νέων προσεγγίσεων στη λεξική περιγραφή των φρασεολογικών μονάδων αυτού του τύπου.

Κύρια θεωρητική βάση της ερευνάς μας αποτελούν μελέτες για τη ρωσική φρασεολογία που έχουν γίνει από Ρώσους γλωσσολόγους στον τομέα της φρασεολογίας (Baranov & Dobrovol'skij), καθώς και αντίστοιχες μελέτες για άλλες γλώσσες, κυρίως τη γερμανική (Fleischer, Steyer). Από αυτές τις μελέτες προκύπτει ότι ο αριθμός τέτοιων φρασεολογισμών είναι συνήθως περιορισμένος (π.χ. στη ρωσική γλώσσα διαφορετικοί ερευνητές εντοπίζουν από 50 έως 90 τέτοιες δομές). Μελετώντας το θέμα των παραγωγικών φρασεολογικών μοντέλων στη νέα ελληνική γλώσσα, έχουμε εντοπίσει προς το παρόν πενήντα ΦΚ στα λεξικά και κείμενα νεοελληνικής λογοτεχνίας. Έτσι, αρχίζει να φαίνεται εφικτή η σύνταξη ενός ειδικού λεξικού που θα εντόπιζε και θα ανέλυε τους ΦΚ της νέας ελληνικής γλώσσας.

Οι ΦΚ αποτελούν αναπόσπαστο κομμάτι της γλωσσικής επικοινωνίας και εντυπωσιακό παράδειγμα εκφραστικότητας στο λόγο. Είναι επίσης και ένα από τα πιο δύσκολα πεδία κατά την εκμάθηση οποιασδήποτε ξένης γλώσσας: το σταθερό συστατικό στοιχείο τους σε μερικές περιπτώσεις αποτελείται μόνο από λειτουργικές λέξεις (κατά κανόνα, όχι λιγότερες από δύο), που δύσκολα γίνονται αντιληπτές, ενώ στη θέση του μεταβλητού συστατικού στοιχείου τους μπορεί να είναι μια ολόκληρη πρόταση. Έτσι, πρέπει να υπογραμμιστεί η αναγκαιότητα μιας συστηματικής περιγραφής των φρασεολογισμών-κατασκευών και η ανάγκη σύνταξης ενός ειδικού λεξικού, κάτι που επιβεβαιώνεται έμμεσα και από τα συμπεράσματα της παρούσας εργασίας.

6 Βιβλιογραφικές Αναφορές

Anastassiadis-Symeonidis, A., Fotopoulou, A. & Kyriacopoulou, T. (2020). Multiword expressions in Modern Greek: synthetic review on their nature. In *Bulletin of Scientific Terminology and Neologisms. Special issue: MWEs in Greek and other languages: from theory to implementation*, p. 15.

Baranov, A.N. & Dobrovol'skij, D.O. (2014). *Osnovy frazeologii (kratkij kurs)* [*Basics of Phraseology (a concise course)*]. Moscow: FLINTA.

Dobrovol'skij, D.O. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch, A. Ziem (Hrsg.) *Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze.* Tübingen: Stauffenburg Linguistik, pp. 110–130.

Dobrovol'skij, D.O. (2016). Grammatika konstrukcij i frazeologija [Construction Grammar and phraseology]. In *Voprosy jazykoznanija*, 3, pp. 7–21.

Fellbaum, Ch. (2016). The treatment of multi-word units in lexicography. In Ph. Durkin (ed.) *The Oxford Handbook of Lexicography.* Oxford: Oxford University Press, pp. 411–424.

Fillmore, Ch., Kay, P. & O'Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. In *Language*, 3(64), pp. 501–538.

Fleischer, W. (1982). *Phraseologie der deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut.

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger, F. Meunier (eds.) *Phraseology. An interdisciplinary perspective.* Amsterdam; Philadelphia: John Benjamins Publishing Company, pp. 27–49.

Greek Web Corpus – *elTenTen14. The Greek Web Corpus.* Προσβάσιμο στο: https://www.sketchengine.eu/eltenten-greek-corpus [18.04.2021].

Κάτος, Γ. (2016). *Λεξικό της λαϊκής και της περιθωριακής μας γλώσσας.* Θεσσαλονίκη: Κέντρο Ελληνικής Γλώσσας. Προσβάσιμο στο: http://georgakas.lit.auth.gr/dictionaries/index.php/anazitisi/g-katou [18.04.2021].

Λεξικό της Κοινής Νεοελληνικής. (1998). Θεσσαλονίκη: Ινστιτούτο Νεοελληνικών Σπουδών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Προσβάσιμο στο: http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides [18.04.2021].

Μίνη, Μ. & Φωτοπούλου Α. (2009). Τυπολογία των πολυλεκτικών ρηματικών εκφράσεων στα λεξικά της Νέας Ελληνικής: όρια και διαφοροποιήσεις. In *Selected Papers from the 18th International Symposium on Theoretical and Applied Linguistics, Thessaloniki 4–6 May 2007.* Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, σσ. 491–503.

Μπαμπινιώτης, Γ. (2005). *Λεξικό της Νέας Ελληνικής Γλώσσας.* Αθήνα: Κέντρο Λεξικολογίας.

Philip, G. (2008). Reassessing the canon. 'Fixed' phrases in general reference corpora. In S. Granger, F. Meunier (eds.)

Phraseology. An interdisciplinary perspective. Amsterdam; Philadelphia: John Benjamins Publishing Company, pp. 95–108.

Σετάτος, Μ. (1994). Φρασεολογήματα και φρασεολογισμοί στην κοινή νεοελληνική. Στο *Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής*, 4(Β), σσ. 167–184.

Sinclair, J. (2004). *Trust the Text. Language, Corpus and Discourse.* London: Routledge.

Steyer, K. (2015). Patterns. Phraseology in a state of flux. In *International Journal of Lexicography*, 28(3), pp. 279–298.

Steyer, K. (2020). Multi-word patterns and networks. How corpus-driven approaches have changed our description of language use. In G. Corpas Pastor, J.-P. Colson (eds.) *Computational Phraseology.* Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 273–295.

Συμεωνίδης, Χ.Π. (2000). *Εισαγωγή στην ελληνική φρασεολογία.* Θεσσαλονίκη: Κώδικας.

Χρηστικό Λεξικό της Νεοελληνικής Γλώσσας. (2014). Αθήνα: Εθνικό Τυπογραφείο.

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Papers

Reports on Lexicographical and Lexicological Projects

New developments in Elexifinder, a discovery portal for lexicographic literature

Kosem I., Lindemann D.

Jožef Stefan Institute, Slovenia
iztok.kosem@cjvt.si, david.lindemann.soraluze@gmail.com

Abstract

In this paper, we present ongoing work on Elexifinder (https://finder.elex.is), a lexicographic literature discovery portal developed in the framework of the ELEXIS (European Lexicographic Infrastructure) project. Since the first launch of the tool, the database behind Elexifinder has been enriched with publication metadata and full texts stemming from the LexBib project, and from other sources. We describe data curation and migration workflows, including the development of an RDF database, and the interaction between the database and Elexifinder. Several new features that have been added to the Elexifinder interface in version 2 are presented, such as a new Lexicography-focused category system for classifying article subjects called LexVoc, enhanced search options, and links to LexBib Zotero collection. Future tasks include getting lexicographic community more involved in the improvement of Elexifinder, e.g. in translation of LexVoc vocabulary, improving LexVoc classification, and suggesting new publications for inclusion.

Keywords: ELEXIS; bibliographical data; metalexicography; lexicographic research

1 Introduction

In 2019, a service called Elexifinder,[1] which enables search for lexicographic research, was made available to the lexicographic community. The tool was developed as part of the European Lexicographic Infrastructure (ELEXIS),[2] a H2020 project funded by the European Commission. Elexifinder has been built using some of the elements of the Event Registry system architecture (Leban et al. 2014; Leban, Fortuna & Grobelnik 2017). The first version included 1,755 publications and 78 videos in 11 different languages, with the majority of publications coming from EURALEX and eLex proceedings. A detailed presentation of this version, including the features available at the time, was made by Kosem and Krek (2019).

At the time of Elexifinder launch, there were already plans in motion to improve it even further. The planned improvements were related to the contents, interface and data preparation workflow. For example, an extensive list of publications to be added to Elexifinder was already compiled at the time, and has been regularly updated since. The way article full texts and publication metadata were collected and recorded was far from optimal, and did not utilize existing sources well enough.

In parallel, at University of Hildesheim, the LexBib project was planned, with the goal of creating a domain ontology and digital bibliography of Lexicography and Dictionary Research. First steps in that project are described in Lindemann, Kliche & Heid (2018); the LexBib metadata and full text collection was put together and made accessible using the Zotero platform,[3] and workflows for a conversion of publication metadata to RDF Linked Data were explored.

Despite slightly different aims (ELEXIS being focussed more on providing a tool to efficiently find the relevant publication, and the LexBib Zotero collection having a more bibliographic focus, that is, to provide validated metadata for the purpose of citations), both initiatives had a great deal of overlap. Therefore, it made perfect sense to merge the efforts and develop a workflow that would benefit both purposes and do away with unnecessary duplication of effort, especially in terms of obtaining publications and recording their metadata.

This paper presents various improvements, some considerable, that have been made to Elexifinder since 2019. We start by looking at the backend, presenting the development of an RDF database, and the interaction between the database and eLexifinder. Next, the new LexVoc SKOS vocabulary of subject headings is presented, including the linking of content-describing terms to publication metadata. Then, we present the new contents added, and the improvements made to the Elexifinder interface. The paper concludes by outlining future plans, such as the translation of LexVoc vocabulary, and the inclusion of more publications.

2 Bibliographical data

2.1 Sources and workflows

[1] Accessible at https://finder.elex.is .
[2] The project homepage is accessible at https://elex.is.
[3] Accessible at https://lexbib.org/zotero.

For the first version of Elexifinder, publication metadata were stored in spreadsheet files, along with manually revised plain text versions of the corresponding full texts. In addition, publication metadata was enhanced with the location of the first author, by manual annotation. We have merged that data with the data present at that time in the LexBib Zotero collection, defined priorities for the inclusion of further bibliographic items, and established a workflow for data migration from Zotero to Elexifinder (see workflow schema in Figure 1).

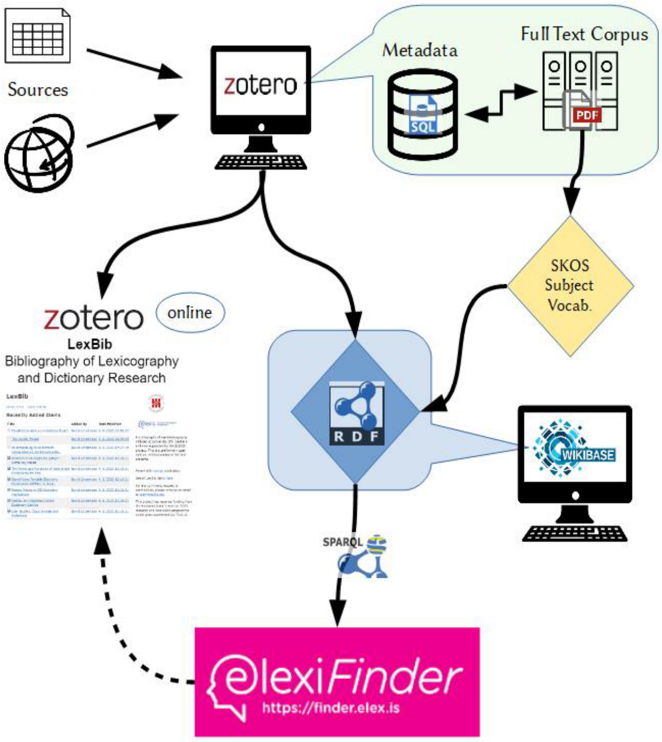


Figure 1: Workflow and data migration scheme

At present, all articles of major journals in the discipline of lexicography, and EURALEX and eLex conference series, a set of edited book volumes, and a set of presentation videos are part of our collection (see complete reference in section 4.1 below). We also have started to integrate bibliographical records stemming from the EURALEX bibliography, compiled by A. Dykstra with the support of individual community members,[4] and OBELEX-meta (Möhrs 2016).[5] Since the enhancement of publication metadata by computational processing of full texts is one of our goals, we have given priority to Open Access publications; we also recorded publications where the full text was accessible due to suitable license agreements. Our copies of full texts are not publicly accessible; they remain restricted for the described text mining purposes in the framework of the project. Through the Zotero platform, and in the Elexifinder tool, we provide download links that lead to the publisher site, which are restricted according to the applicable license.

All publication metadata records have been manually validated. In nearly all cases, the metadata sets resemble all categories necessary for citations, and in most cases, also article abstracts. The Zotero platform allows an export of single items or item batches in several citation styles, or as a structured dataset, e.g. in the bibtex format. Wherever possible, i.e. where that information has been available in the full texts themselves, or journal issue or edited volume back matters, we have manually included the location of the first author. That is now the case for around 98% of the approximately 7,000 bibliographic items included in the LexBib collection.

2.2 Author disambiguation

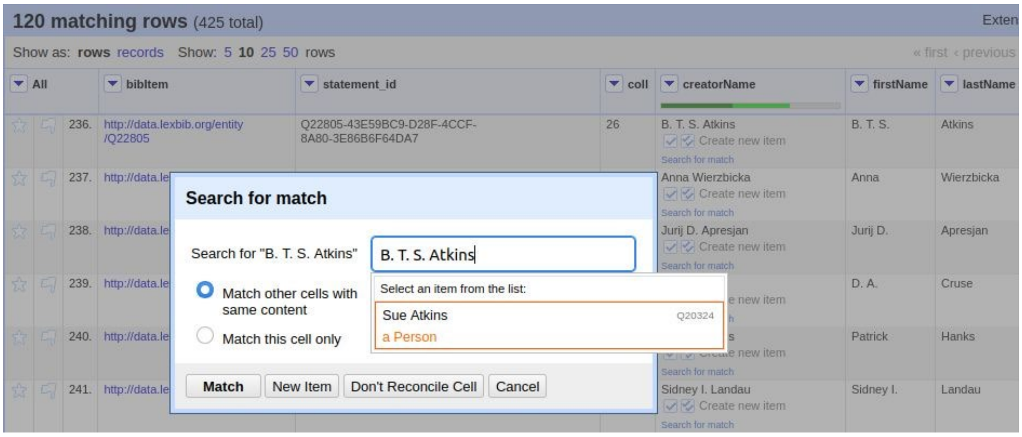
The most challenging task in the curation workflow of bibliographical data is, beyond any doubt, the disambiguation of author and editor name variants. In our collection, some persons appear with up to six different name variants, which is due to capitalization of first names, the inclusion of middle names, hyphenation of double first or last names, alternative spellings, last name ordering errors, spelling errors, etc. Around 18,000 author/editor statements in our data, which presented around 5,000 different names, have been reduced (disambiguated) to around 4,000 persons.

The conversion of literal author/editor values (i.e. name strings) to statements that point to disambiguated person items, where name variants of the same person come together, is the necessary step for allowing searches and search result displays

[4] Accessible at http://euralex.pbworks.com/w/page/7230036/FrontPage.

[5] Accessible at https://www.owid.de/obelex/meta.

that involve all the articles written by that person, regardless of the name variants stated in the bibliographical records, and for linking of bibliographical records to metadata concerning the authors and editors. In other words, that step is what converts metadata consisting of literal values (as in Zotero, or library records in MARC standard) to Linked Data.
A widely used standard for representing Linked Data is RDF.[6] Statements, such as e.g. the links from a bibliographic item to its creators, are represented as semantic triples. We have adapted an existing Zotero RDF export script for our use case, and migrated all publication metadata to an RDF triple store. We have clustered name variants belonging to identical persons, using the OpenRefine software application,[7] so that, for updates of the collection, the recorded name variants of a person are considered; for updates, we are using the same application (see a screenshot detail in Figure 2).


120 matching rows (425 total)
Exten
Show as: rows records Show: 5 10 25 50 rows
« first ‹ previous

All	biblItem	statement_id	coll	creatorName	firstName	lastName
236.	http://data.lexbib.org/entity/Q22805	Q22805-43E59BC9-D28F-4CCF-8A80-3E86B6F64DA7	26	B. T. S. Atkins Create new item Search for match	B. T. S.	Atkins
237.	http://data.le			Anna Wierzbicka Create new item Search for match	Anna	Wierzbicka
238.	http://data.le			Jurij D. Apresjan Create new item Search for match	Jurij D.	Apresjan
239.	http://data.le			e new item	D. A.	Cruse
240.	http://data.le			s e new item Search for match	Patrick	Hanks
241.	http://data.le			Sidney I. Landau Create new item Search for match	Sidney I.	Landau

Search for match

Search for "B. T. S. Atkins" B. T. S. Atkins

● Match other cells with same content
○ Match this cell only

Select an item from the list:
Sue Atkins Q20324
a Person

Figure 2: Auhor disambiguation in the OpenRefine application.

In addition to author disambiguation, generally speaking, the choice of a database that stores semantic triples as central data repository[8] allows us to extend the bibliographical database towards a knowledge graph that involves all relevant kinds of entities (see Lindemann, Klaes & Zumstein 2019). The knowledge graph includes an enrichment of entities with data harvested from other resources in the Linked Open Data Cloud, and/or the definition of links to entities available in these. For the time being, we concentrate on a disambiguation of natural persons, and on defining links between bibliographical records and content-describing terms (see next section). Furthermore, SPARQL database queries[9] provide a straightforward way to obtain structured datasets from RDF data in custom formats, such as the JSON export needed as ingest for the Elexifinder tool. At the same time, the SPARQL endpoint of our RDF database[10] allows the community to access the LexBib data for other research purposes.

3 Subject indexation of metalexicographical literature with LexVoc

We have started to develop LexVoc, a controlled vocabulary of Lexicography-related terms that shall be used as content descriptors, and linked to the corresponding bibliographical items. We have defined English preferred and alternative lexicalizations, and represented relations between terms according to the W3C SKOS standard,[11] that is, following a widely used practice,[12] and, at the same time, choosing a format that can be straightforwardly loaded into our RDF database. Sources for the controlled vocabulary have been the following:

1) An updated and extended version of the index of "Bibliografía Temática de la Lexicografía" (Córdoba Rodríguez 2003),[13] translated to English.
2) The typology of dictionaries by Engelberg and Storrer (2016).
3) The glossary of lexicographic terms by Kipfer (2013).
4) The index of the volume "Using Online Dictionaries" (Müller-Spitzer 2014).

We have defined relations between terms stemming from sources (1) to (4), so that terms can be represented as nodes in a graph, with SKOS relations as arcs. In the second step, we have extended the vocabulary with a manually revised subset of salient term candidates,[14] extracted from a corpus compiled using all English full texts present in the collection used for Elexifinder version 2 (see section 4.1 for full reference).
We are currently performing experiments for extending the vocabulary further, using term extraction results from subsets of our English full texts, e.g. of recent publications about electronic lexicography, in order to cover specialized state-of-

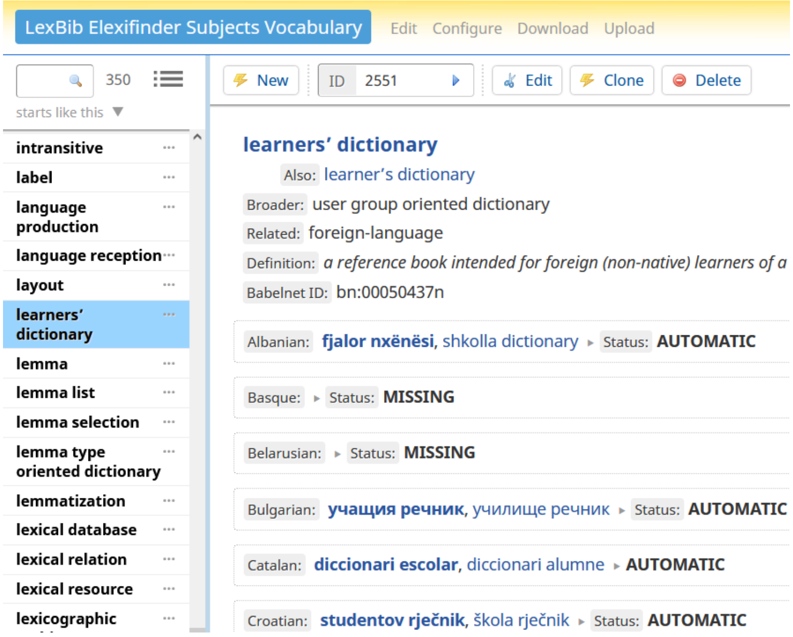
[6] See https://www.w3.org/RDF/.
[7] See https://openrefine.org/.
[8] We have employed Ontotext GraphDB and Wikibase database solutions.
[9] See https://www.w3.org/TR/rdf-sparql-query/.
[10] The SPARQL endpoint is available at https://lexbib.elex.is/query/sparql.
[11] See https://www.w3.org/2004/02/skos/.
[12] See, for example, the Library of Congress Subject Headings vocabulary, accessible at https://id.loc.gov/authorities/subjects.html.
[13] Accessible at https://www.udc.es/grupos/lexicografia/bibliografia/tematica.html.
[14] Salience is calculated according to a TF/IDF measure, in this case with EnTenTen18 as reference corpus. We have used the Sketch Engine for corpus compilation and processing (see https://www.sketchengine.eu/).

the-art terminology. The vocabulary can be explored in a constantly updated graph view,[15] and accessed using SPARQL. In addition to that vocabulary, we also use multilingual terms that denote natural languages as content-describing terms, i.e. to make the object language(s) of a research article explicit. We have obtained these terms from Wikidata.[16]

For automatic extraction of full text bodies from PDF documents, we have used the GROBID tool (Romary & Lopez 2015).[17] In subcollections where the GROBID default algorithm fails to isolate the text body from headers, footers, title, abstract, author affiliation data, and references, etc., we have resorted to manual cleaning of PDF plain text versions produced with standard 'pdf2txt' tools.[18]

LexVoc is now at a stage of development that allows first experiments of automatic indexation of articles. For this, we have performed a discovery of vocabulary terms in lemmatized full texts ("gazetteer approach"). Information on single terms, frequency data, and the bibliographic items they are associated to as content descriptors is available using SPARQL.[19] We are very interested in feedback regarding the structure of the vocabulary, and in suggestions for further sources to be included. We have set up a dedicated discussion group on the LexMeet platform.[20]

It is our goal to obtain a multilingual version of the LexVoc vocabulary in order to apply the indexation process to non-English text, on the one hand, and to provide to the users localized terms as search criteria on the other, that is, to enable users to search in their preferred language for articles indexed with certain terms, regardless of the text language. We want to cover as many languages as possible, but will be first focussing on the languages official in the countries of ELEXIS partners and observers.



intransitive
label
language production
language reception
layout
learners' dictionary
lemma
lemma list
lemma selection
lemma type oriented dictionary
lemmatization
lexical database
lexical relation
lexical resource
lexicographic

learners' dictionary
Also: learner's dictionary
Broader: user group oriented dictionary
Related: foreign-language
Definition: *a reference book intended for foreign (non-native) learners of a*
Babelnet ID: bn:00050437n

Albanian: **fjalor nxënësi**, shkolla dictionary ▸ Status: **AUTOMATIC**
Basque: ▸ Status: **MISSING**
Belarusian: ▸ Status: **MISSING**
Bulgarian: **учащия речник**, училище речник ▸ Status: **AUTOMATIC**
Catalan: **diccionari escolar**, diccionari alumne ▸ **AUTOMATIC**
Croatian: **studentov rječnik**, škola rječnik ▸ Status: **AUTOMATIC**

Figure 3: Lexonomy screenshot

The translation process of LexVoc will be carried out using the Lexonomy application,[21] where we have set up a set of XML-based multilingual dictionaries, one for each language, the lemmata of which are the English SKOS vocabulary terms. Contributors will access an editing form, where translation equivalents can be filled in or modified, and annotated with a status, as shown in Figure 3. During the editing process, the contents of those bilingual dictionaries will be regularly mirrored to a single multilingual resource that can be accessed by the interested public.

To facilitate the translation task, we have extracted translation candidates using the BabelNet[22] API (term labels with status "automatic" in Figure 3), so that the task for contributors consists of validating candidate translation equivalents, or providing translations from scratch, where no translation equivalent candidate could be provided. The search for contributors has been started at the time of writing this paper.

[15] Accessible at https://lexbib.elex.is/wiki/LexVoc.
[16] Multilingual labels of Wikidata items instances of class Q33742, "natural language".
[17] See https://grobid.readthedocs.io/en/latest/.
[18] The GROBID tool is trained on standardized research paper formats, as found in journals and proceedings. Book chapters are often not correctly parsed, as they usually present a different structure. Thus, for book chapters, we chose the manual approach. Different text encodings found in PDF are also problematic: Diacritics and other special characters may not be correctly recognized by standard tools.
[19] See https://lexbib.elex.is/wiki/Project:SPARQL/examples.
[20] Accessible at https://meet.elex.is/groups/lexicographic-concepts-vocabulary.
[21] Accessible at https://lexonomy.elex.is.
[22] See BabelNet homepage at https://babelnet.org/.

4 Elexifinder version 2

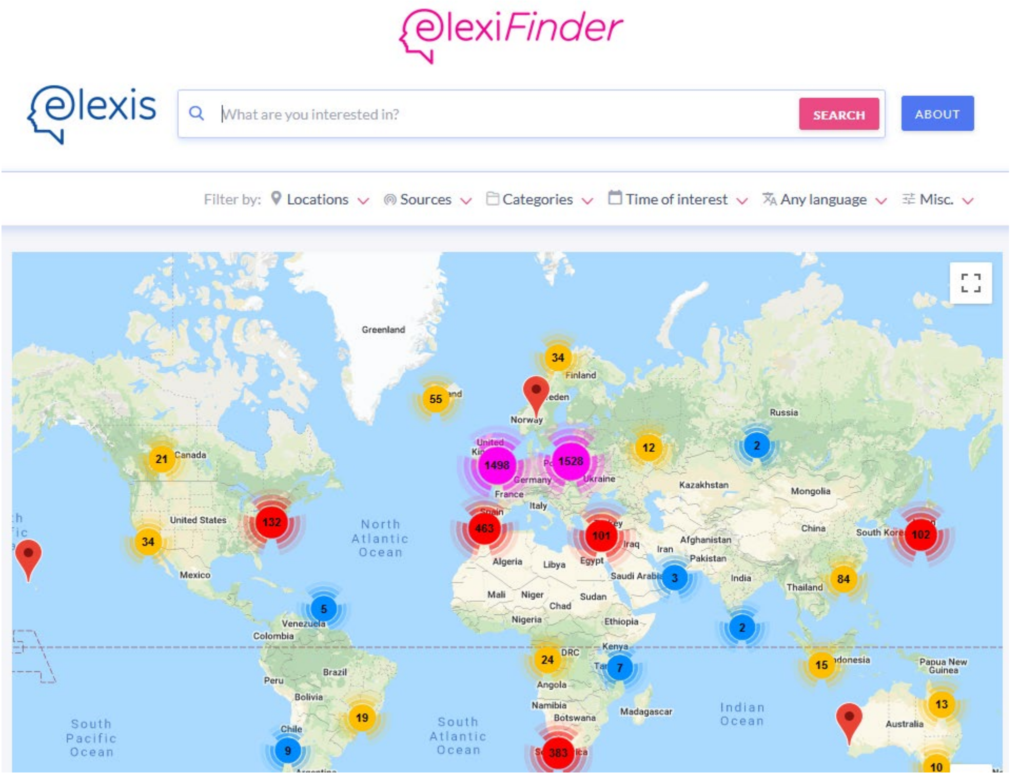


Figure 4: Elexifinder search portal at https://finder.elex.is

4.1 Contents

The most recent version of Elexifinder, updated in early 2021, contains 6,482 publications and 86 videos; in other words, 4,727 publications and 8 videos were added since version 1. The main share of new publications has come in the form of journal papers, and chapters in collective volumes and handbooks. The full list of the contents is as follows:[23]

- Conference proceedings (2,018 articles):
  - eLex (2009-2019)
  - Euralex (1983-2020)
  - Globalex (2016-2020)
- Journals (4,145 articles):
  - International Journal of Lexicography (1988-2019)
  - Lexikos (1991-2019)
  - LexicoNordica (1994-2017)
  - Lexicographica (1985-2019)
  - Nordiske Studier i Leksikografi (1992-2018)
  - Lexicon (1995-2018)
  - Asialex (2014-2020)
  - Slovenščina 2.0 (2013-2020)
  - Revista de Lexicografía (1994-2019)
- Collective Volumes and Handbooks (319 articles):
  - Gouws et al. (eds.). 2013. Dictionaries. An International Encyclopedia of Lexicography (HSK 5/4).
  - Teubert (ed.). 2007. Text Corpora and Multilingual Lexicography.
  - Fuertes Olivera (ed.). 2010. Specialised Dictionaries for Learners.
  - Müller-Spitzer. 2014. Using Online Dictionaries.
  - Jackson (ed.). 2013. Bloomsbury Companion to Lexicography
  - Braun et al. (eds.). 2003. Internationalismen II: Studien zur interlingualen Lexikologie und Lexikographie.
  - Domínguez Vázquez et al. (eds.). 2014. Zweisprachige Lexikographie zwischen Translation und Didaktik.
  - Domínguez Vázquez et al. (eds.). 2014. Lexicografía de las Lenguas Románicas.
  - Gottlieb & Mogensen (eds.). 2007. Dictionary Visions, Research and Practice.
  - Wiegand (ed.). 2000. Wörterbücher in der Diskussion IV: Vorträge aus dem Heidelberger Lexikographischen Kolloquium.
  - Wiegand (ed.). 2002. Perspektiven der pädagogischen Lexikographie des Deutschen II, Untersuchungen anhand

[23] Full reference of edited volumes is found on Zotero (http://lexbib.org/zotero).

    des »de Gruyter Wörterbuchs Deutsch als Fremdsprache«.
  - Sterkenburg (ed.). 2003. A Practical Guide to Lexicography.
- Videos (86 items):
  - eLex 2011
  - Euralex 2018
  - WNLex workshop (2018)
  - 15 lexicographic presentations in Slovenia (in Slovene and English)

This dramatic increase in the number of publications has been accompanied by a similar increase in the number of languages represented in Elexifinder. It now contains publications written in 20 different languages:

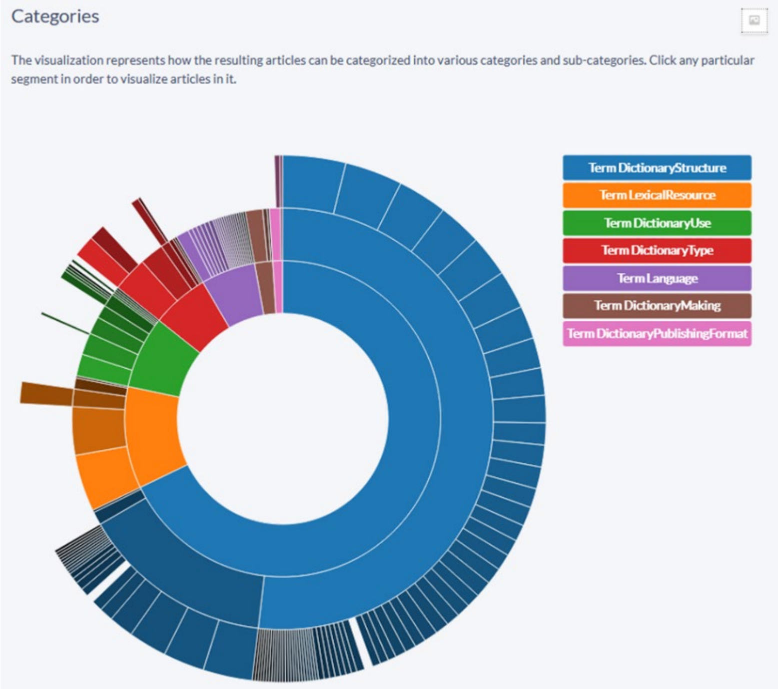
- English: 3729
- German: 829
- Danish: 436
- Spanish: 390
- Swedish: 362
- French: 195
- Norwegian Bokmål: 178
- Afrikaans: 136
- Slovene: 81
- Italian: 40
- Nynorsk: 36
- Portuguese: 12
- Russian: 10
- Dutch: 7
- Modern Greek: 5
- Catalan: 4
- Belarusian: 1
- Finnish: 1
- Croatian: 1

English still dominates as the language of articles and videos. However, the distribution of publications per language has become more balanced. Most notably, in version 1, articles and videos in English represented 84% of total contents, while in version 2 the share has dropped to 57%.

4.2 Interface

Improvements have also been made to the Elexifinder interface, both in terms of functionality and user-friendliness. The two major upgrades are related to the improved author disambiguation workflow and subject indexation, mentioned in Sections 2 and 3 respectively. The author disambiguation procedure, which for Elexifinder allows the selection of one name representation for all author name variants, means that it is now much more straightforward to find all publications of a certain author. In version 1, this was done by using the Sources/Authors filter option and selecting all relevant name variants; in version 2, not only does the user select only one name, it is now also possible to obtain the list of publications by entering/finding the author's name in the main search window.

Elexifinder offers a category-based system search, which allows to browse articles according to certain content-describing terms, and/or to perform cascaded searches, i.e. to filter sets of displayed search results. In version 2, DMOZ all-domain categories, which were initially used as a temporary solution, have been replaced by the lexicography-oriented controlled vocabulary described in Section 3 above (see Figure 5 for a display of categories relevant in articles authored or co-authored by Patrick Hanks).



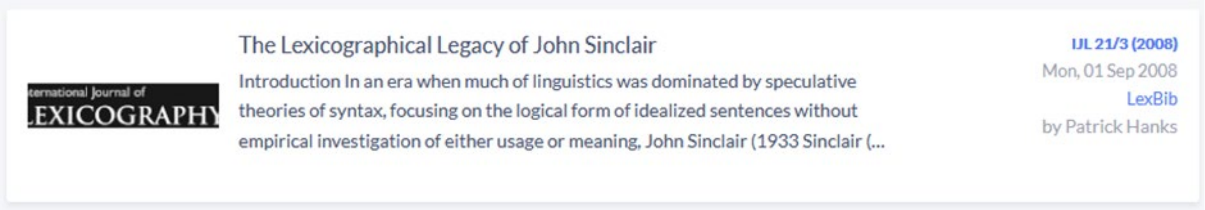
Categories

The visualization represents how the resulting articles can be categorized into various categories and sub-categories. Click any particular segment in order to visualize articles in it.

Figure 5: Categories relevant for author Patrick Hanks

With publications in many languages now represented in Elexifinder - and more languages to be added - it becomes necessary for the users to be able to find research publications on a topic of interest in all the languages. This need is addressed by a newly added cross-linguistic searching via concepts, which are identified by wikification, "a process of entity linking that uses Wikipedia as the knowledge base" (Leban, Fortuna & Grobelnik 2016). In other words, publications in different languages dealing with similar topics are indexed, and found, with the same concepts (in English), provided that the concepts and their translations are found on Wikipedia. While the LexVoc subject vocabulary (see Section 3) is being translated, this approach already offers an immediate solution for multilingual information retrieval.

Another search-related improvement is the option of searching the publications and videos published in a certain source. Using the search by group in the Sources/Sources filter, the users can now for example limit their search to only eLex conference proceedings or Lexikos journal contents, etc.



The Lexicographical Legacy of John Sinclair

Introduction In an era when much of linguistics was dominated by speculative theories of syntax, focusing on the logical form of idealized sentences without empirical investigation of either usage or meaning, John Sinclair (1933 Sinclair (...

IJL 21/3 (2008)
Mon, 01 Sep 2008
LexBib
by Patrick Hanks

Figure 6: Elexifinder bibliographic item display

An important addition to the result display is the link for each publication to its corresponding entry in the LexBib collection on Zotero (see Figure 6). The decision to add this link was made in order to meet the needs of users who require the complete metadata of a publication, for example for citing purposes.

Lastly, each publication now also comes with a source logo image (see Figure 6). We have decided to add this feature in order to make the sources of publications more easily and quickly identifiable. This conveniently complements the information of the source in the top right corner of the item display, where the name of the source, including the year of publication and issue if relevant, has to be provided in a relatively short form due to space constraints. This abbreviated source information is, at the same time, a hyperlink that leads to the website associated with the source item.

5 Conclusions and Outlook

Elexifinder has come a long way in terms of contents and functionality; however, there is still much to be done. In addition to adding more publications and making the tool even more user-friendly, immediate challenges to be addressed are dissemination among the lexicographic community, and sustainability.

As the majority of major lexicographic conferences and journals, which mainly contain papers in English, are now found in Elexifinder,[24] we have shifted our focus to acquiring (open access) lexicographic research in other languages, and to lexicographically-relevant research published in volumes or presented at conferences with a scope that is not strictly lexicographic. As identifying such publications is a difficult task, we have asked members of the lexicographic community for help, with initially turning to the lexicographers and researchers coming from ELEXIS partner and observer institutions, in order to test the approach. For the submission of suggestions for new content, and related discussions, we have decided to use LexMeet,[25] a newly developed platform for the lexicographic community to discuss issues, ideas, project results, collaborations etc.

As mentioned in Section 3, we have also launched a project of translating the controlled vocabulary of English terms. Contributors will be using the Lexonomy dictionary writing software on the ELEXIS website for editing translation equivalents. The LexMeet platform will be used for discussions, either general ones by the entire contributor group, or language-specific by contributors working on a specific language. The multilingual dataset will also be publicly accessible throughout the translation process, and when finalized, published in a CLARIN repository.

As far as the dissemination is concerned, a longer and global dissemination campaign is on the way, with the aim of making the community aware of the tool, and helping them with using it. At the same time, we will aim to obtain as much feedback as possible, for example by using an online survey, in order to identify further ways in which to improve Elexifinder.

Finally, it is important to dedicate some time to thinking about sustainability of the tool and its contents, especially as the end of ELEXIS project is drawing near. The greatest sustainability challenge is connected to contents; we need to find a way to keep them regularly updated. One of the solutions currently explored is to make the Elexifinder workflow a community-driven project, where lexicographers not only submit requirements or new bibliographical datasets, but also divide the task of data curation, and upload to Elexifinder.

6 References

Córdoba Rodríguez, F. (2003). *Bibliografía Temática de La Lexicografía*. A Coruña: Universidade da Coruña.

Engelberg, S., Storrer, A. (2016). Typologie von Internetwörterbüchern und -portalen. In A. Klosa & C. Müller-Spitzer (eds.) *Internetlexikografie. Ein Kompendium*. Berlin/New York: de Gruyter, pp. 31–63.

Kipfer, B.A. (2013). Glossary of Lexicographic Terms. In H. Jackson (ed.) The Bloomsbury Companion to Lexicography. London: Bloomsbury, pp. 391–406.

Kosem, I., Krek, S. (2019). ELEXIFINDER: A Tool for Searching Lexicographic Scientific Output. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, 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. Brno: Lexical Computing CZ s.r.o., pp. 506–18.

Leban, G., Fortuna, B., Brank, J. & Grobelnik, M. (2014). Event Registry: Learning about World Events from News. In *Proceedings of the 23rd International World Wide Web Conference, WWW14, Seoul, Korea, April 7-11, 2014*, pp. 107–10.

Leban, G., Fortuna, B. & Grobelnik, M. (2016). Using News Articles for Real-Time Cross-Lingual Event Detection and Filtering. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval (NewsIR16), Padova, Italy, March 20, 2016*, pp. 33–38.

Leban, G., Fortuna, B. & Grobelnik, M. (2017). Event Extraction from Media Texts. In C. Sammut & G.I. Webb (eds.) Encyclopedia of Machine Learning and Data Mining. Boston, MA: Springer US, pp. 416–22.

Lindemann, D., Klaes, C. & Zumstein, P. (2019). Metalexicography as Knowledge Graph. Edited by Maria Eskevich, Gerard De Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek & Milan Dojchinovski. *Open Access Series in Informatics* 70.

Lindemann, D., Kliche, F. & Heid, U. (2018). Lexbib: A Corpus and Bibliography of Metalexicographcal Publications. In *Proceedings of EURALEX 2018*. Ljubljana, pp. 699–712.

Möhrs, C. (2016). Online Bibliography of Electronic Lexicography. The Project OBELEXmeta. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the 17th EURALEX International Congress: Lexicography and Linguistic Diversity*. Tbilisi: Tbilisi State University, pp. 906–9.

Müller-Spitzer, C., ed. (2014). *Using Online Dictionaries*. Lexicographica Series Maior 145. Berlin: De Gruyter.

Romary, L. & Lopez, P. (2015). GROBID - Information Extraction from Scientific Publications. *ERCIM News, Scientific Data Sharing and Re-Use* 100 (January).

Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The authors wish to thank Christiane Klaes from University of Hildesheim, who in the framework of her MA thesis project has defined the algorithm used for author name variant clustering (see section 2).

[24] One notable omission in the existing dataset is Dictionaries, a journal of the Dictionary Society of North America, but we already have its whole content in our database and will be adding it with the next Elexifinder update.

[25] Accessible at https://meet.elex.is/.

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Posters

EURALEX XIX
Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Posters

Lexicography and Corpus Linguistics
Lexicography and Language Tech

Crowdsourcing Pedagogical Corpora for Lexicographical Purposes

Zingano Kuhn T., Todorović B.Š., Holdt Š.A., Zviel-Girshin R., Koppel K., Luís A.R., Kosem I.

CELGA-ILTEC/University of Coimbra, University of Belgrade/Faculty of Philology, Centre for Language Resources and Technologies/University of Ljubljana, Ruppin Academic Centre, Institute of the Estonian Language, CELGA-ILTEC/University of Coimbra, Centre for Language Resources and Technologies, University of Ljubljana/
Jožef Stefan Institute
tanarazingano@outlook.com, branislava.sandrih@fil.bg.ac.rs, spela.ArharHoldt@ff.uni-lj.si, rinazg@gmail.com, kristina.koppel@eki.ee, aluis@fl.uc.pt, iztok.kosem@cjvt.si

Corpora are valuable sources for the development of language learning materials (e.g., books, grammars, dictionaries, exercises), because they contain language as produced in natural contexts. Even though corpora are getting larger, mainly due to crawling data from the web, their pedagogical use remains rather challenging. Not all texts are appropriate for language learning or teaching purposes as they can potentially contain sensitive or offensive content, in addition to exhibit structural problems, errors, among other problems. Corpus cleaning for pedagogical purposes is however a very time-consuming task if done manually. In this paper we present a new and more effective method for creating problem-labelled pedagogical corpora for a group of languages, namely Portuguese, Serbian, Slovene, Dutch and Estonian, by means of crowdsourcing. First, we report on an experiment aimed at verifying the adequacy of crowdsourcing as a technique for corpus labelling. We then outline the lessons learned and discuss how these have led us to explore an alternative way of compiling pedagogical corpora through gamification.

Keywords: corpus creation; good example sentences; pedagogical corpora; crowdsourcing

1 Introduction

Corpora have been widely used for the development of language learning material, including learners' dictionaries, and other pedagogical resources. This is no surprise, since corpora show how language is authentically used in everyday life and thus provide valuable information for the learners' own language development. Römer (2009), Boulton (2017), and Vyatkina and Boulton (2017), to name but a few, have pointed out an impressive number of publications on corpus use for pedagogical purposes. Corpora that are built specifically for language learning purposes are usually called pedagogical corpora: "The pedagogical corpus, as its name suggests, is primarily intended to serve as a resource for teaching rather than research, although many can serve both functions" (Chambers 2016: 364). The wide variety of applications of pedagogical corpora clearly demonstrates their usefulness for language learning and teaching. According to Römer (2009) (who goes back to the distinction proposed by Leech in 1997), indirect application of pedagogical corpora refers to the work carried out by researchers and didactic material developers, while direct application involves practical activities with the corpus by learners and teachers. One of the main characteristics of a pedagogical corpus regards its design process. This is because, as Braun points out, the 'pedagogic mediation of corpora' is necessary since the structure of many existing corpora, designed with linguistic research goals in mind, conflicts with the pedagogic requirements for corpus design and use (Braun 2005). One form of pedagogic mediation of corpora is through the close monitoring of the content of the corpus to identify possible structural (grammar and spelling) problems as well as sensitive/offensive content. Although preserving the original material in the corpus can be especially useful from the point of view of authenticity, the potentially problematic examples need to be presented with some guidance from the teachers. One way to facilitate the creation of language learning materials in general, and more specifically, of lexicographical resources, as well as enhance the use of corpora in the classrooms, is by marking the potentially problematic examples in pedagogical corpora. This way, teachers, material developers, lexicographers, among others, can choose to filter out certain examples according to their needs and context of use.

The main objective of our project is to create such labelled pedagogical corpora and use them for different purposes, among which are SkELLs (Sketch Engine for Language Learning)[1] for Dutch, Estonian,[2] Serbian, Slovene, and Portuguese. SkELL (Baisa & Suchomel 2014) is a free corpus tool with a pedagogical corpus which offers selected Sketch Engine functionalities (word sketch, examples, and thesaurus). Some tailored and more learner-friendly settings[3]

[1] https://skell.sketchengine.eu

[2] One can already use SkELL for Estonian. SkELL queries from the Estonian Corpus for Learners 2020 which was built using GDEX (Kilgarriff et al. 2008). One of the classifiers of GDEX for Estonian was a blacklist of words (incl. vulgarisms, swear words, potentially sensitive/offensive words), which were all omitted in the corpus building process (Koppel 2020).

[3] The interface of the tool is easy to navigate and clear, i.e., free of complex, rarely usable features. The search functionalities are as

also make SkELL language learning suited, thus serving as a complement to learners' dictionaries. As SkELL is fully automatically created, the included corpus must already be processed for aforementioned potential problems. State-of-the-art approaches to automated corpus filtering typically include the removal of structural noise and preselected problematic lexica (i.e., with the use of blacklists, as proposed by Kilgarriff et al. 2014). However, to reach a satisfactory quality, additional and more sophisticated approaches are needed, together with a clear understanding of what types of problems need to be addressed.

In order to make the corpus labelling process more efficient, as well as gather empirical data on the types of language examples the wider community perceives as problematic for teaching purposes, we have proposed and evaluated a method that applies crowdsourcing techniques. The goals of this paper are threefold: we report on an experiment that was performed, outline the lessons learned, and discuss how these have led us to propose an alternative way to compile pedagogical corpora with the help of the crowd.

This paper is organized as follows: Section 2 provides a brief review on the main methods applied to corpus cleaning and shows that, for our purposes, corpus labelling, rather than corpus cleaning, is required. Section 3 introduces the previous experiment that has been carried out to verify if crowdsourcing can be an adequate technique for corpus labelling, discusses the results of this experiment and presents some of the lessons learned. In Section 4, an alternative way of crowdsourcing corpus labelling, namely, through a game, is proposed. Section 5 wraps up this paper by pointing out some of the most significant challenges that have been faced so far and outlining what the next steps are.

2 From Corpus Cleaning to Corpus Labelling

Most of the literature about corpus cleaning refers to cleaning data from unnecessary information. For example, crawling data from the web implies extracting unnecessary tags, structural elements, meta-information, comments, links, commands and scripts (Spousta, Marek & Pecina 2008; Graën, Batinić & Volk 2014) or removing non-human-generated and quoted text (Styler 2011, Suchomel 2020). Many new approaches to web page cleaning were encouraged by the CLEANEVAL 2007 contest organized by ACL Web as Corpus interest group. Competitors used heuristic rules as well as different machine learning methods, including Support Vector Machines (Bauer & Knill 2007), decision trees, genetic algorithms, and language models (Hofmann & Weerkamp 2007). Another understanding of what corpus cleaning entails is cleaning the noise in the form of typos from large corpora. Reynaert (2006) talks about corpus induced corpus clean-up and presents a multilingual, language-independent and context-sensitive spelling checking and correction system, where the lexicon employed by the system is not a 'trusted' dictionary but contains noise in the form of recurrent typos found in any word type list derived from a large corpus of texts.

With regards to the compilation of pedagogical corpora specifically, many of them are carried out by linguistics institutes and university departments, often involving entire teams of linguistics experts. In this context, the main approach for creating a 'clean' corpus is to take an existing (web) corpus and select all sentences within a certain language-dependent range of words that are considered inappropriate for language learners (or a certain socio-cultural group). These are mostly swear words and sensitive words regarding politics, religion, sexuality, crime, illness, death, among others (Efthymiou, Gavriilidou & Papadopoulou 2014; Allan 2019). The goal is to exclude sentences containing these words from the (new) corpus. The easiest way of doing this is using a blacklist with 'sensitive/offensive' keywords (see Koppel et al. 2019), which is the method used for the creation of SkELL corpora, but this method has the disadvantage that either too few or too many sentences are excluded. In the first case, a relatively large new corpus still contains data that might be inappropriate for language learners in autonomous learning contexts. In the second case, a relatively small (and maybe too small) corpus is in itself appropriate but does not show the learner the subtleties of the language, since by eliminating sentences containing these words altogether, the corpus lacks representation of the neutral use of such words.[4] It is true that a careful selection of textual sources in the first place could help avoid some of the above-mentioned problems. If more reliable publishers, and possibly texts, are selected for compilation of pedagogical corpora, spelling/grammar problems might be filtered out. However, this also means that more time needs to be spent, while the variety of textual types and genres will inevitably be reduced as well. As is known, one of the advantages of using web corpora for pedagogical purposes is the wide spectrum of language use that can be found.

Although both types of corpus cleaning are unquestionably helpful, with regards to the purposes of this project, they are not sufficient, because the pedagogical corpora also need to be marked in terms of sensitive/offensive language and grammar/spelling mistakes. It should not be forgotten that in addition to the use of these corpora for publication of SkELL for Dutch, Estonian, Serbian, Slovene and Portuguese, and for dictionary making, it is expected that teachers will be able to use them for materials development. This means that all the sentences should be marked with tags for sensitivity, offensiveness, and structural mistakes, so that sentences can be filtered out according to the context of application.

For finding the balance between the two cases, manual assessment and selection seems unavoidable. Although this method should produce a pedagogical corpus with a higher level of quality, the amount of work and time that it takes for linguistics experts to create a 'clean' corpus has motivated a group of researchers within the COST action enetCollect[5] (Agerri et al. 2018; Lyding et al. 2018) to find an alternative solution by using crowdsourcing in the compilation process.

simple as possible (e.g., the search is case insensitive, it finds all parts of speech for a given word form), and the results are displayed in a readable and learning-oriented manner (e.g., instead of a list of concordances, whole sentences are displayed; the linguistic metalanguage is minimised; and special visualisations of language data are provided, such as wordclouds of similar words).

[4] Manual analysis of the blacklists created by the Sketch Engine team for automatic creation of web corpora for Dutch, Estonian, Serbian, Slovene, and Portuguese, and which would be also used for creation of SkELL corpora for those languages, has revealed that they contain many polysemous words that have both offensive and neutral senses.

[5] https://enetcollect.eurac.edu

Thus, an experiment was set up to test the viability of such a method. This experiment will be presented in the next section.

3 The Crowdsourcing Experiment

Crowdsourcing (or citizen science) is a practice where ordinary people, i.e., the crowd, contribute to creating content, solving problems, or even to doing some research. The crowd does not necessarily need to have expertise on the subject (be Čibej, Fišer & Kosem 2015; Nicolas et al. 2020). Benjamin (2015) has pointed out two characteristic features of crowdsourcing: 1) splitting the process into microtasks that can be completed with little effort, and 2) gamification where emphasis is placed on pleasure rather than effort. Crowdsourcing has been used in lexicography, e.g., for selecting keywords, formulating definitions, cross-editing the entries, providing examples, collocations, synonyms, word associations etc. (see, e.g., Čibej, Fišer & Kosem 2015; Arhar Holdt et al. 2020; Vainik 2018). It has also been used both in building the corpora (see, e.g., Ambati & Vogel 2010; Lane et al. 2010; Post, Callison-Burch & Osborne 2012) as well as in annotating the corpora (Bontcheva et al. 2014; Gut & Bayerl 2004). But as far as we know, it has not been used to mark general web corpora with potential offensive/sensitive content and structural problems so as to create annotated pedagogical corpora that can be used for dictionary making and language learning.

The main purpose of the experiment was to have the crowd help filter out offensive sentences from web corpora. Our secondary objectives were to have the crowd identify problematic sentences in corpora that, in principle, should be offensiveness/sensitivity-free, and to learn what the crowd considers to be offensive or sensitive. These specific objectives stemmed from our knowledge that automatic extraction of sentences based on blacklists fails to filter out sensitive content, that polysemous words can have neutral and offensive senses, and that offensiveness is a subjective matter.

Pybossa[6] was chosen as the crowdsourcing platform because a) it is free and b) because the custom tasks (interface) can be written in Javascript. In addition, one of the team members of the research project has a robust experience with using Pybossa in other crowdsourcing projects (Dekker & Schoonheim 2018a, 2018b) and has direct access to a local installation (INL) which ensures that the output data can be kept safely.

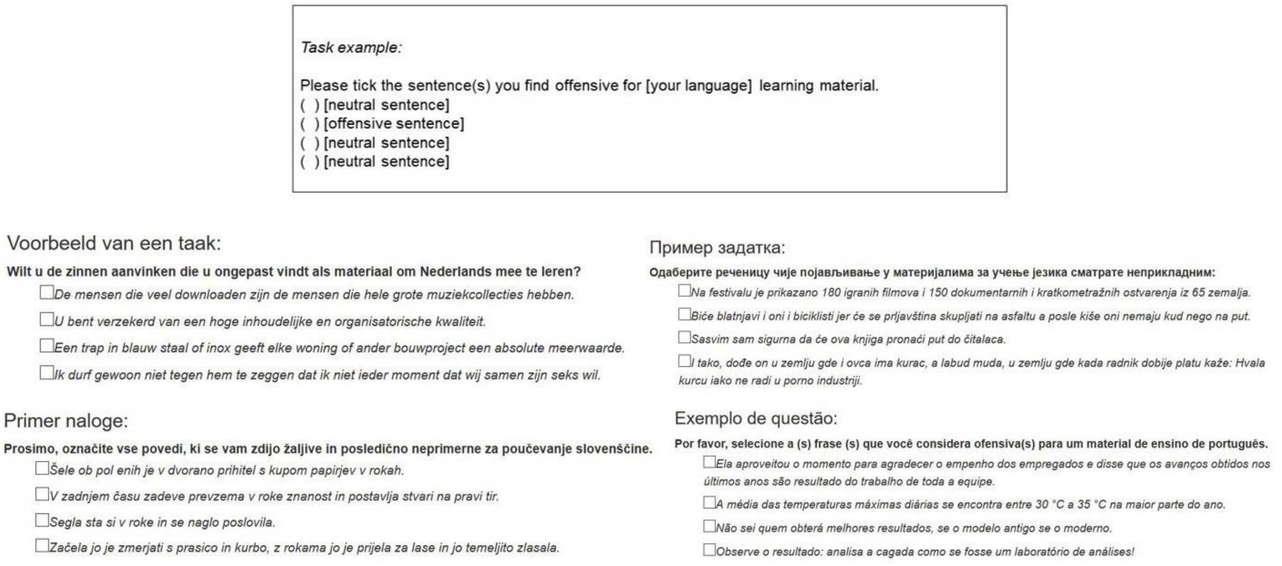
A multilanguage (Portuguese, Serbian, Dutch and Slovene) crowdsourcing project[7] was created with a common landing page, where the crowd was first asked to pick their language and then was transferred to the corresponding language home page. The individual languages' homepage had all the same structure and texts, which had been written together in English and later translated to each language. In addition, the Pybossa interface, i.e., buttons, messages, etc. also needed to be translated to each of the experiment languages. This presentation page contained a short introduction to the experiment, in which the purpose of the task and justification were provided and had the purpose to motivate participation by showing the participants that their contribution would benefit the community (i.e., social motivation, Čibej, Fišer & Kosem 2015). In addition, there was an invitation to participate, which contained the following: i) an example of the task that should be performed (see Figure 1); ii) a request of how many tasks we would like them to answer and the expected time that should take; iii) information about the institutions[8] promoting the experiment and about enetCollect; iv) a disclaimer informing the anonymous status of their answers, together with an example of the type of offensive sentence they could encounter and e-mail for contact; v) an informed consent to participate. Such detailed instructions are needed because of several reasons. Firstly, anticipating what kind of contribution is expected from the participants and how long that will take them may increase engagement. Secondly, showing that known academic institutions support the experiment ensures users that this is a reliable experiment. Thirdly, a clear description of how data provided by the participants will be handled and the provision of an informed consent conveys security. Finally, an example of highly offensive content allows participants to be psychologically prepared for the task and avoids dropouts.

Probably, what is more challenging for researchers creating a crowdsourcing experiment is the formulation of the right question (microtask design, Čibej, Fišer & Kosem 2015). This question needs to encourage participants to provide only the answers the task is aiming to obtain, and nothing else, but in the most straightforward and simple way possible. Figure 1 shows the model (in English) of the task example and illustrates how it was presented on each language's home page.

[6] https://pybossa.com/

[7] https://taalradar.ivdnt.org/corpusfiltering/

[8] The Centre for the Studies of General and Applied Linguistics at University of Coimbra (CELGA-ILTEC), Portugal; the Dutch Language Institute (INT) in Leiden, Netherlands; the Society for Language Resources and Technologies in Serbia (JeRTeh); and the Centre for Language Resources and Technologies, University of Ljubljana (CJVT), Slovenia.



Task example:

Please tick the sentence(s) you find offensive for [your language] learning material.
( ) [neutral sentence]
( ) [offensive sentence]
( ) [neutral sentence]
( ) [neutral sentence]

Voorbeeld van een taak:

Wilt u de zinnen aanvinken die u ongepast vindt als materiaal om Nederlands mee te leren?

□*De mensen die veel downloaden zijn de mensen die hele grote muziekcollecties hebben.*

□*U bent verzekerd van een hoge inhoudelijke en organisatorische kwaliteit.*

□*Een trap in blauw staal of inox geeft elke woning of ander bouwproject een absolute meerwaarde.*

□*Ik durf gewoon niet tegen hem te zeggen dat ik niet ieder moment dat wij samen zijn seks wil.*

Пример задатка:

Одаберите реченицу чије појављивање у материјалима за учење језика сматрате неприкладним:

□*Na festivalu je prikazano 180 igranih filmova i 150 dokumentarnih i kratkometražnih ostvarenja iz 65 zemalja.*

□*Biće blatnjavi i oni i biciklisti jer će se prljavština skupljati na asfaltu a posle kiše oni nemaju kud nego na put.*

□*Sasvim sam sigurna da će ova knjiga pronaći put do čitalaca.*

□*I tako, dođe on u zemlju gde i ovca ima kurac, a labud muda, u zemlju gde kada radnik dobije platu kaže: Hvala kurcu iako ne radi u porno industriji.*

Primer naloge:

Prosimo, označite vse povedi, ki se vam zdijo žaljive in posledično neprimerne za poučevanje slovenščine.

□*Šele ob pol enih je v dvorano prihitel s kupom papirjev v rokah.*

□*V zadnjem času zadeve prevzema v roke znanost in postavlja stvari na pravi tir.*

□*Segla sta si v roke in se naglo poslovila.*

□*Začela jo je zmerjati s prasico in kurbo, z rokama jo je prijela za lase in jo temeljito zlasala.*

Exemplo de questão:

Por favor, selecione a (s) frase (s) que você considera ofensiva(s) para um material de ensino de português.

□*Ela aproveitou o momento para agradecer o empenho dos empregados e disse que os avanços obtidos nos últimos anos são resultado do trabalho de toda a equipe.*

□*A média das temperaturas máximas diárias se encontra entre 30 °C a 35 °C na maior parte do ano.*

□*Não sei quem obterá melhores resultados, se o modelo antigo se o moderno.*

□*Observe o resultado: analisa a cagada como se fosse um laboratório de análises!*

Figure 1: Task example model in English and task examples in Dutch, Serbian, Slovene and Portuguese.

The experiment was advertised via e-mail, messages, and newsletter (Dutch) to all kinds of public, from close friends and family to members of our institutions and university students, as we were not targeting language specialists only.

3.1 Methodology

The experiment followed the same methodology of data preparation and Pybossa task design for all languages. Starting with data preparation, first, a list of the 100 most frequent nouns was compiled, which was further edited according to the characteristics of each language, arriving at a list of 38 nouns (lemmas) (see Table 1). Next, sentences containing these nouns were retrieved from the correspondent corpus in Sketch Engine (see Table 1) via API. For that, two extraction processes were applied - one with the GDEX function (Kilgarriff et al. 2008) in Sketch Engine enabled and another with the GDEX function disabled - which resulted in dataset 1 and dataset 2. GDEX stands for Good Dictionary Examples and is a function in the Sketch Engine tool that, based on predefined criteria, identifies example sentences in a corpus, placing the best ones at the top of the list of concordance lines in order to facilitate the lexicographer's process of example selection. It should be mentioned that the GDEX configurations have built-in blacklists that contain malicious or offensive content. Thus, sentences in dataset 1, which were extracted via process 1, 'passed' the GDEX control and were, therefore, considered potentially good, as this functionality enabled certain structural and semantic controls. Sentences in dataset 2, which were extracted via process 2, on the other hand, had not been filtered by the GDEX function, so it was not possible to determine whether they were good or 'bad' examples. A third step was added to the data preparation: the sentences extracted with GDEX-off were further filtered by language-dependent special blacklists, created separately from those built into the Sketch Engine, which were named 'curse lists', containing only explicitly offensive or sensitive words (see Table 1). Sentences containing words or expressions from this list were then automatically annotated as potentially inappropriate and included as ground truth for further analysis (Dekker et al. 2019; Zingano Kuhn et al. 2019), comprising dataset 3.

	Dutch	Serbian	Slovene	Portuguese
lemma list	Removal of mistagged nouns, proper nouns and numerals lemmas-nl.txt	lemmas-sr.txt	100 most frequent common nouns lemmas-sl.txt	Removal of proper nouns lemmas-pt.txt
corpus	NlTenTen 2 billion[9]	srWaC[10]	Gigafida[11]	pttenten_18_fl4_50 M (50-million-word sample from PtTenTen 3.8 bi)[8]
GDEX configuration	Based on CW_minimaal SketchEngine GDEX configuration. It is a minimal configuration, which favours collocations. Replaced optimal_length (9,12) and max length 30 by a hard length limit of between 7 and 40, to match the Portuguese configuration.			Portuguese.gdex configuration available in SketchEngine
curse list	Only swear words and manually expanded with personal knowledge curselist-nl.txt	curselist-sr.txt	Internally prepared list, using words labelled as vulgar from existing Slovene dictionaries curselist-sl.txt	Only swear words, with no polysemous or cultural-related words curselist-pt.txt

Table 1: Data preparation details.[12]

Moving now to the crowdsourcing experiment on Pybossa, two sets of tasks were designed and assigned randomly via the landing page. Set of tasks A contained only sentences from dataset 2, i.e., sentences that had not been GDEX-filtered. This means no pre-assumptions could be made as to their offensiveness status. Set of tasks B contained sentences from dataset 1, i.e., that had been GDEX-filtered, so were potentially good sentences, with some sentences from dataset 3, i.e., sentences that certainly contained offensive or sensitive words. Each set of tasks contained 4,560 sentences per language. Ideally, each sentence should be judged by three different people, so 13,680 judgements were needed per set of tasks. This means around 300 contributors were necessary: 150 for set of tasks A and 150 for set of tasks B. Therefore, our calculation was that each potential participant should judge around 90 sentences. Since each task contained 4 sentences, this resulted in approximately 23 tasks per participant. We estimated that this would be an optimal number of tasks per participant that would benefit the experiment, without being too time-consuming (we estimated that it would take 10 minutes to answer 23 tasks).

3.2 Results and Lessons Learned

The Pybossa output was collected after the experiment was online for two months. The level of engagement was very low for the Serbian, Slovene and Portuguese experiments (43, 12 and 32 contributors, respectively). For Dutch, numbers were more promising (131 contributors), although still far below from the total number required to have all sentences judged by three people. Despite this, the analysis of the outcome has revealed some very interesting insights.

For each sentence, we analyzed the cases in which the crowd's input contradicted our assumptions about appropriate or inappropriate content (Dekker et al. 2019; Zingano Kuhn et al. 2019). Analyzing the crowd's responses, we noticed the following cases:

- *TP (True positives)* - sentences annotated as potentially inappropriate and considered inappropriate by the crowd majority.
- *FN (False negatives)* - sentences annotated as potentially inappropriate and considered appropriate by the crowd majority.
- *FP (False positives)* - sentences annotated as potentially appropriate and considered inappropriate by the crowd majority.
- *TN (True negatives)* - sentences annotated as potentially appropriate and considered appropriate by the crowd majority.
- *UKN (Unknown)* - the number of participants who found the sentence inappropriate was equal to the number of participants who found the sentence appropriate, and vice versa.

While true positive results and true negative results confirmed our assumptions, the false negative and false positive ones

[9] NlTenTen and PtTenTen are web corpora compiled by the Sketch Engine team as part of the TenTen family (Jakubíček et al. 2013).
[10] srWaC is a Serbian corpus made up of texts collected from the Internet. https://www.sketchengine.eu/srwac-serbian-corpus/
[11] Gigafida is the reference corpus of written Slovene language. The current version 2.0 is described in Krek et al. 2020, and available at https://viri.cjvt.si/gigafida/.
[12] All input files are given on GitHub: https://github.com/Branislava/corpuscleanup_v1

have given us an opportunity to learn what participants think. The manual analysis of the *FP* sentences (*False Positives*) from each language has revealed that these sentences were mostly sophisticated (i.e., not directly formulated) cases of misogyny or religiously-offensive content. *False Positives* were also attested with sentences spreading propagation of violence towards children or containing topics related to war and politics. Since these sentences did not explicitly exhibit blacklisted words, they were not detected by the system in the first place.

Interestingly, the participants did not consider sentences with explicitly rude content necessarily inappropriate. These cases represented our false negatives. We assumed that this was due to two reasons: 1) the crowd found sentences including obscene lexica not necessarily bad learning material, and 2) some annotators were more concentrated on the structure and language accuracy, and less on the pedagogical implications of the inclusion of such sentences in language learning materials.

Another finding originated from the feedback provided by the participants after the task. The feedback was optional, and we primarily expected reports on technical problems or similar. However, among the Slovene participants, two reported that they found the task rather purposeless due to the lack of problems in the data. This indicates that for non-web corpora, as is the Gigafida reference corpus, the material for the crowdsourcing task needs to be chosen with more emphasis on the problems: their lack had a demotivational effect on the participants.

We generally concluded that participants were willing to help, but also often inclined to interpret the task in their own way. Even though in our case the experiment was specifically focused on marking what was strictly 'offensive' to the participants, they often did more than this. For example, they marked the sentences that they found inappropriate for a learner's material, such as incomplete sentences, complex sentences, sentences containing spelling and grammar errors or even sentences containing too many foreign terms.

4 Gamifying corpus labelling

Based on the modest results and on the lessons learned from the experiment, we concluded that a new way of motivating the crowd and a more specific task were required. We then opted to follow the 'Games with a Purpose' (GWAP) (von Ahn, 2006) approach, "i.e., games that are fun to play and at the same time collect useful data for tasks that computers cannot yet perform" (Hacker & Ahn 2009: 1208). GWAPs have been often designed to annotate or clear language data for the creation of various lexical infrastructures, for example JeuxDeMots (Lafourcade 2007), Phrase Detectives (Poesio et al. 2013), Wordrobe (Venhuizen et al., 2013), ZombiLingo (Guillaume, Fort & Lefebvre 2016), Game of Words (Arhar Holdt et al. 2020, Kosem et al. 2020). Thus, at this second stage of the project, a game for web corpora labelling is under development.

The model of the game is inspired by Matchin (Hacker & Ahn 2009). The main idea of Matchin is to elicit users' preferences about images without asking them directly, but rather by asking what their opponent player would prefer. Players are rewarded when their predictions match. Taking into consideration Hacker and Ahn's claim that "asking partners in a two-player game to guess which of two options their partner will choose represents a viable mechanism for extracting user preferences and data' (Hacker & Ahn 2009: 1208), we have decided to build on the mechanism of Matchin to collect information on corpus examples. According to Hacker and Ahn (2009), their game has been extremely successful, with tens of thousands of players. It is our hope that additional game modes and a variety of gamification elements, such as scoring, players scoreboard, avatar, etc. will contribute to motivate the crowd to play our game.

The main purpose of our game is to have players identify problematic corpus sentences (choosing between two sentences offered), and then categorize the identified sentences according to the type of problematic content. According to Sabou et al. (2014), categories should be kept between 2 to 5, to avoid cognitive overload. To define the categories, manual assessment of 100 automatically extracted sentences from corpora of Slovene and Serbian has been performed, leading to the introduction of the following five categories: offensive, vulgar, sensitive content, spelling/grammar problems, lack of context/incomprehensible. A help pop-up page will be available for players to see example sentences of each category in order to help them make categorization decisions. At the moment of writing, game modes and players interface are being developed for all languages (Dutch, Serbian, Slovene, Estonian and Portuguese).

The game development project is organized into three phases: data preparation, game preparation, and preparation of machine learning for automatic corpus labelling, as can be seen in the diagram below (Figure 2). The first phase involves preparing the datasets that will feed the game. For that, the corpora of all languages will be GDEXed with especially created pedagogically-driven GDEX configurations for each language, consisting of a common set of criteria and some language-dependent criteria. For instance, one thing that has come out from the previous experiment is that sentence length (in words) is important and has to be determined per language. These sentences will be filtered by blacklists, resulting in two types of output (i.e., the potentially good and bad sentences) that will be used as the input datasets for the game. The second phase is game preparation, which involves a) the development of different game modes, b) gamification aspects, such as scoring and motivation, and c) the player interface. In addition, a researcher interface will be built to allow easy access to the database containing the labelled sentences. In the third phase, machine learning models for all languages involved will be trained to automatically identify problematic content (manually categorized by the players) in web corpora. We expect this automatic identification of problematic content will facilitate the compilation of larger, clean corpora.

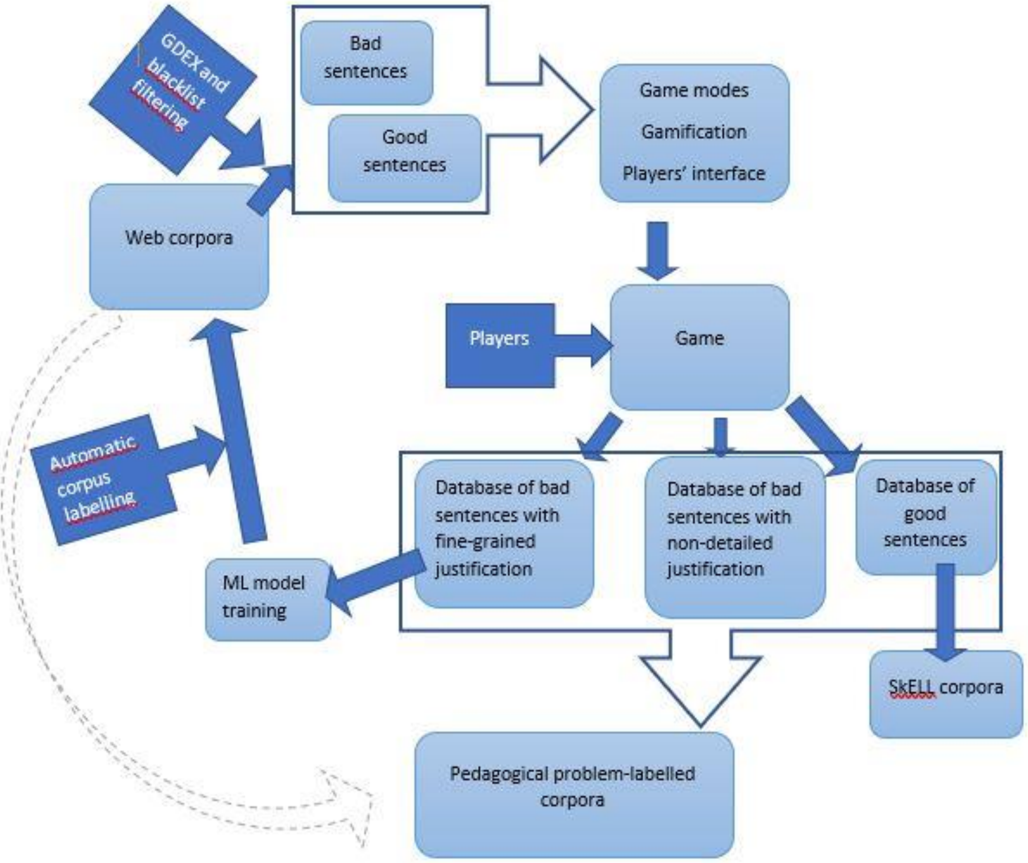

GDEX and blacklist filtering
Bad sentences
Good sentences
Game modes
Gamification
Players' interface
Web corpora
Players
Game
Automatic corpus labelling
Database of bad sentences with fine-grained justification
Database of bad sentences with non-detailed justification
Database of good sentences
ML model training
SkELL corpora
Pedagogical problem-labelled corpora

Figure 2: Game development project.

The game will be available as a webpage and as a mobile phone app. The players' answers, including the submitted labelling of sentences, will be logged and stored in a database. This way, language teachers and teaching material creators will be able to compile labelled corpora for pedagogical purposes from this output, i.e., corpora that still contain problematic sentences, but that can be used since the labels enable them to (de)select content/structure that is considered inappropriate or not (yet) suitable for the category of language learners involved. In addition, lexicographers will be able to compile filtered corpora containing only unmarked sentences, for instance, all sentences that have not been marked can be used for compiling the SkELL corpora.

5 Concluding Remarks

The progress in the field of automatic detection of good corpus examples has been considerable, and the tools have been used extensively especially in lexicography, and to a lesser extent in language pedagogy, one problem being the lack of availability of (suitable) pedagogical corpora. The approach we propose in this paper is to create pedagogical corpora from larger web corpora, using crowdsourcing. As our experiment with the labelling of corpus sentences has confirmed, crowdsourcing can be a very helpful and efficient method for these purposes. With the help of the crowd, sentences with offensive/sensitive content can be filtered out from web corpora. At the same time, the method also provides a valuable insight into what the crowd, i.e., the community, considers as (in)appropriate content.

Nonetheless, our experiment also revealed that improvements to the methodology were needed, particularly in terms of having more motivating tasks to increase the level of engagement by the participants and providing more focused questions to guarantee the input provided by the participants is relevant. The project is now exploring an alternative way of using crowdsourcing by adopting the 'Games with a Purpose' approach. In this new stage of the project, a game for web corpora labelling is under development. While the gamification approach addresses some of the issues encountered during the experiment, it brings new challenges related to game development and design, and dissemination.

If this gamification experiment turns out to be successful, it will open a new way of creating pedagogical corpora with the help of crowdsourcing. These corpora will have many different possible uses, especially in language learning, but also in other fields. For example, in lexicography, such corpora can be considered invaluable sources of good candidate examples, and on their basis, dictionary creation could become considerably faster. It is our ultimate goal to provide examples of good practice and prepare workflows that can serve as the benchmark for other languages, especially under-resourced ones.

6 References

Allan, K. (Ed.) (2019). *The Oxford Handbook of Taboo Words and Language*. Oxford University Press.

Ambati, V. & Vogel, S. (2010). Can crowds build parallel corpora for machine translation systems?. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk*.

Arhar Holdt, Š., Logar, N., Pori, E. & Kosem, I. (2020). "Game of Words": Play the Game, Clean the Database. In *Proceedings of the XIX EURALEX International Congress*.

Agerri, R., Maritxalar, M., Lyding, V., & Nicolas, L. (2018). enetCollect: A New European Network for combining Language Learning with Crowdsourcing Techniques. *Procesamiento Del Lenguaje Natural, 61*, 171-174. doi:http://dx.doi.org/10.26342/2018-61-25

Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In A. Horák, P. Rychlý (eds) *Proceedings of the Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014*. Brno: Tribun EU, pp. 63-70.

Bauer, M.W. Knill, C. (eds.) (2007). *Management reforms in international organizations.* Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG

Behzadan, V., Aguirre, C., Bose, A. & Hsu, W. (2018). Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE.

Benjamin, M. (2015). Crowdsourcing microdata for cost-effective and reliable lexicography. In L. Li, J. Mckeown, L. Liu (eds.) *Proceedings of AsiaLex 2015, Hong Kong*. Hong Kong Polytechnic University, pp. 213-221.

Bontcheva, K., Roberts, I., Derczynski, L. & Alexander-Eames, S. (2014). The GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Boulton, A. (2017). Corpora in language teaching and learning: Research timeline. *Language Teaching*, Cambridge University Press (CUP), 50 (4), pp.483-506.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*, pp. 47-64.

Chambers, A. (2016). Written language corpora and pedagogic applications. In F. Farr, L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology*, pp. 362-375.

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubíček, J. Kallas, S. Krek (eds.) *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age. Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana, Brighton: Trojina, Institute for Applied Slovene Studies, Lexical Computing Ltd., pp. 70-83.

Dekker, P., & Schoonheim, T. (2018a). Crowdsourcing Language Resources for Dutch using PYBOSSA: Case Studies on Blends, Neologisms and Language Variation. In *Proceedings of the enetCollect WG3&WG5 Meeting, 24-25 October 2018*. Leiden, Netherlands.

Dekker, P., & Schoonheim, T. (2018b). When to use PYBOSSA? Case studies on crowdsourcing for Dutch. Presentation at *enetCollect WG1 hands-on workshop Gothenburg*, December 2018.

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 *eLexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts*. Brno: Lexical Computing.

Efthymiou, A., Gavriilidou, Z. & Papadopoulou, E. (2014). Labeling of Derogatory Words in Modern Greek Dictionaries. In N. Lavidas, T. Alexiou & A. Sougari (Ed.), *Major Trends in Theoretical and Applied Linguistics 2*. Versita Ltd, 78 York Street, London W1H 1DP, Great Britain.: De Gruyter Open Poland, pp. 27-40.

Graën, J., Batinić, D. & Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *Konvens 2014, Hildesheim, 8 October 2014 - 10 October 2014*.

Guillaume, B., Fort, K. & Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *International Conference on Computational Linguistics (COLING).*

Gut, U. & Bayerl, P.S. (2004). Measuring the reliability of manual annotations of speech corpora. In *Speech Prosody 2004, International Conference*.

Hacker, S. & Von Ahn, L. (2009). Matchin: eliciting user preferences with an online game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Hofmann, K. & Weerkamp, W. (2007). Web corpus cleaning using content and structure. In *Proceedings of the Web as Corpus Workshop (WAC3), Cleaneval Session*. Louvain-la-Neuve, Belgium.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference*. Lancaster, UK.

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), pp.7-36.

Kilgarriff, 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, Spain: Documenta Universitaria, pp. 425-432.

Koppel, K. (2020). *Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele* [Corpus-Based Automatic Detection of Example Sentences for Dictionaries for Estonian Learners]. (Doktoritöö, Tartu Ülikool). Tartu: Tartu Ülikooli Kirjastus.

Koppel, K., Kallas, J., Khokhlova, M., Suchomel, V., Baisa, V. & Michelfeit, J. (2019). SkELL corpora as a part of the language portal Sõnaveeb: problems and perspectives. In: I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreira, M. Janssen, 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. 763−782.

Kosem, I., Martelli, F., Navigli, R., Jakubíček, M. & Kallas, J. (2020). *Crowdsourcing Module. Deliverable 4.3 of the*

European Lexicographic Infrastructure Project. https://elex.is/wp-content/uploads/2020/03/ELEXIS_D4_3_Crowdsourcing_module.pdf [25/07/2021]

Krek, S., Holdt, Š.A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. & Dobrovoljc, K. (2020). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th international symposium on natural language processing*.

Lane, I., Eck, M., Rottmann, K. & Waibel, A. (2010). Tools for collecting speech corpora via Mechanical-Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with amazon's mechanical turk*.

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*.

Nicolas, L., Lyding, V., Borg, C., Forăscu, C., Fort, K., Zdravkova, K., Kosem, I., Čibej, J., Holdt, Š.A., Millour, A. & König, A. (2020) Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. & Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1), pp.1-44.

Post, M., Callison-Burch, C. & Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Reynaert, M. (2006). Corpus-Induced Corpus Clean-up. In *LREC* 2006.

Römer, U. (2009). Using general and specialised corpora in language teaching: Past, present and future. In M.C. Campoy, B. Belles-Fortuno, M. L. Gea-Valor (Eds.), *Corpus-Based Approaches to English Language Teaching*. Continuum Publishing Corporation, pp.18-35.

Sabou, M., Bontcheva, K., Derczynski, L. & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *LREC 2014*.

Spousta, M., Marek, M. & Pecina, P. (2008). Victor: the web-page cleaning tool. In *4th Web as Corpus Workshop (WAC4)-Can we beat Google*.

Styler, W. (2011). *The EnronSent corpus*. Boulder: University of Colorado at Boulder Institute of Cognitive Science.

Suchomel, V. (2020). Better Web Corpora for Corpus Linguistics and NLP. Doctoral theses. Masaryk University, Faculty of Informatics, Brno, Czech Republic.

Vainik, E. (2018). Compiling the Dictionary of Word Associations in Estonian: from scratch to the database. *Eesti Rakenduslingvistika Ühingu aastaraamat*, (14), pp.229-245.

Venhuizen, N., Evang, K., Basile, V. & Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6), pp.92-94.

Vyatkina, N., & Boulton, A. (2017). Corpora in language teaching and learning. *Language Learning and Technology*, *21*(3), 1-8.

Zingano Kuhn, T.Z., Dekker, P., Šandrih, B., Zviel-Girshin, R., Arhar, Š., Holdt, T.S & Schoonheim,T. (2019). Crowdsourcing Corpus Cleaning for Language Learning Resource Development. In *EuroCALL 2019: European Association of Computer Assisted Language Learning*.

Acknowledgements

This work was supported by the Portuguese national funding agency, FCT - Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020). The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene). Many ideas result from the framework of the CA160105 enetCollect COST Action. We thank all that made our work possible.

EURALEX XIX
Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Posters

Pedagogical Lexicography

Αρχές για τη δημιουργία ενός εξειδικευμένου λεξικού για ποιητικούς νεολογισμούς: μελέτη περίπτωσης στην ΟΔΥΣΕΙΑ του Νίκου Καζαντζάκη

Μαθιουδάκης Ν.

Πανεπιστήμιο Γρανάδας
nikosmathious@gmail.com

Περίληψη

Το φαινόμενο της νεολογίας αποκαλύπτεται ως μια ανανεωτική δύναμη της γλώσσας, αλλά ταυτόχρονα αναδεικνύεται και ως μια ιδιότυπη λειτουργία του λογοτεχνικού ύφους, καθώς η δημιουργία νέων λέξεων ή/και εννοιών από τους λογοτέχνες αποτελεί χαρακτηριστικό τους γνώρισμα και στοιχείο της ποιητικής γραμματικής τους, της αποτύπωσης της προσωπικής τους σφραγίδας. Επομένως, οι ποιητικοί νεολογισμοί είναι μέρος του συνόλου των νεολογισμών, αν και θεωρούνται συνήθως εφήμερες δημιουργίες των λογοτεχνών. Σκοπός της έρευνάς μας είναι η θεμελίωση των αρχών για τη μελέτη των ποιητικών νεολογισμών υπό το πρίσμα του λεξικογραφικού πεδίου, ελλείψει εξειδικευμένων λεξικών αναφορικά με τους ποιητικούς νεολογισμούς νεοελλήνων συγγραφέων. Ειδικότερα, παρουσιάζονται οι αρχές δημιουργίας ενός ψηφιακού λεξικού για τους ποιητικούς νεολογισμούς της καζαντζακικής ΟΔΥΣΕΙΑΣ, καθώς παρουσιάζει εξαιρετικό ενδιαφέρον το ιδιοσυγκρασιακό και ιδιοτυπικό λεξιλόγιο του Νίκου Καζαντζάκη, μιας και στο ποιητικό του έργο εγκιβωτίζει περίπου 5.000 νεολογικές αθησαύριστες λέξεις, που είναι κυρίως σύνθετοι και πολυσύνθετοι σχηματισμοί. Τέλος, περιγράφουμε διεξοδικά τη μακροδομή και τη μικροδομή του λεξικού, σημειώνοντας ενδεικτικά παραδείγματα λημμάτων των καζαντζακικών ποιητικών νεολογισμών.

Έννοιες-κλειδιά: ποιητικοί νεολογισμοί, γλώσσα της λογοτεχνίας, νεοελληνική λογοτεχνία, εξειδικευμένη λεξικογραφία, εξειδικευμένα λεξικά, ΟΔΥΣΕΙΑ, Καζαντζάκης

1 Εισαγωγή

Το φαινόμενο της νεολογίας παρατηρείται τόσο στον γραπτό όσο και στον προφορικό λόγο. Σχετίζεται άμεσα με την ανθρώπινη επικοινωνία και, μάλιστα, αποτελεί μια καθημερινή αναγέννηση της γλώσσας. Οι νεολογισμοί γεννιούνται καθημερινά, ορισμένοι εκ των οποίων γίνονται εύκολα αντιληπτοί, μιας και θεωρούνται ασυνήθιστοι και παράξενοι τύποι λέξεων. Σύμφωνα με την Αναστασιάδη-Συμεωνίδη (1986: 26), οι νεολογισμοί, προϊόν της νεολογίας, αποτελούν αναπόσπαστο κομμάτι του ατομικού λεξιλογίου και έχουν άμεση σχέση με τη γλωσσική ικανότητα κάθε ανθρώπου. Αυτόματα, λοιπόν, οι νέες αυτές λεξικές μονάδες επηρεάζονται από το κοινωνικό-πνευματικό περιβάλλον, αφού η κοινωνία και το λεξιλόγιο είναι έννοιες άμεσα συνδεδεμένες στη γλωσσική λειτουργία.

Με τον όρο *νεολογισμός* αναφερόμαστε σε κάθε νέα λέξη (ή κάθε νέα σημασία λέξης), η οποία εμπλουτίζει μια γλώσσα, καθώς χρησιμοποιείται με τρόπο είτε ενεργητικό είτε παθητικό από τη γλωσσική κοινότητα. Ο Picone (1996) υποστηρίζει ότι νεολογισμός είναι κάθε νέα λέξη, μόρφημα ή έκφραση και κάθε νέα σημασία για μια προϋπάρχουσα λέξη, έκφραση ή προϋπάρχον μόρφημα που εμφανίζεται σε μια γλώσσα. Αντίστοιχα, η Lehrer (2005) θεωρεί τους νεολογισμούς ως νέες λέξεις που συχνά εμφανίζονται και εισάγονται σε μια γλώσσα ενώ άλλες λέξεις χρησιμοποιούνται, είτε για μικρό χρονικό διάστημα είτε πιθανότατα μόνο για μια φορά.

Επομένως, ο νεολογισμός ως το άμεσο παράγωγο του φαινομένου της νεολογίας είναι κάθε μη συστηματικά κωδικοποιημένος σχηματισμός (ή κάθε νέα σημασία ενός προϋπάρχοντος σχηματισμού) σε κάποιο βασικό λεξικό μιας γλώσσας, ο οποίος χρησιμοποιείται είτε προφορικά είτε γραπτά και σύμφωνα με διάφορα κριτήρια, χρήζει ανάγκης για καταγραφή και μελέτη. Πολλοί μελετητές διερευνούν τις διαδικασίες δημιουργίας νεολογισμών, στον γραπτό και τον προφορικό λόγο, και προσπαθούν να τυποποιήσουν τις συνθήκες δημιουργίας των νέων λεξικών σχηματισμών. Προσεγγίζοντας, ειδικότερα, τη γραπτή νεολογία διαπιστώνουμε πως υπάρχει ένα πλήθος νεολογισμών σε κάθε έκφανση της καθημερινότητάς μας, όπως ακριβώς συμβαίνει και στη λογοτεχνία, η οποία θεωρείται αναπόσπαστο κομμάτι της ανθρώπινης ανάπτυξης του πολιτισμού.

Στην παρούσα εργασία, ως μελέτη περίπτωσης, ασχολούμαστε με το επικό ποίημα ΟΔΥΣΕΙΑ[1] του Νίκου Καζαντζάκη, η οποία αποτελεί θεματικά τη συνέχεια της *Οδύσσειας* του Ομήρου. Σε μια έκταση 33.333 στίχων, διαρθρωμένη σε 24 ραψωδίες, ο Καζαντζάκης ως ένας λεξιθήρας, ένας ερασιτέχνης γλωσσολόγος, ο οποίος αποστρέφεται την κοινή και χιλιοειπωμένη λέξη, εγκιβωτίζει στο έπος του περισσότερες από 5.000 νεολογικές αθησαύριστες λέξεις εν δυνάμει ποιητικούς νεολογισμούς (Μαθιουδάκης 2012) – στην πλειονότητά τους σύνθετες και πολυσύνθετες λέξεις (Μαθιουδάκης & Καρασίμος 2014, 2015) –, αναδεικνύοντας με εξαιρετικό τρόπο τόσο τον ιδιοσυγκρασιακό του χαρακτήρα που αποτυπώνεται αδιαμφισβήτητα στο λεξιλόγιο και στα χαρακτηριστικά της ΟΔΥΣΕΙΑΣ όσο και την προσωπική συγγραφική του σφραγίδα, το ατομικό του υφολογικό αποτύπωμα. Στόχος μας είναι η περιγραφή του σχεδιασμού και της κατασκευής ενός εξειδικευμένου ψηφιακού λεξικού με τους ποιητικούς καζαντζακικούς νεολογισμούς, ελλείψει εξειδικευμένων λεξικών παλαιότερων και νεότερων ελλήνων συγγραφέων, προκειμένου να αποτελέσει χρηστικό εργαλείο, αλλά και πρότυπο ανάπτυξης παρόμοιων λογοτεχνικών λεξικών.

[1] Επισημαίνεται ιδιαιτέρως εμφατικά πως στη μελέτη υιοθετείται ο τίτλος του καζαντζακικού έπους ΟΔΥΣΕΙΑ σε κεφαλαιογράμματη γραφή με ένα σίγμα («Σ») και όρθια γράμματα, από τη μία ως ένδειξη σεβασμού στη βούληση του ίδιου του συγγραφέα να τιτλοφορήσει με τον συγκεκριμένο τρόπο γραφής το ποίημά του στην πρώτη του έκδοση (1938) και από την άλλη ως διακριτό σημείο προς αποφυγή σύγχυσης με τον τίτλο του γνωστού έπους του Ομήρου, *Οδύσσεια*, η οποία γράφεται με δύο σίγμα.

2 Ποιητικοί νεολογισμοί

Η νεολογία είναι, λοιπόν, η ανανεωτική πνοή που διέπει τη γλώσσα, εμπλουτίζοντάς τη με νεοτερισμούς και καινοτόμες δημιουργικές εκφράσεις σε όλα τα επίπεδά της (Καραντζόλα & Φλιάτουρας 2004). Κυρίως, όμως, όπως σημειώνει και η Aitchison (2006: 292), «το λεξιλόγιο είναι ίσως το μέρος της γλώσσας που αλλάζει περισσότερο σε κάθε γενιά», επιταχύνοντας και άλλες παράλληλες γλωσσικές και γενικότερες αλλαγές.

Οι *λογοτεχνικοί* ή αλλιώς *ποιητικοί νεολογισμοί* αποτελούν, όπως είναι φυσικό, δημιουργικό παράγωγο της νεολογίας και μέρος των νεολογισμών (κυρίως του γραπτού λόγου), καθώς είναι οι δημιουργικές λογοτεχνικές εκφράσεις των συγγραφέων. Η Αναστασιάδη-Συμεωνίδη (1986) χαρακτηρίζει τους ποιητικούς νεολογισμούς ως γεμάτες φαντασία αλλά εφήμερες δημιουργίες των λογοτεχνών, προτείνοντας την εξειδικευμένη μελέτη τους, αφού οι λογοτέχνες, θεωρητικά τουλάχιστον, έχουν απεριόριστες δυνατότητες για νεολογική δημιουργία. Από την άλλη πλευρά, ο Χαραλαμπάκης, τόσο στη μελέτη του περί νεολογισμών (2011) όσο και σε ιδιαίτερες μελέτες του περί ύφους (2001), δίνει ιδιαίτερη έμφαση στις νεολογικές ποιητικές δημιουργίες ελλήνων λογοτεχνών, αναγνωρίζοντάς τες ως ιδιαίτερα στοιχεία ύφους. Επίσης, σύμφωνα με τον Guilbert (1973, 1975), οι συγγραφείς θεωρούνται ως σημαντικοί παράγοντες ανανέωσης και δημιουργίας της γλώσσας, καθώς οι νεολογικές λέξεις που βρίσκονται ενσωματωμένες σε ένα λογοτεχνικό κείμενο συχνά είναι μέρος της προφορικής παράδοσης και του γραπτού λόγου, αποτυπώνοντας μια πληθώρα κοινωνικο-πολιτικο-πολισμικών πτυχών της γλώσσας, επιφορτίζοντας την έννοια του νεολογισμού με έναν σημαίνοντα ρόλο στο πλαίσιο ενός λογοτεχνικού κειμένου (Oleynikova 2016).

Όσον αφορά την ανανεωτική πνοή που δίνει ο ποιητής στη γλώσσα μέσα από διάφορους μηχανισμούς, ο Νάκας επεξηγεί το φαινόμενο της νεολογίας που το αναφέρει και ως «νεολεξία», επισημαίνοντας ότι το φαινόμενο αυτό είναι σύνηθες στην ποίηση, όπως και σε άλλες πτυχές του ανθρώπινου βίου, αλλά, ωστόσο, υποστηρίζει ότι «ποιητής δεν σημαίνει απαραίτητα και γλωσσοπλάστης» (2003: 187).

Όπως είναι αναμενόμενο, αρκετοί μελετητές, ανάμεσά τους η Χριστοφίδου (2001), η Κόλλια (2007, 2011), ο Χαραλαμπάκης (2011) και ο Μαθιουδάκης (2012, 2020) ασχολούνται με την αναζήτηση νεολογισμών και νεολογικών λέξεων σε λογοτεχνικά κείμενα, χαρακτηρίζοντας τους συγκεκριμένους ποιητικούς νεολογισμούς ως στοιχείο ύφους του συγγραφέα. Επίσης, οι συγκεκριμένοι ερευνητές κάνουν προσπάθειες να διαπιστώσουν τις περιπτώσεις κατά τις οποίες οι αναγνώστες μπορούν να κατανοήσουν τη σημασία νέων σχηματισμών ή ήδη υπαρκτών με άλλη σημασία, προκειμένου να απολαύσουν την ανάγνωση του λογοτεχνικού κειμένου.

Επομένως, η δημιουργία ενός λεξικού για τους ποιητικούς νεολογισμούς και μάλιστα εξειδικευμένου κρίνεται αναγκαία τόσο για τους επιστήμονες και τους εκπαιδευτικούς αλλά και για το ευρύτερο αναγνωστικό κοινό. Τέλος, αξίζει να σημειωθεί πως ο Χαραλαμπάκης, τον Νοέμβριο του 2015[2], στο πλαίσιο της απονομής του βραβείου «Δαίδαλος» στον πρόσωπό του για το «Χρηστικό λεξικό της νεοελληνικής γλώσσας» (εκδ. Ακαδημία Αθηνών) από την Εταιρεία Συγγραφέων, ανακοίνωσε πως η Ακαδημία Αθηνών ετοιμάζει τη σύνταξη ενός λεξικού «που θα περιλαμβάνει όχι μόνο τους νεολογισμούς αλλά και τις σημασίες που αποκτούν οι λέξεις στους κορυφαίους παλαιότερους και νεότερους λογοτέχνες μας, πεζογράφους και ποιητές», αναλύοντας διεξοδικά τους λόγους για τους οποίους ένα τέτοιο λεξικό πρόκειται να είναι χρήσιμο και για την επιστήμη και για την κοινωνία.

3 Εξειδικευμένη λεξικογραφία και Εξειδικευμένα λεξικά

3.1 Οροθεσία

Σύμφωνα με τους Fuertes-Olivera & Tarp 2014, η *εξειδικευμένη λεξικογραφία* (*specialised lexicography*) ορίζεται ως ο κλάδος της λεξικογραφίας που ασχολείται με τη θεωρία και την πρακτική εξειδικευμένων λεξικών, δηλαδή με λεξικά, γλωσσάρια, λεξιλόγια, αλλά και κάθε πληροφοριακό εργαλείο, που καλύπτουν περιοχές εκτός της γενικής πολιτιστικής γνώσης και της λεγόμενης *γλώσσας για γενικούς σκοπούς* (*language for general purposes – LGP*). Έτσι, η εξειδικευμένη λεξικογραφία μπορεί να χαρακτηρισθεί και ως ειδικού σκοπού και αντιπροσωπεύει διάφορους κλάδους, όπως και τις κοινωνικές και τις ανθρωπιστικές επιστήμες. Παρόλο που συχνά αναφέρεται ως *λεξικογραφία εξειδικευμένης γλώσσας* (*specialised language – LSP*), η εξειδικευμένη λεξικογραφία υπερβαίνει κατά πολύ μια απλή περιγραφή των διάφορων εξειδικευμένων γλωσσών, ενώ επίσης αντιμετωπίζει εγγενώς την ουσία των γλωσσών αυτών, προκειμένου να παρέχει άμεση και έγκαιρη πρόσβαση σε ουσιώδη γνωστικά επιτεύγματά τους.

Οι Gouws & Tarp (2019) σημειώνουν πως η θέση ενός σύγχρονου λεξικογράφου, και ειδικότερα λεξικογράφων εξειδικευμένων λεξικών, είναι να αναπτύσσει έναν ιδιότυπο κοινωνικό ρόλο, προκειμένου να αντιμετωπίζει, αφενός, τις ανάγκες των χρηστών ενός λεξικού και αφετέρου, να λαμβάνει υπόψη τις τρέχουσες τεχνολογικές δυνατότητες. Για τον λόγο αυτό, οι ίδιοι μελετητές (2019: 259) υποστηρίζουν πως «ορισμένοι οραματιστές λεξικογράφοι ήρθαν στο προσκήνιο με προτάσεις για την αντιμετώπιση των προβλημάτων της αποσυγκειμενοποίησης (decontextualization) και για να προχωρήσουν πέρα από το μη φυσικό περιβάλλον (unnatural environment) ενός λεξικού προς την παροχή εξατομικευμένων πληροφοριών απευθείας σε ένα πλαίσιο όπου ο χρήστης βιώνει μια πληροφορία που χρειάζεται». Επομένως, οι σύγχρονοι λεξικογράφοι είναι σε θέση να κάνουν μερικά από τα ανεκπλήρωτα όνειρα του παρελθόντος πραγματικότητα, προβάλλοντας το γεγονός πως η πρόκληση του μέλλοντος είναι να καταστεί το αδύνατο δυνατό.

Στο πλαίσιο, μάλιστα, της σύνταξης και δημιουργίας ενός ψηφιακού λεξικού, αναπτύσσοντας αρχές της ψηφιακής λεξικογραφίας, ο Tarp (2011) προτάσσει ένα μελλοντικό λεξικογραφικό μοντέλο ως πιθανόν δρόμο για την «εξατομίκευση των αναγκών ικανοποίησης». Επίσης, ο Tarp (2012) επισημαίνει πως το προφίλ κάθε χρήστη, η περιγραφή της κάθε

[2] Βλ. Μαγνητοσκοπημένη την εκδήλωση «Απονομή βραβείων "Διδώ Σωτηρίου" και "Δαίδαλος" 2015, της Εταιρείας Συγγραφέων»: https://www.youtube.com/watch?v=Za4hXPr-dqo (και συγκεκριμένα στο 54:14 κ.εξ.) Πρβ. σχετικό άρθρο στο *www.bookia.gr* (με ημεροχρονολογία ανάρτησης 05/01/2016) [20/04/2021].

κατάστασης και το κάθε φιλτράρισμα είναι οι ειδικές τεχνικές για την απόκτηση ενός πιο εξατομικευμένου προϊόντος. Έτσι, προτείνει ότι σε κάθε μεμονωμένο χρήστη ενός λεξικογραφικού ηλεκτρονικού εργαλείου μπορεί να δίνεται η δυνατότητα για εξατομικευμένη αναζήτηση και διαμόρφωση των αποτελεσμάτων, σύμφωνα με τις ανάγκες του, στην οθόνη του (Tarp 2012: 261).

Επομένως, για τη δημιουργία ενός εξειδικευμένου λεξικού βασιζόμαστε στις σύγχρονες θεωρίες της εξειδικευμένης λεξικογραφίας (Bergenholtz & Tarp 1995, Sterkenburg 2003, Nielsen & Tarp 2009, Fuertes-Olivera & Tarp 2014) – με τη χρήση των νέων τεχνολογιών για την ανάπτυξη του λεξικού σε ένα ψηφιακό περιβάλλον. Σύμφωνα με τους Gouws & Tarp (2019), στη σύγχρονη λεξικογραφία η συγκειμενοποίηση και η εξατομίκευση αποτελεί ένα μείζον ζήτημα, καθώς οι λεξικογράφοι είχαν συχνά ανεκπλήρωτα όνειρα για νέες δυνατότητες στο ψηφιακό περιβάλλον. Έτσι, η εμφάνιση των διαδικασιών προσαρμογής της συγκειμενοποίησης και της εξατομίκευσης σε διαφορετικές περιοχές και περιβάλλοντα καταδεικνύει πως οι διαδικασίες αυτές εισάγουν μια νέα λεξικογραφική πρακτική, ακόμα και σε περιβάλλοντα, όπως για παράδειγμα στη λογοτεχνία.

Με βασικό γνώμονα, λοιπόν, το γεγονός ότι η λογοτεχνία εγγενώς αποτελεί μια εξειδικευμένη γλώσσα (ποιητική γλώσσα), της οποίας το λεξιλόγιο αποτελεί μέρος της εξειδικευμένης λεξικογραφίας για ειδικές ανάγκες (ποιητικοί νεολογισμοί/ποιητικές λέξεις), η σύνταξη ενός πρότυπου λεξικού αποτελεί βασικό εργαλείο τόσο για την ανάγνωση και την κατανόηση των λογοτεχνικών κειμένων όσο και για τη διδασκαλία της λογοτεχνίας γενικότερα. Ειδικότερα, η προσέγγιση των ποιητικών νεολογισμών και η λημματοποίησή τους πρόκειται να φέρει στο φως πλήθος γλωσσολογικών πληροφοριών για τους μηχανισμούς της ελληνικής γλώσσας, κυρίως σε μορφολογικό αλλά και σημασιολογικό επίπεδο.

3.2 Λεξικά και γλωσσάρια ελλήνων λογοτεχνών

Η λεξιλογική προσέγγιση της νεοελληνικής λογοτεχνίας, και κυρίως από λεξικογραφική σκοπιά, αποτελεί ένα ζήτημα που δεν έχει αναπτυχθεί ιδιαιτέρως, αλλά ούτε και με επιστημολογικά συστηματικό τρόπο. Ο Χαραλαμπάκης, αφενός, εύστοχα έχει επισημάνει ότι «λιγοστά είναι τα λεξικά ή οι πίνακες λέξεων μεμονωμένων συγγραφέων και ακόμα λιγότερες οι μελέτες που αναφέρονται στους "διαλεκτισμούς", τα διαλεκτικά δηλ. ή ιδιωματικά στοιχεία που χρησιμοποιούν λίγο ή πολύ όλοι σχεδόν οι λογοτέχνες μας» (2001: 170). Ο Καψωμένος (2004), αφετέρου, θεωρεί ότι η συστηματική μελέτη και η δημιουργία βάσεων δεδομένων με την αξιοποίηση της ηλεκτρονικής τεχνολογίας βοηθά στην ανάπτυξη της γλωσσικής στατιστικής και της υφολογίας, ανανεώνοντας το θεωρητικό και μεθοδολογικό υπόβαθρο για την έρευνα και την ανάλυση του ύφους ενός λογοτεχνικού κειμένου.

Έτσι, από τη μία πλευρά, μέχρι σήμερα, διατίθενται πίνακες λέξεων για αρκετούς σημαντικούς συγγραφείς και κείμενα της νεοελληνικής λογοτεχνίας (Γιαννίκου et al. 2007), όπως – γίνεται ενδεικτική αναφορά σε ορισμένους με αλφαβητική σειρά – για τον Εγγονόπουλο (Κουμπής 1999), τον Ελύτη (Μαυρομάτης 1981), τον Καβάφη (Lorando et al. 1970, Κοκόλης 1976), τον Κάλβο (Gentilini 1970), τον Καρυωτάκη (Pelacchi 1971, Peri 1983), τον Μακρυγιάννη (Κυριαζίδης et al. 1992), τον Σεφέρη (Κοκόλης 1975), τον Σολωμό (Καψωμένος et al. 1983), καθώς και συμφραστικοί πίνακες για τη *Θυσία του Αβραάμ* (Φιλιππίδου 1986), τον *Διγενή Ακρίτα* (Beaton et al. 1995), τον *Ερωτόκριτο* του Κορνάρου (Φιλιππίδου & Holton 1996), το ποιητικό έργο του Σεφέρη (Καζάζης & Σιστάκου 2003). Και φυσικά σε ψηφιακό περιβάλλον μοναδική αξιομνημόνευτη περίπτωση είναι η ψηφιακή πλατφόρμα[3] *Ανεμόσκαλα* (υπό τη σκέπη του Κέντρου Ελληνικής Γλώσσας), η οποία παρουσιάζει συμφραστικούς πίνακες λέξεων για μείζονες νεοέλληνες ποιητές.

Όμως, από την άλλη πλευρά, και συγκριτικά με ανάλογες περιπτώσεις ξένων λογοτεχνών, εξαιρετικά περιορισμένες είναι οι συστηματικές προσπάθειες καταγραφής σε εξειδικευμένα λεξικά του λεξιλογίου συγκεκριμένων ελλήνων συγγραφέων. Στην ελληνική βιβλιογραφία υπάρχει μόνο το *Λεξικό Σολωμού* του Καψωμένου (1983), που ουσιαστικά είναι ένας πίνακας λέξεων του ελληνόγλωσσου σολωμικού έργου, και το *Γλωσσάρι στο έργο του Νίκου Καββαδία* (με ερμηνευτικά σχόλια) του Τράπαλη (2010). Σε ψηφιακό περιβάλλον έχουμε μόνο το γλωσσάρι του Καββαδία, όπως παρουσιάζεται από τον Τράπαλη, βρίσκεται σε ιστότοπο[4] υπό τη σκέπη του Πανεπιστημίου Αθηνών, στον οποίο παρουσιάζονται οι λέξεις και οι σημασίες τους με τη μορφή αλφαβητικού καταλόγου, καθώς και ορισμένα τοπωνύμια. Οι υπόλοιπες τυχόν λεξιλογικές προσεγγίσεις είναι ερασιτεχνικού χαρακτήρα και αναδημοσιεύονται σε προσωπικά ιστολόγια.

Και συγκεκριμένα, στην περίπτωση του Καζαντζάκη, δεν έχουμε πολλές έρευνες λεξιλογικού ή/και λεξικογραφικού ενδιαφέροντος. Ειδικότερα, η ΟΔΥΣΕΙΑ του Καζαντζάκη μελετήθηκε αρκετά, κυρίως με προσπάθειες αποτύπωσης γλωσσικών και υφολογικών στοιχείων (ενδεικτικά αναφέρονται Bien 1972, Γιακουμάκη 1982, Χαραλαμπάκης 2001, 2010 & Μαθιουδάκης 2012), αλλά όμως δεν έχει ερευνηθεί λεξιλογικά σε μεγάλη έκταση. Ο λεξικογραφικός πλούτος του έργου παρουσιάζεται μερικώς από τον Πρεβελάκη (1932, στο Μαρινάκης 2004), ο οποίος ασχολήθηκε με τις ραψωδίες Α-Κ, και από τη Γιακουμάκη (1982) η οποία ασχολήθηκε με τις ραψωδίες Α-Δ, ενώ ο Μαθιουδάκης (2012) παρουσιάζει τη μοναδική ολοκληρωμένη λεξικογραφική μελέτη για το καζαντζακικό έπος, σε επίπεδο διδακτορικής έρευνας. Τέλος, πέρα από τις ερευνητικές προσεγγίσεις των μελετητών, έχουμε και μια λεξικογραφική τοποθέτηση από τον ίδιο τον Καζαντζάκη, ο οποίος στην πρώτη έκδοση του έργου (1938) επισυνάπτει και ένα «Λεξιλόγιο», ένα γλωσσάρι με περίπου 1.500 «άγνωστες» λέξεις.

3.3 Λεξικά και γλωσσάρια ξένων λογοτεχνών

Στο σημείο αυτό πρέπει να αναφέρουμε, έστω και ακροθιγώς, ορισμένες έρευνες στον διεθνή χώρο, μετά την εξέταση του ζητήματος στον ελληνικό χώρο, αναφορικά με λεξικά και γλωσσάρια ξένων καταξιωμένων λογοτεχνών. Ενδεικτικά, παρουσιάζονται εξειδικευμένα λεξικά για τον C. Dickens (Sutherland 2012), τον J. Joyce (Chenier online), τον W. Shakespeare (Schmidt 1902, Wells 1998, Onions 1986, Crystal & Crystal 2002), τον W. B. Yeats (Conner 1999) και αρκετές

[3] Βλ.: https://www.greek-language.gr/Resources/literature/tools/concordance/index.html [20/04/2021].

[4] Βλ.: http://users.uoa.gr/~nektar/arts/tributes/nikos_kabbadias/meletes_trapalhs_glossari_a.htm [20/04/2021].

άλλες.
Σε ψηφιακό περιβάλλον αναφέρονται ενδεικτικά τρία παράδειγμα: του L. Carroll, του C. Dickens και του W Shakespeare. Πρώτο παράδειγμα: Δύο πολύ συνοπτικά γλωσσάρια αναφορικά με το έργο *Η Αλίκη στη χώρα των θαυμάτων* του Carroll παρουσιάζονται στον ιστότοπο *Alice in Wonderland*[5] και στον ιστότοπο *Alice in Wonderland-Wiki*[6]. Δεύτερο παράδειγμα: Ένα ενδεικτικό λεξιλόγιο του Dickens καταγράφεται σε δύο γλωσσάρια: αφενός στον ιστότοπο *The Charles Dickens page*[7] και αφετέρου στον ιστότοπο *Charles Dickens, Victorian Literature and Vocabulary*[8]. Τρίτο παράδειγμα: ίσως η πιο ενδιαφέρουσα περίπτωση απόδοσης ψηφιακού λεξικού αφορά τον Shakespeare· δύο εκτεταμένα γλωσσάρια που βασίζονται σε δύο έγκριτες παλαιότερες εκδόσεις για το έργο του άγγλου βάρδου. Επομένως, έχουμε αφενός ένα «λεξικό» που βασίζεται στον Schmidt (1902), όπως παρουσιάζεται στον ιστότοπο *Perseus*[9] του Tufts University, και αφετέρου ένα «γλωσσάρι» που βασίζεται στους Crystal & Crystal (2002), όπως παρουσιάζεται στον ιστότοπο *Shakespeare's Words*[10]. Αξίζει να σημειωθεί πως μονάχα το ψηφιακό γλωσσάρι των Crystal & Crystal παρουσιάζει ιδιαίτερο ενδιαφέρον και αποτελεί χρηστικό εργαλείο, καθώς έχει σαφή επιστημονικά κριτήρια, ενώ το κάθε λήμμα περιέχει σημασία και διασυνδέεται με τους στίχους στους οποίους εμφανίζεται η λέξη.

4 Μεθοδολογία

Σύμφωνα με τον Ξυδόπουλο (2008) και τη Μότσιου (1994), ένα λεξικό ή ένα λεξιλόγιο ή γενικά ένα λεξικογραφικό σώμα δεδομένων αποτελείται κυρίως από δύο μέρη, τη μακροδομή και τη μικροδομή, όσον αφορά στην αναλυτική περιγραφή της δομής και του περιεχομένου του. Η *μακροδομή* (*macrostructure*), από τη μια πλευρά, περιλαμβάνει τον κατάλογο των λημμάτων που περιέχονται στο συγκεκριμένο λεξικό και περιγράφει τις σημαντικές πληροφορίες που αφορούν στο σύνολο των λημμάτων, αποδίδοντας τα βασικά χαρακτηριστικά τους. Η *μικροδομή* (*microstructure*), από την άλλη πλευρά, είναι η εσωτερική δομή και οι πληροφορίες του κάθε λήμματος, δηλαδή της κάθε *λέξης-κεφαλής* (*headword*), η οποία ουσιαστικά αναφέρεται στον κανονικό τύπο μιας λέξης μέσα σε έναν κατάλογο. Στη συγκεκριμένη μελέτη χρησιμοποιείται ο όρος *λήμμα* (*lemma*) για να προσδιοριστεί ο λιγότερο χαρακτηρισμένος ή βασικός τύπος μιας λέξης ή εναλλακτικά ο όρος *λέξη-κεφαλή* (*headword*).

4.1 Ψηφιακή ανάπτυξη λεξικού

Το ψηφιακό λεξικό της ΟΔΥΣΕΙΑΣ του Νίκου Καζαντζάκη αποτελεί ένα πρωτοποριακό εργαλείο και πρόκειται να αποτελέσει έναν εύχρηστο και λειτουργικό ψηφιακό κόμβο αναφοράς τόσο για την ερευνητική όσο και την εκπαιδευτική κοινότητα. Η πλατφόρμα υλοποιείται με χρήση ανοικτών λογισμικών (open-source software) βασισμένα σε σύγχρονες τεχνολογίες κατασκευής θησαυρών λημμάτων και άλλων τεχνολογιών του σημασιολογικού ιστού (semantic web), με ενσωματωμένες λειτουργίες επισημείωσης, τεκμηρίωσης (in-context) και αναζήτησης των λημμάτων.
Σύμφωνα με το πρότυπου ψηφιακής επισημείωσης/τεκμηρίωσης (mark-up) και υποδομών υποστήριξης της κωδικοποίησης της πληροφορίας, το περιεχόμενο του λεξικού προετοιμάζεται με κωδικοποίηση (encoding) των λημμάτων και σύνδεση του κάθε λήμματος με το απόσπασμα αναφοράς (contextualisation). Οι εργασίες σχεδιασμού και κατασκευής[11] χωρίζονται σε τρεις ενότητες:

(α) «Αρχιτεκτονική της Πληροφορίας» (Information Architecture): η οποία αφορά το επίπεδο οργάνωσης των δεδομένων και της δομικής τους απεικόνισης, καθώς και στη μελέτη των αναλυτικών σεναρίων χρήσης (ανάγνωση, αναζήτηση, διαμοιρασμός κ.ο.κ.) και λειτουργικών/τεχνικών απαιτήσεων και προδιαγραφών.

(β) «Σχεδιασμός» (Design): η οποία αφορά τον εικαστικό και γραφιστικό σχεδιασμό της πλατφόρμας, για τη δημιουργία της οπτικής της ταυτότητας, αλλά και τη μελέτη της εμπειρίας χρήσης. Για το σχεδιασμό θα μελετηθούν θέματα τυπογραφίας, χρωμάτων, χαράξεων, σχεδιαστικών στοιχείων, προσαρμογής σε πολλαπλές οθόνες χρήσης και σενάρια χρήσης στο σχεδιαστικό περιβάλλον.

(γ) «Ανάπτυξη» (Development): η οποία αφορά την κατασκευή της τεχνικής υποδομής και την υλοποίηση της πλατφόρμας στο επιλεγμένο περιβάλλον ανάπτυξης. Στο πλαίσιο την ενότητας αυτής θα υλοποιηθούν όλες οι εργασίες προγραμματισμού και παραμετροποίησης του συστήματος, με βάση τις προδιαγραφές που θα έχουν προκύψει από τα αποτελέσματα των εργασιών της «Αρχιτεκτονικής της Πληροφορίας» και του «Σχεδιασμού», σε συνδυασμό με τις λειτουργικές και τεχνικές απαιτήσεις που ορίζουν οι διεθνείς πρακτικές ανάπτυξης ψηφιακών θησαυρών όρων και λημμάτων.

Σημαντικό είναι να σημειωθεί πως, ακολουθώντας το παράδειγμα του ιστότοπου *Shakespeare's Words* – *www.shakespeareswords.com*, το σχεδιαζόμενο ψηφιακό λεξικό για τον Νίκο Καζαντζάκη πρόκειται να αναπτυχθεί σε έναν δυναμικό ιστότοπο με domain name *kazantzakiswords.gr*, σύμφωνα με την ακόλουθη μακροδομή και μικροδομή.

[5] Βλ.: *https://www.alice-in-wonderland.net/resources/background/glossary/* [20/04/2021].
[6] Βλ.: *https://aliceinwonderland.fandom.com/wiki/Glossary_of_Alice_in_Wonderland_Terms* [20/04/2021].
[7] Βλ.: *https://www.charlesdickenspage.com/charles-dickens-glossary.html* [20/04/2021].
[8] Βλ.: *https://victorianvocabulary.weebly.com/dickens-glossary.html* [20/04/2021].
[9] Βλ.: *http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.03.0079* [20/04/2021].
[10] Βλ.: *https://www.shakespeareswords.com/Public/Glossary.aspx* [20/04/2021].
[11] Οι λεπτομέρειες του σχεδιασμού και της κατασκευής του ψηφιακού λεξικού προέκυψαν μετά από συζήτηση με την ερευνητική ομάδα του *thinking.gr*.

4.2 Μακροδομή

Η μακροδομή περιλαμβάνει περίπου 5.000 λήμματα, τα οποία είναι ταξινομημένα κατά απόλυτη αλφαβητική σειρά. Ο συνολικός αριθμός των λημμάτων αποτελείται από δύο βασικές κατηγορίες, (α) αυτή των ρηματικών τύπων και (β) των ονοματικών τύπων, καθώς και από μία τρίτη κατηγορία (γ) αυτή των επιρρημάτων. Ο προσδιορισμός των λεξικών τύπων ως προς τα μέρη του λόγου έγινε σύμφωνα με τη λειτουργία τους μέσα στο λογοτεχνικό περιβάλλον στο οποίο βρίσκονται. Οι ρηματικοί τύποι ως μέρη του λόγου συμπεριλαμβάνουν τα ρήματα και τις μετοχές που έχουν ρηματικό χαρακτήρα. Οι ονοματικοί τύποι ως μέρη του λόγου διακρίνονται σε δύο υποκατηγορίες, στα ουσιαστικά και στα επίθετα. Στα επίθετα, επίσης, εντάσσονται – για καθαρά συμβατικούς λόγους – και οι μετοχές που έχουν επιθετικό χαρακτήρα. Τέλος, τα επιρρήματα ως μέρη του λόγου δημιουργούν μεγάλο προβληματισμό στον καθορισμό τους, εξαιτίας δυσκολιών ακριβούς σύνταξης που δημιουργούνται λόγω της έντονης ποιητικότητας του κειμένου, όπου σε αρκετές περιπτώσεις υπάρχει ασάφεια ως προς το αν ο λεξικός τύπος είναι επίθετο σε παρακείμενο ουσιαστικό, ή επίρρημα.

Τα λήμματα αποτελούν το σύνολο των τύπων λέξεων που συλλέχτηκαν μέσα από την ΟΔΥΣΕΙΑ του Νίκου Καζαντζάκη με τη διαδικασία της ανάγνωσης λέξη προς λέξη και είναι μορφολογικοί σχηματισμοί, οι οποίοι δεν εντοπίστηκαν σε κάποιο από τα βασικά λεξικά της Κοινής Νέας Ελληνικής. Ο λεξικογραφικός έλεγχος έγινε αρχικά[12] στο *Λεξικόν της Νέας Ελληνικής Γλώσσης* (1933) της Πρωίας, το *Μέγα λεξικόν όλης της ελληνικής γλώσσης* (1964) του Δ. Δημητράκου, το *Λεξικό της Κοινής Νεοελληνικής* (1998) του Μ Τριανταφυλλίδη και το *Λεξικό της Νέας Ελληνικής Γλώσσας* (2005) του Γ. Μπαμπινιώτη και στη συνέχεια[13] στο *Λεξικόν της Νέας Ελληνικής Γλώσσης* (1971) του Ι. Σταματάκου, το *Νέο ελληνικό λεξικό της σύγχρονης δημοτικής γλώσσας* (1995) του Ε. Κριαρά, το *Χρηστικό λεξικό της Νεοελληνικής Γλώσσας* (2014) του Χ. Χαραλαμπάκη (Ακαδημίας Αθηνών). Κατά συνέπεια, θεωρούμε ότι αποτελούν μη κωδικοποιημένες λέξεις (Αναστασιάδη-Συμεωνίδη 1986), οι οποίες δεν έχουν καταγραφεί σε αυτά τα λεξικά, ή, με άλλα λόγια, λέξεις που δεν έχουν ενσωματωθεί στο δυναμικό του λεξιλογίου της νέας ελληνικής γλώσσας και δεν έχουν καθιερωθεί στην καθημερινή ομιλία.

Επομένως, τα λήμματα του συγκεκριμένου εξειδικευμένου λεξικού αποτελούν μια μορφή νεολογισμών (neologisms), αφού είναι αθησαύριστες λέξεις (undictionaried words), όπως σημειώνει η Αναστασιάδη-Συμεωνίδη (1986: 53), στην κατηγορία των νεολογισμών είναι δυνατόν να ανήκει κάθε είδους άγνωστη λέξη, την οποία συναντά ένας αναγνώστης σε ένα κείμενο (λογοτεχνικό ή μη) παλαιότερων εποχών. Επιπροσθέτως, σημειώνεται πως τα λήμματα χαρακτηρίζονται με έντονο νεολογικό χαρακτήρα, ο οποίος δικαιολογείται εξαιτίας της λεξικογραφικής έρευνας που πραγματοποιήθηκε σε επτά από τα σημαντικότερα λεξικά της Κοινής Νέας Ελληνικής.

4.3 Μικροδομή

Η μικροδομή [14] εμπεριέχει τη λεπτομερή περιγραφή των λεξικογραφικών πληροφοριών του λήμματος, οι οποίες διακρίνονται σε δέκα μέρη (Σχήμα 1) και είναι: (α) το λήμμα, (β) το μέρος του λόγου, (γ) το ερμήνευμα, (δ) η μορφολογική ανάλυση, (ε) η συχνότητα εμφάνισης, (στ) ο τύπος της λέξης, (ζ) η θέση, (η) το συγκείμενο, (θ) ο κωδικός και (ι) η αγγλική μετάφραση του λήμματος. Πιο συγκεκριμένα, οι πληροφορίες που παρέχονται σε κάθε λήμμα-λέξη είναι:

(1) *Λήμμα* ή *κύριο λήμμα* ή *λέξη-κεφαλή* (*lemma* or *headword*), που αποτελεί τον βασικό ή κανονικό τύπο της λέξης και είναι με έντονη γραφή (bold).

(2) *Μέρος του λόγου* (*part of speech*), που ανήκει ο τύπος λέξης, σύμφωνα με το συγκεκριμένο λήμμα.

(3) *Ερμήνευμα* (*explanation*), που είναι η σημασία/ερμηνεία του λήμματος.

(4) *Μορφολογική ανάλυση* (*morphological analysis*), που είναι η ανάλυση του λήμματος στα μορφολογικά συστατικά του και είναι με πλάγια γραφή (italics). (Στο συγκεκριμένο πεδίο σημειώνεται και πιθανή παραπομπή σε άλλο λήμμα του λεξικού.)

(5) *Συχνότητα εμφάνισης* (*frequency*), που αποτελεί τον αριθμό συχνότητας εμφάνισης του λέξης μέσα στο ποίημα.

(6) *Λεκτικός τύπος* ή *τύπος της λέξης* (*word-form*), όπως αυτός βρέθηκε μέσα στην ΟΔΥΣΕΙΑ του Καζαντζάκη.

(7) *Θέση* (*position*), που είναι το ακριβές σημείο εμφάνισης της λέξης. Η περιγραφή *Έργο* (*work*) αναφέρεται στην ΟΔΥΣΕΙΑ. Η περιγραφή *Ραψωδία* (*rhapsody*) αναφέρεται στο σημείο μέσα στο ποίημα που βρίσκεται ο συγκεκριμένος τύπος λέξης, σύμφωνα με το καθορισμένο παράδειγμα, αναγράφοντας το γράμμα της ραψωδίας από το Α έως το Ω, ενώ η περιγραφή *Στίχος* (*line*) στην αριθμημένη γραμμή της κάθε ραψωδίας μέσα στο ποίημα που βρίσκεται ο συγκεκριμένος τύπος λέξης, σύμφωνα με το καθορισμένο παράδειγμα.

(8) *Συγκείμενο* (*context*)[15], το οποίο είναι το περιβάλλον κείμενο ή περικείμενο ή τα συμφραζόμενα, που βρίσκεται ενσωματωμένος ο συγκεκριμένος τύπος λέξης σε επίπεδο ενός μετρικού στίχου.

(9) *Κωδικός* (*code*), ο οποίος είναι η κωδικοποίηση για την άμεση διασύνδεση του λεκτικού τύπου με τον συγκεκριμένο στίχο μέσα στο υπόλοιπο ποίημα. Η κωδικοποίηση γίνεται σε λατινικό αλφάβητο και με τον

[12] Όπως σημειώνει ο Μαθιουδάκης 2012 & 2020.

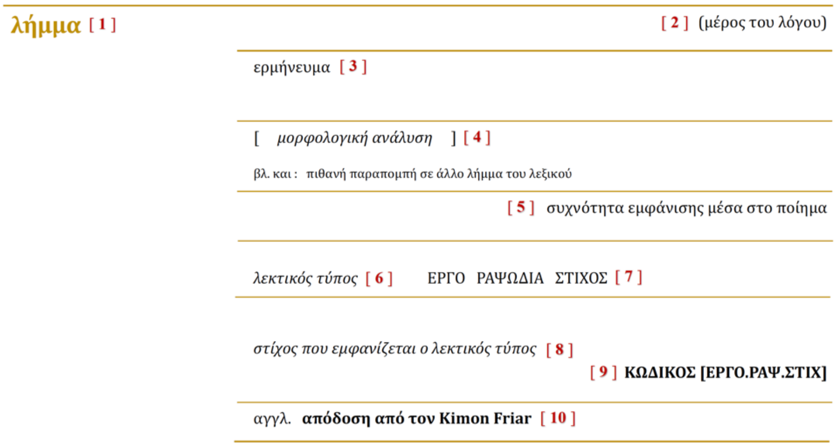
[13] Πιθανότατα να γίνει λεξικογραφικός έλεγχος των λημμάτων και στο *Αναλυτικόν ορθογραφικόν λεξικόν της νεοελληνικής γλώσσης* (1967) του Θ. Βοσταντζόγλου, καθώς και στο Αρχείο του Ιστορικού Λεξικού της Ακαδημίας Αθηνών.

[14] Η μικροδομή του λεξικού βασίστηκε σε μεγάλο βαθμό στη δημιουργία δομής για το «Σώμα Αθησαύριστων Λέξεων» (ΣΑΛ) της ΟΔΥΣΕΙΑΣ του Νίκου Καζαντζάκη· βλ. Μαθιουδάκης 2012 & 2020.

[15] Οι στίχοι παρουσιάζονται σύμφωνα με την επικρατούσα έκδοση του έργου (Καζαντζάκης 1967 κ.εξ.).

ακόλουθο σχηματισμό: ΕΡΓΟ.ΡΑΨ.ΣΤΙΧ. Το πρώτο μέρος σηματοδοτεί το έργο με την ένδειξη ODI., το δεύτερο μέρος τη ραψωδία με μια ένδειξη I-XXIV (ένα έως είκοσι τέσσερα σε λατινικούς αριθμούς) και το τρίτο μέρος τον στίχο με την ένδειξη σε αραβικούς αριθμούς (1, 2, 3, 4, 5 κ.ο.κ).

(10) *Αγγλική μετάφραση* (*english translation*), που είναι η απόδοση του λήμματος στην αγγλική έκδοση της ΟΔΥΣΕΙΑΣ του Νίκου Καζαντζάκη από τον Kimon Friar (1958).



λήμμα [1]

[2] (μέρος του λόγου)

ερμήνευμα [3]

[*μορφολογική ανάλυση*] [4]

βλ. και : πιθανή παραπομπή σε άλλο λήμμα του λεξικού

[5] συχνότητα εμφάνισης μέσα στο ποίημα

λεκτικός τύπος [6] ΕΡΓΟ ΡΑΨΩΔΙΑ ΣΤΙΧΟΣ [7]

στίχος που εμφανίζεται ο λεκτικός τύπος [8]

[9] **ΚΩΔΙΚΟΣ [ΕΡΓΟ.ΡΑΨ.ΣΤΙΧ]**

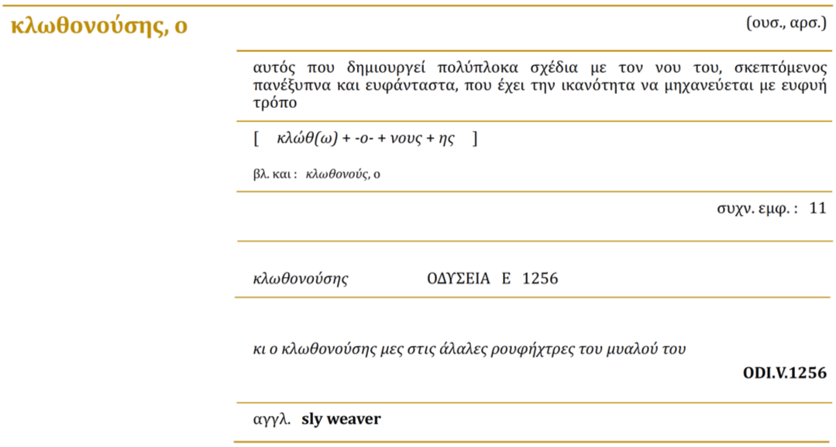
αγγλ. **απόδοση από τον Kimon Friar** [10]

Σχήμα 1: Συμβατική απεικόνιση μικροδομής λήμματος.

5 Παραδείγματα λημμάτων

Ακολουθώντας τη μικροδομή του ψηφιακού λεξικού (Σχήμα 1), σημειώνονται ενδεικτικά τρία λήμματα από τους ποιητικούς νεολογισμούς της ΟΔΥΣΕΙΑΣ του Νίκου Καζαντζάκη. Οι συγκεκριμένες λέξεις αποτελούν τρεις χαρακτηρισμούς προς τον κεντρικό ήρωα του έπους, τον Οδυσσέα, ο οποίος σκιαγραφείται με περισσότερους από διακόσια[16] επιθετικούς προσδιορισμούς μέσα στο ποίημα (Πρεβελάκης 1958, Μαθιουδάκης & Καμπάκη-Βουγιουκλή 2011).

Συγκεκριμένα, τα τρία λήμματα είναι «κλωθονούσης» (Σχήμα 2), «κοσμοπαρωρίτης» (Σχήμα 3) και «χαρομάχος» (Σχήμα 4), τα οποία φανερώνουν την ιδιότυπη γλωσσοπλαστική διάθεση του Νίκου Καζαντζάκη, και σχηματικά παρουσιάζονται όπως παρακάτω:



κλωθονούσης, ο

(ουσ., αρσ.)

αυτός που δημιουργεί πολύπλοκα σχέδια με τον νου του, σκεπτόμενος πανέξυπνα και ευφάνταστα, που έχει την ικανότητα να μηχανεύεται με ευφυή τρόπο

[*κλώθ(ω) + -ο- + νους + ης*]

βλ. και : *κλωθονούς, ο*

συχν. εμφ. : 11

κλωθονούσης ΟΔΥΣΕΙΑ Ε 1256

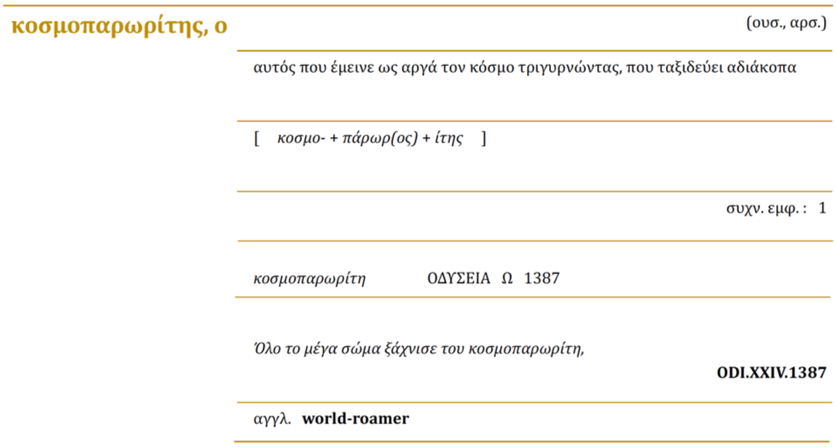
κι ο κλωθονούσης μες στις άλαλες ρουφήχτρες του μυαλού του

ODI.V.1256

αγγλ. **sly weaver**

Σχήμα 2: Συμβατική απεικόνιση του λήμματος «κλωθονούσης».

[16] Πρεβελάκης, 1958: 312, σχόλιο 177.



κοσμοπαρωρίτης, ο (ουσ., αρσ.)

αυτός που έμεινε ως αργά τον κόσμο τριγυρνώντας, που ταξιδεύει αδιάκοπα

[*κοσμο- + πάρωρ(ος) + ίτης*]

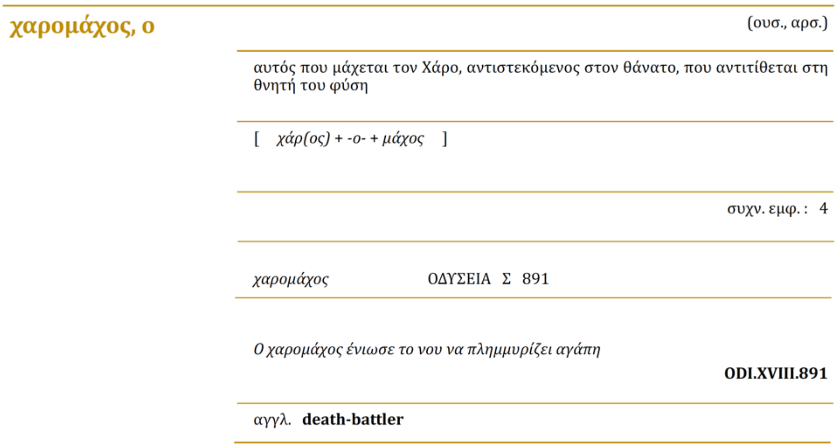
συχν. εμφ. : 1

κοσμοπαρωρίτη ΟΔΥΣΕΙΑ Ω 1387

Όλο το μέγα σώμα ξάχνισε του κοσμοπαρωρίτη,
ODI.XXIV.1387

αγγλ. **world-roamer**

Σχήμα 3: Συμβατική απεικόνιση του λήμματος «κοσμοπαρωρίτης».



χαρομάχος, ο (ουσ., αρσ.)

αυτός που μάχεται τον Χάρο, αντιστεκόμενος στον θάνατο, που αντιτίθεται στη θνητή του φύση

[*χάρ(ος) + -ο- + μάχος*]

συχν. εμφ. : 4

χαρομάχος ΟΔΥΣΕΙΑ Σ 891

Ο χαρομάχος ένιωσε το νου να πλημμυρίζει αγάπη
ODI.XVIII.891

αγγλ. **death-battler**

Σχήμα 4: Συμβατική απεικόνιση του λήμματος «χαρομάχος».

6 Επίλογος

Το φαινόμενο της ποιητικής νεολογίας αποτελεί ένα ζήτημα διαχρονικό το οποίο αναπτύσσεται και μεταπλάθεται σύμφωνα με τις ανάγκες της κάθε εποχής. Συνεχώς συντίθενται νέες λέξεις μορφολογικά (ή ακόμα και ήδη υπάρχουσες λέξεις αλλάζουν σημασία), προκειμένου να ικανοποιήσουν το συγγραφικά γλωσσικό αισθητήριο του κάθε λογοτέχνη, ποιητή ή πεζογράφο. Σε άμεση αντίστιξη, βρίσκεται η ανάγκη δημιουργίας εξειδικευμένων λεξικών ή/και γλωσσαρίων για να καταγραφούν οι ιδιότυποι λογοτεχνικοί σχηματισμοί, καθώς αποτελούν μέρος της καθημερινότητάς μας, ενώ μπορούν να αποτελέσουν ένα ανεξερεύνητο πεδίο μελέτης, αποκαλύπτοντας στοιχεία γλωσσολογίας και υφολογίας για το υπό διερεύνηση λογοτεχνικό κείμενο, αλλά και αναδεικνύοντας στοιχεία για την ίδια τη γλώσσα μας και τους μηχανισμούς της, ειδικότερα στο επίπεδο της παραγωγής και της σύνθεσης.

Στην περίπτωση της καζαντζακικής ΟΔΥΣΕΙΑΣ, παρά το πλήθος των σύνθετων και πολυσύνθετων ποιητικά νεολογικών σχηματισμών, ο κρητικός λογοτέχνης ισχυρίζεται πως είχε δημιουργήσει μόλις μερικές νέες λέξεις (Καζαντζάκης 1938, Πρεβελάκης 1958) και μάλιστα μόνο στις περιπτώσεις που χρειαζόταν να δημιουργηθεί μια νέα λέξη. Αντίθετα, αρκετοί μελετητές (Ανδριώτης 1959, Γιακουμάκη 1982, Μαθιουδάκης 2020) σημειώνουν ότι η ΟΔΥΣΕΙΑ έχει πλήθος ιδιολέκτων, γλωσσοπλαστικών γεννημάτων νεολογισμών. Επιπρόσθετα, ο Καζαντζάκης συνεχώς τονίζει ότι οι περισσότερες λέξεις είναι από το στόμα του λαού, γι’ αυτό κι έχουν μια τέτοια πλαστικότητα (Πρεβελάκης 1958 & 1984, Μαθιουδάκης 2020), ενώ προτιμά τη χρήση ιδιωματικών τύπων ως πιο έντονα εκφραστικούς. Χρησιμοποιεί τις λέξεις με μια ιδιαίτερη αγάπη ως φορείς των ξεχωριστών του νοημάτων. Προτιμά να συνθέτει και να παράγει λέξεις, γιατί γοητεύεται από την σπανιότητα της ύπαρξής τους. Ρήματα, επίθετα, ουσιαστικά, επιρρήματα γεννιούνται και αναγεννιούνται από ένα λογοτεχνικό μυαλό που απέχει πολύ από την επιστήμη της γλωσσολογίας – όπως αυτή αναπτύσσεται θεωρητικά –, αλλά ο Καζαντζάκης παρουσιάζεται ως ένας γλωσσολόγος του δρόμου, ένας γλωσσοταξιδευτής που συλλέγει και δημιουργεί λέξεις, όχι αυθαίρετα, αλλά με πάθος: ένα πάθος που πηγάζει από την πραγματική αγάπη του για τη γλώσσα.

Σε αριθμό 5.000 αθησαύριστοι ποιητικοί νεολογισμοί εγκιβωτίζονται στο ποιητικό έπος, με κύριο στόχο τη διάσωσή τους και την απόδοσή τους ως παρακαταθήκη στις μελλούμενες γενιές, καθιστώντας με τον τρόπο αυτό το έπος του θησαυρό λεξιλογικού πλούτου. Ο Καζαντζάκης προτρέχει να μας πληροφορεί σε διάφορες επιστολές του πως «κούρσεψε» τις λέξεις του από τα ταξίδια του, αποδίδοντας τις ρίζες τους στην ατόφια δημοτική με έντονη λαϊκή διάθεση. Για την πράξη του αυτή τον ονοματίζω «λαϊκό γλωσσολόγο»[17] με αγάπη για τη νεοελληνική γλώσσα και την πατρίδα του. Ο Καζαντζάκης προτρέχει να μας πληροφορεί στη διατύπωση του γλωσσ(ολογ)ικού του «πιστεύω» πως «πέντε ή έξι μονάχα αναγκάστηκ[ε] να φκιάσ[ει]», θεωρώντας πως η χρήση των μηχανισμών της γλώσσας μας για την παραγωγή και τη σύνθεση νέων λέξεων δεν αποτελεί πρωτότυπη δημιουργία. Στην ίδια διατύπωση μας πληροφορεί πως οι λέξεις είναι «παρμένες απ' όλα τα μέρη της Ελλάδας», προσπαθώντας να δημιουργήσει μια πανελλήνια γλώσσα με πανδιαλεκτική διάθεση. Για τις πράξεις του αυτές τον θεωρώ «λεξιδημιουργό» και «γλωσσοποιό»[18], καθώς δεν είναι (αφού ούτε ο ίδιος θεωρούσε τον εαυτό του) γλωσσοπλάστης, με τη στενή σημασία του όρου.

Παρά το γεγονός πως ο ίδιος θεωρεί τον εαυτό του περισσότερο συλλέκτη-καταγραφέα παρά γλωσσοπλάστη-δημιουργό, αναρωτιόμαστε ποιο επιχείρημα θα έβρισκε ο Καζαντζάκης, αν ζούσε σήμερα, να μας πείσει ότι οι 5.000 περίπου ποιητικοί νεολογισμοί, που έχουμε αποθησαυρίσει μελετώντας την ΟΔΥΣΕΙΑ είναι όλες παρμένες από το στόμα του λαού; Κάποιες ίσως να είναι, αλλά οι περισσότερες πιστεύουμε ότι είναι δημιουργήματα του ποιητή. Πιθανολογούμε, λοιπόν, ότι ο Καζαντζάκης δεν θεωρούσε γλωσσοπλαστικές και δικές του τις λέξεις που έφτιαχνε σύμφωνα με τους κανόνες σύνθεσης και παραγωγής λέξεων.

7 Βιβλιογραφικές αναφορές

Αναστασιάδη-Συμεωνίδη, Α. (1986). *Η νεολογία στην κοινή νεοελληνική*. Θεσσαλονίκη: Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής, ΑΠΘ.

Ανδριώτη, Ν. Π. (1959). Η γλώσσα του Καζαντζάκη. Στο *Νέα Εστία*, τχ. 779, (Χριστούγεννα 1959), σσ. 90-95.

Apperson, G. L. (2015). *A Jane Austen Dictionary*. UK: Cambridge University Press.

Aitchison, J. (2006). *Γιατί αλλάζει η γλώσσα. Πρόοδος ή παρακμή;*, (μτφρ. Ν. Βέργης). Αθήνα: Εκδόσεις Πατάκη.

Beaton, R., Kelly, J., & Λεντάρη, Τ. (1995). *Πίνακας συμφραζομένων του Διγενή Ακρίτη. Σύνταξη Ε΄*. Ηράκλειο: Πανεπιστημιακές Εκδόσεις Κρήτης.

Bergenholtz, H., & Tarp, S. (eds.). (1995). *Manual of Specialised Lexicography: The preparation of specialised dictionaries*. Amsterdam: John Benjamins Publishing Company.

Γιαννίκου, Α., Μπαλτζή, Σ., Παπαϊωάννου, Δ., Πετρωτού, Χ., Παπαστεφάνου, Β., & Ρίζου, Β. (2007). Αρχείο Ελλήνων Συγγραφέων. Στο *5ο Διεπιστημονικό Διαπανεπιστημιακό Συνέδριο του ΕΜΠ και του ΜΕΚΔΕ του ΕΜΠ*. Μέτσοβο: Μετσόβιο Κέντρο Διεπιστημονικής Έρευνας ΜΕΚΔΕ του ΕΜΠ.

Chenier, N. Online. *Joyce Word Dictionary (JWD)*. Διαθέσιμο στο: https://joycewords.com/ [20/04/2021].

Conner, L. I. (1999). A Yeats Dictionary: Persons and Places in the Poetry of William Butler Yeats. New York: Syracuse University Press.

Crystal, D. & Crystal, B. (2002). *Shakespeare's Words: A Glossary and Language Companion*. UK: Penguin Books

Fuertes-Olivera, P. A. & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin: De Gruyter.

Gentilini, A. (1970). Γλωσσάριο των Ωδών. Στο M. F. Pontani (επιμ.), *Ανδρέα Κάλβου Ωδαί*. Αθήνα: Ίκαρος, σσ. 175-κε΄.

Guilbert, L. (1975). *La créativité lexicale*. Paris: Larousse.

Guilbert, L. (1973). Théorie du néologisme. In *Le Néologisme dans la langue et dans la literature*, vol. 25. Paris: Cahiers de l'Asspciation Internationale des études françaises.

Καζαντζάκης, Ν. (1964 κ.εξ.). *Οδύσσεια*. Αθήνα: Εκδόσεις Καζαντζάκη.

Kazantzakis, N. (1958) *The Odyssee. A modern sequel*, (tr. Kimon Friar). New York: Simon and Schuster 1958.

Καζαντζάκης, Ν. (1938). *ΟΔΥΣΕΙΑ*. Αθήνα: εκδ. Πυρσός.

Καραντζόλα, Ε., & Φλιάτουρας, Α. (2004). *Γλωσσική αλλαγή*. Αθήνα: Νήσος.

Καψωμένος, Ε. Γ. (2004). Γλωσσική στατιστική και ύφος στον Ερωτόκριτο. Στο *Πεπραγμένα Θ΄ Διεθνούς Κρητολογικού Συνεδρίου, Β1*. Ηράκλειο: Εταιρεία Κρητικών Ιστορικών Μελετών, σσ. 89-104.

Καψωμένος, Ε. Γ., Αντωνίου, Μ., Λαδογιάννη, Γ., Στρουγγάρη, Μ., & Τριάντου, Ι. (1983). *Λεξικό Σολωμού. Πίνακας λέξεων του ελληνόγλωσσου σολωμικού έργου*. Ιωάννινα: Επιστημονική Επετηρίδα της Φιλοσοφικής Σχολής (Πανεπιστήμιο Ιωαννίνων) – Δωδώνη.

Κοκόλης, Ξ. Α. (1976). *Πίνακας λέξεων των 154 ποιημάτων του Κ. Π. Καβάφη*. Αθήνα: Ερμής.

Κοκόλης, Ξ. Α. (1975). *«Λέξεις-άπαξ»: στοιχείο ύφους, θεωρητική εξέταση-καταγραφή στα «ποιήματα» του Γ. Σεφέρη*. Αθήνα: Εξάντας.

Κόλλια, Ε. (2011). Οι νεολογισμοί ως στοιχείο ύφους στο μυθιστόρημα *Μαριάμπας* του Γιάννη Σκαρίμπα. Στο *Πρακτικά 6ης Συνάντησης Εργασίας Μεταπτυχιακών Φοιτητών Τμήματος Φιλολογίας, 13-15 Μαΐου 2011*. Αθήνα: Εθνικό Καποδιστριακό Πανεπιστήμιο Αθηνών, 178-186.

[17] Βλ. Συνέντευξη του Ν. Μαθιουδάκη στον Δ. Δουλγερίδη στην εφημ. *Τα Νέα* (ημερογρονολογία ανάρτησης 19/09/2020)· περισσότερα στο https://www.tanea.gr/print/2020/09/19/lifearts/o-kazantzakis-os-laikos-glossologos/ [20/04/2021].

[18] Βλ. Συνέντευξη του Ν. Μαθιουδάκη στον Π. Φρούντζο στην εφημ. *Documento – Docville* (ημερογρονολογία ανάρτησης 01/12/2020)· περισσότερα στο https://www.documentonews.gr/article/nikos-mathioydakhs-oewrw-ton-kazantzakh-glwssopoio-kai-oxi-glwssoplasth/ [20/04/2021].

Κόλλια, Ε. (2007). Οι νεολογισμοί ως στοιχείο ύφους στο έργο *Το Σόλο του Φίγκαρω* του Γιάννη Σκαρίμπα. Στο *Πρακτικά Α΄ Πανελλήνιου Συνεδρίου για τον Γιάννη Σκαρίμπα*. Χαλκίδα: Διάμετρος, 75-86.

Κουμπής, Α. (1999). *Πίνακας λέξεων των ποιημάτων του Νίκου Εγγονόπουλου*. Ηράκλειο: Πανεπιστημιακές Εκδόσεις Κρήτης.

Κυριαζίδης, Ν. Ι., Καζάζης, Ι. Ν., & Bréhier, J. (1992). *Τα ελληνικά του Μακρυγιάννη με τον υπολογιστή*. Αθήνα: Εκδόσεις Παπαζήση.

Lehrer, A. (2005). Understanding trendy neologisms. In *Italian Journal of Linguistics-Revista di Linguistica*, 369-382.

Lorando, G., Marcheselli, L., & Gentilini, A. (1970). *Lessico di Cavafis*. Padova, Liviana: Università di Padova, Studi Bizantini e Neogreci.

Μαθιουδάκης, Ν. (2020). *Η Οδύσ[σ]εια των λέξεων. Νεολογικά Αθησαύριστα στο έπος του Νίκου Καζαντζάκη*. Αθήνα: Κάπα Εκδοτική.

Μαθιουδάκης, Ν. (2012). Νεολογικά Αθησαύριστα στην ΟΔΥΣΕΙΑ του Νίκου Καζαντζάκη. Στρατηγικές κατανόησης, Ασάφεια και Βεβαιότητα [Διδακτορική διατριβή]. Κομοτηνή: Δημοκρίτειο Πανεπιστήμιο Θράκης.

Μαθιουδάκης, Ν., & Καμπάκη-Βουγιουκλή, Π. (2011). Η επιθετική ταυτότητα του Οδυσσέα στο έπος του Νίκου Καζαντζάκη: μια πρόταση μέσω των ασαφών συνόλων. Στο Κ. Α. Δημάδης (επιμ.), *Πρακτικά του Δ΄ Ευρωπαϊκού Συνεδρίου Νεοελληνικών Σπουδών, Γρανάδα, 9-12 Σεπτεμβρίου 2010*, (τόμ. Α΄). Αθήνα: Ευρωπαϊκή Εταιρεία Νεοελληνικών Σπουδών (ΕΕΝΣ), σσ. 295-314.

Μαθιουδάκης, Ν., & Καρασίμος, Α. (2015). Διαλεκτικά στοιχεία στη διαδικασία σύνθεσης: Μελέτη στην Οδύσεια του Καζαντζάκη. Ανακοίνωση στην 36η Συνάντηση του Τομέα Γλωσσολογίας του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, 24-25 Απριλίου 2015.

Mathioudakis, N., & Karasimos, A. (2014). Dialectic and Idiomatic Aspects in Odyssey by Nikos Kazantzakis. In G. Kotzoglou, K. Nikolou, E. Karantzola, K. Frantzi, I. Galantomos, M. Georgalidou, V. Kourti-Kazoullis, Ch. Papadopoulou & E. Vlachou (ed.), *Selected Papers of the 11th International Conference on Greek Linguistics (11th ICGL)*. Rodes/Greece: University of Aegean, 1070-1088.

Μαρινάκης, Θ. (2004). Με πυξίδα ένα γλωσσάρι: Το Γλωσσάριο του Παντελή Πρεβελάκη για την Οδύσσεια του Νίκου Καζαντζάκη (Ραψωδίες Α-Κ) [Μεταπτυχιακή εργασία]. Ρέθυμνο: Πανεπιστήμιο Κρήτης – Τμήμα Φιλολογίας.

Μαυρομάτης, Δ. Κ. (1981). *Πίνακας λέξεων του «Άξιον Εστί» του Οδυσσέα Ελύτη*. Ιωάννινα: Πανεπιστήμιο Ιωαννίνων – Φιλοσοφική Σχολή – Έδρα Νεοελληνικής Φιλολογίας.

Μότσιου, Β. (1994). *Στοιχεία λεξικολογίας. Εισαγωγή στη νεοελληνική λεξικολογία*. Αθήνα: Νεφέλη.

Νάκας, Θ. (2003). *Γλωσσοφιλολογικά, Γ΄. Μελετήματα για τη γλώσσα και τη λογοτεχνία*. Αθήνα: Παρουσία – Εκδόσεις Πατάκη.

Nielsen, S., & Tarp, S. (eds.). (2009). *Terminology and Lexicography Research and Practice*. Amsterdam: John Benjamins Publishing Company.

Ξυδόπουλος, Γ. Ι. (2008). *Λεξικολογία. Εισαγωγή στην ανάλυση της λέξης και του λεξικού*. Αθήνα: Εκδόσεις Πατάκη.

Oleynikova, G. (2016). The Role of Author's Neologisms in Literary Text. In *Journal of Danubian Studies and Research*, vol. 6, no 2 (2016).

Onions, C. T. (1986). *A Shakespeare glossary*. Oxford: Clarendon.

Pelacchi, G. (1971). Lessico della poesia di Kariotakis [Dissertation]. Padova: Istituto di Studi Bizantini e Neogreci.

Peri, M. (1983). *Πίνακας λέξεων του Καρυωτάκη*. Padova, Liviana: Università di Padova, Studi Bizantini e Neogreci.

Picone, M. D. (1996). *Anglicisms, Neologisms and Dynamic French*. Amsterdam & Philadelphia: John Benjamins.

Πρεβελάκης, Π. (1984). *Τετρακόσια γράμματα του Καζαντζάκη στον Πρεβελάκη*. Αθήνα: Εκδόσεις Καζαντζάκη.

Πρεβελάκης, Π. (1958). *Ο Ποιητής και το Ποίημα της Οδύσσειας*. Αθήνα: Βιβλιοπωλείο της Εστίας.

Schmidt, A. (1902). *Shakespeare Lexicon and Quotation Dictionary: A Complete Dictionary of All the English Words, Phrases, and Constructions in the Works of the Poet (vol. 1 A-M & vol. 2 N-Z)*. Berlin: Georg Reimer.

Sterkenburg, P. van (ed.). (2003). *A Practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company.

Sutherland, J. (2012). *The Dickens Dictionary: An A-Z of England's Greatest Novelist*. UK: Icon Books Ltd.

Tarp, S. (2012). Online Dictionaries: Today and Tomorrow. In *Lexicographica*, 28 (1), 253-267.

Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purposes: Towards the Indi-vidualization of Needs Satisfaction. In P. A. Fuertes-Olivera & H. Bergenholtz (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, pp. 54-70.

Tarp, S., & Gouws, R. (2019). Lexicographical Contextualization and Personalization: A New Perspective. In *Lexikos*, vol. 29, 250-268.

Τράπαλης, Γ. (2010). *Γλωσσάρι στο έργο του Νίκου Καββαδία: με ερμηνευτικά σχόλια*. Αθήνα: εκδ. Άγρα.

Φιλιππίδου, Ν. (1986). *«Η θυσία του Αβραάμ» στον υπολογιστή. Λεξιλογικοί πίνακες και υφολογικά σχόλια*. Αθήνα: Ερμής.

Φιλιππίδου, Ν., & Holton, D. (1996). *Του κύκλου τα γυρίσματα. Ο Ερωτόκριτος σε ηλεκτρονική ανάλυση: Συμφραστικός πίνακας λέξεων: Concordance: Α-Ι, Κ-Ο & Π-Ω*. Αθήνα: Ερμής.

Χαραλαμπάκης, Χ. (2011). Οι νεολογισμοί του Κωστή Παλαμά. Στο *Κωστής Παλαμάς. Πρακτικά ημερίδας, Αθήνα 16 Δεκεμβρίου 2009*. Αθήνα: Ακαδημία Αθηνών, σσ. 57-71.

Χαραλαμπάκης, Χ. (2010). Το γλωσσικό ύφος του Νίκου Καζαντζάκη. Στο *Ημερίδα για τον Νίκο Καζαντζάκη, 27 Νοεμβρίου 2007*. Αθήνα: Ακαδημία Αθηνών, σσ. 131-140.

Χαραλαμπάκης, Χ. 2001. *Νεοελληνικός λόγος. Μελέτες για τη γλώσσα, τη λογοτεχνία και το ύφος*. Αθήνα: εκδ. Νεφέλη.

Χριστοφίδου, Α. (2001). *Ο ποιητικός νεολογισμός και οι λειτουργίες του. Κειμενική-γλωσσολογική προσέγγιση στο έργο του Οδυσσέα Ελύτη*. Αθήνα: εκδ. Gutenberg.

Wells, S. ([1998] 2013). *An A-Z Guide to Shakespeare*. UK: Oxford University Press.

8 Λεξικά

Δημητράκος, Δ. (1964). *Μέγα λεξικόν όλης της ελληνικής γλώσσης*, τόμ. Α-ΙΕ. Αθήναι: Εκδόσεις Δομή.

Κριαράς, Ε. (1995). *Νέο ελληνικό λεξικό της σύγχρονης δημοτικής γλώσσας. Ορθογραφικό, ερμηνευτικό, ετυμολογικό, συνωνύμων, αντιθέτων, κυρίων ονομάτων*. Αθήνα: Εκδοτική Αθηνών

Μπαμπινιώτης, Γ. ([2005] 2008). *Λεξικό της Νέας Ελληνικής Γλώσσας: με σχόλια για τη σωστή χρήση των λέξεων: ερμηνευτικό, ορθογραφικό, ετυμολογικό, συνωνύμων-αντιθέτων, κυρίων ονομάτων, επιστημονικών όρων, ακρωνυμίων*. Αθήνα: Κέντρο Λεξικολογίας.

Πρωίας. (1933). *Λεξικόν της Νέας Ελληνικής Γλώσσης: Ορθογραφικόν και Ερμηνευτικόν: Συνταχθέν υπό επιτροπής φιλολόγων και επιστημόνων. Έκδοσις Νεοτάτη*, τόμ. Α-Γ, (επιμ. Γ. Ζευγώλη). Αθήναι: Εκδοτικός Οίκος Σταμ. Π. Δημητράκου.

Σταματάκου, Ι. Δρ. (1971). *Λεξικόν της Νέας Ελληνικής Γλώσσης. Καθαρευούσης και δημοτικής και εκ της νέας ελληνικής εις την αρχαίαν*, τόμ. 1-3. Αθήνα: Βιβλιοπρομηθευτική.

[Τριανταφυλλίδη, Μ.] Ινστιτούτο Νεοελληνικών Σπουδών – Ίδρυμα Μανόλη Τριανταφυλλίδη. (1998). *Λεξικό της Κοινής Νεοελληνικής*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης. Online version: *www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/index.html* [20/04/2021].

[Χαραλαμπάκης. Χ.]. (2014). *Χρηστικό λεξικό της Νεοελληνικής Γλώσσας*. Αθήνα: Ακαδημία Αθηνών.

Βοσταντζόγλου Θ. ([1967] 1976). *Αναλυτικόν ορθογραφικόν λεξικόν της νεοελληνικής γλώσσης (καθαρευούσης και δημοτικής)*. Αθήνα: Δομή.

cography tain ions Speech ologisms Idioms NLP Etymology Glossary Lemma gin Corpora Dictionary Word Meaning Lexicon ble Pronounciation Definition Headword

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

amples Entry

www.euralex2020.gr

Lexicology Dictionary Use Lexical Resources

Software Demonstrations

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Software Demonstrations

Lexicography
for Specialised Languages, Terminology

Introducing Terminologue: a cloud-based, open-source terminology management tool

Měchura M. B.[1], **Ó Raghallaigh B.**[1]

[1] *Fiontar & Scoil na Gaeilge, Dublin City University, Ireland*
valselob@gmail.com, brian.oraghallaigh@dcu.ie

Abstract

This software demonstration introduces Terminologue www.terminologue.org, a cloud-based, open-source terminology management tool. When installed, Terminologue allows users to create, edit and publish termbases via the web. Self-registration is provided, and there are no limits to the number of termbases that can be created or to the number of entries in a termbase. The web-based interface allows registered users to manage their account, to configure their termbases, to modify a termbase's metadata fields, and to edit termbase data. Each entry represents a concept and entries are edited via a tabbed widget which allows users to focus on the different elements of the entry in turn. The overall interface is optimised for both desktop and mobile screens. Data import and export tools are provided, as well as termbase download and upload. Users can be assigned different access levels from read-only to full administrator level. An extranet interface allows lists of entries to be shared with external subject-area experts for review and comment.

Keywords: terminology; terminography; open source; software as a service

1 Introduction

Terminologue is a web-based platform for building, managing and publishing termbases. The software is open-source and can be obtained from GitHub.[1] Documentation is provided for users wishing to install and host their own instance of Terminologue. Alternatively, users who simply wish to use Terminologue can create an account on www.terminologue.org and proceed. This instance of the software is hosted in the cloud by Dublin City University (DCU) and provided as is free of charge. Terminologue is a child of the Léacslann platform (Měchura 2012) and a sibling of the Lexonomy dictionary writing and publishing system (Měchura 2017).

The terminologue software was developed by the Gaois Research Group in Fiontar & Scoil na Gaeilge, DCU, on behalf of Foras na Gaeilge, the state-funded body responsible for the promotion of the Irish language throughout the whole island of Ireland. It was developed to manage the National Terminology Database for Irish (NTD), which is administered by Foras na Gaeilge and contains Irish-language terms approved by the National Terminology Committee (Měchura & Ó Raghallaigh 2010). For this reason, Terminologue is not standards-based but rather was designed to suit the requirements of the NTD, and is based on previous work on this project dating back to its launch in 2005 (Měchura 2006). Nonetheless, it broadly corresponds to established practice, i.e., there are concepts, inside concepts there are terms, definitions, etc. And while Terminologue provides a generic public interface for publishing termbases, the NTD is published via a custom interface available at www.tearma.ie.

Terminologue's back end is written in Node.js, so it runs on both Linux and Windows. Node.js is a JavaScript runtime designed for building scalable network applications.[2] Termbases are stored as SQLite databases and terminological entries are stored internally as JavaScript Object Notation (JSON) data. Terminologue's front end is written in EJS and CSS. The front end uses the Screenful library for screen layout, access management, on-screen interactivity and search.[3] In addition to Irish and English, the Screenful and Terminologue user interfaces have been translated by users into Czech, Swedish, Welsh, Finnish, Dutch, Russian, Spanish and Arabic. The screen flows from right to left when the Arabic interface is selected.

Terminologue users can create a termbase by clicking on the link to create a termbase on the home page when signed in. Any number of termbases can be created and they can also be deleted. When creating a new termbase, the user gives it a name and a URL. If the user makes their termbase public, it will be available at this URL. Termbases can be created from scratch or from a number of templates. Templates are provided for simple monolingual, bilingual and multilingual termbases, and the settings preconfigured by the templates can be changed later.

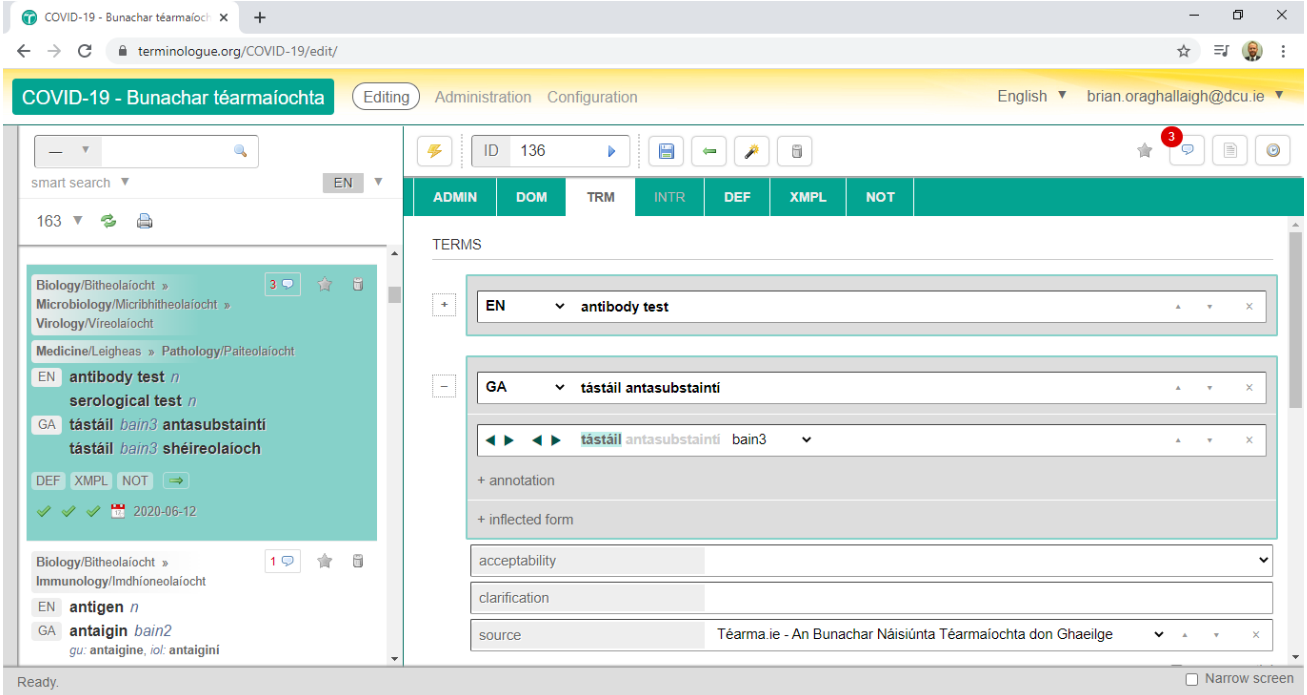
Termbases are stored on the server as individual SQLite files. These files can be downloaded via the web interface and users of www.terminologue.org are encouraged to do so to back up their work. Terminologue also allows users to import data from a TBX formatted file, and to export their data into a TBX file, albeit not in a lossless way. Individual entries, lists of entries, or entire termbases can be downloaded in TBX or TXT format. The cloud-based instance of the software currently hosts *c.*750 users and *c.*1,000 termbases. We know from users who have been in touch that it is in use in a number of universities as a teaching tool on translation and terminology courses.

[1] https://github.com/gaois/terminologue
[2] https://nodejs.org/en/about/
[3] https://github.com/michmech/screenful

2 Entry Structure

Once a termbase has been created, users can go to its homepage and from there to the editing interface. This is where terminological entries are created and edited. There is a list of entries on the left-hand side. The sample entries can optionally be deleted, and any number of new entries can be created. There is no upper limit on the number of entries a termbase can contain. To open an entry, a user must click on it and it will appear on the space on the right hand-side. What is seen here is a formatted rendering of the entry. This is what the entry will look like if and when it is decided to make the termbase public. To edit the entry, click the Edit button at the top, and Terminologue will open the entry for editing.

Regardless of whether or not a termbase is created from a template, Terminologue entries have the same basic structure. This structure comprises administrative data (e.g., whether checked or not), domains, designations (with each designation capable of comprising a term, an acceptability label, a clarification, and sources), intros (i.e., concept disambiguators), definitions, examples, notes, collections, extranets and cross references. Terms comprise an identifier, a language label, text, a list of annotations, and a list of inflections. Each entry represents a concept.


COVID-19 - Bunachar téarmaíoch
terminologue.org/COVID-19/edit/
COVID-19 - Bunachar téarmaíochta
Editing
Administration
Configuration
English
brian.oraghallaigh@dcu.ie
smart search
EN
163
ID
136
3
ADMIN
DOM
TRM
INTR
DEF
XMPL
NOT
TERMS
Ready.
Narrow screen

Biology/Bitheolaíocht » Microbiology/Micribhitheolaíocht » Virology/Víreolaíocht
Medicine/Leigheas » Pathology/Paiteolaíocht
EN antibody test *n*
serological test *n*
GA tástáil *bain3* antasubstaintí
tástáil *bain3* shéireolaíoch
DEF XMPL NOT
✓ ✓ ✓ 2020-06-12

Biology/Bitheolaíocht » Immunology/Imdhíoneolaíocht
EN antigen *n*
GA antaigin *bain2*
gu: antaigine, *iol:* antaiginí

EN antibody test
GA tástáil antasubstaintí
tástáil antasubstaintí bain3
+ annotation
+ inflected form
acceptability
clarification
source Téarma.ie - An Bunachar Náisiúnta Téarmaíochta don Ghaeilge

Figure 1: The Terminologue user interface showing the tabbed entry editing widget.

The tabs at the top allow these elements to be edited. The entry's terms can be edited under the TRM tab, its definitions under the DEF tab and so on. For example, to add a new term to the entry, a user must click the plus sign under the TRM tab and fill in the wording of the term. Once the term has been added more information can be added to it, for example a part-of-speech label, or one or more inflected forms. Once a user has finished editing an entry, they have to click the Save button at the top. Entries are stored as JSON data in the *entries* table of the termbase's SQLite database. A copy of the entry JSON is always saved to the *history* table prior to new changes being saved. Older versions of entries can be revived from the history. Each time an entry is saved, indexes stored in the database are automatically updated.

Editing an entry basically means typing text into boxes and selecting values from lists. Many of the lists of values, such as part-of-speech labels, domain labels and so on, constitute the termbase's metadata and can be configured individually for each termbase in the Administration section. If a termbase is started from a template, then it will already have some metadata preconfigured here. Termbase-level settings and functions can be found in the Configuration section. These include user access, name and blurb, languages, alphabetical order, automatic changes (where certain changes trigger automatic administrative labelling of entries), publishing, TBX export and import, SQLite download, URL change, delete and empty.

3 Significant Features

The previous section introduced Terminologue and has, we hope, shown that Terminologue has the features a terminologist typically requires for his or her work. In the next section we are going to zoom in on a few features in Terminologue which we think deserve special attention.

3.1 The Tabbed Editing Interface

When editing a concept, there is a tab for editing its terms, another tab for editing its definitions, and so on: all these items are 'entry-level' items, in terms of the TMF metamodel (Steurs et al. 2015). There is no notion of 'language level' and 'term level' in Terminologue, neither in the user interface nor in the internal data structure. Instead, concepts are composed of items of different types, such as terms and definitions, and these have metadata to indicate what language they are in.

3.2 Optional Sharing of Terms among Concepts

Each individual term can be linked to more than one concept, while any changes made to the term become immediately visible in all concepts it is linked to. This feature is useful because, in the NTD, terms are often polysemous (they refer to more than one concept) and richly grammatically annotated (with part-of-speech labels and lists of inflected forms). Being able to link a term to more than one concept saves work (as it is not necessary to duplicate the grammatical annotations) and prevents inconsistencies.

3.3 Inline Grammatical Annotation

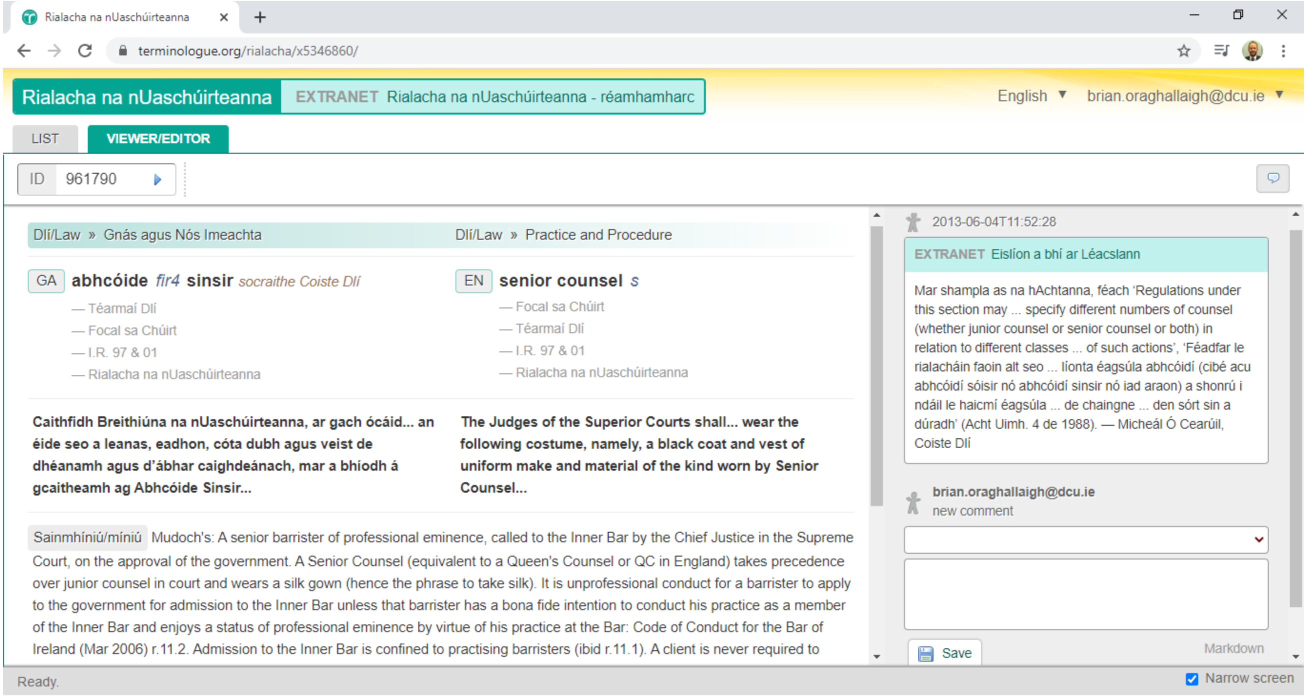
There is a tradition in Irish-language terminography of attaching grammatical labels to individual words inside multi-word terms, for example to the head noun of a noun phrase, as opposed to the entire noun phrase. Terminologue supports this by providing a widget for setting start and end character indexes for annotations. The annotation character indexes are lost when exporting to TBX, however.

3.4 Linguistically Smart Search Features

When making a termbase (or indeed any lexical resource) available to the public online, experience has shown (Měchura 2008) that members of the public often perform searches that are, in one way or another, defective. A high percentage of search requests are inflected forms of words, incorrectly spelled words, or incomplete terms. Linguistically sophisticated algorithms are required to match such search requests successfully to terms in the database. Terminologue has features for spelling error detection (using Levenshtein distance), for lemmatisation (using a large database of inflected forms of words in many different languages), and for partial matching.

3.5 Extranets

Organisations involved in terminological work often enlist the help of external subject-area experts. Cooperation with external experts usually happens on an extranet, which is basically an online environment where the experts can look at (unpublished) terminological entries and add their comments. Terminologue has such a feature, where lists of entries can be viewed and commented on by invited external users, facilitating an integrated workflow from initial drafting to final validation and publishing.


Rialacha na nUaschúirteanna
terminologue.org/rialacha/x5346860/
Rialacha na nUaschúirteanna
EXTRANET Rialacha na nUaschúirteanna - réamhamharc
English ▼
brian.oraghallaigh@dcu.ie ▼
LIST
VIEWER/EDITOR
ID 961790
Dlí/Law » Gnás agus Nós Imeachta
Dlí/Law » Practice and Procedure
GA
EN
2013-06-04T11:52:28
EXTRANET Eislíon a bhí ar Léacslann
brian.oraghallaigh@dcu.ie
new comment
Save
Markdown
Ready.
☑ Narrow screen

abhcóide *fir4* **sinsir** *socraithe Coiste Dlí*
— Téarmaí Dlí
— Focal sa Chúirt
— I.R. 97 & 01
— Rialacha na nUaschúirteanna

senior counsel *s*
— Focal sa Chúirt
— Téarmaí Dlí
— I.R. 97 & 01
— Rialacha na nUaschúirteanna

Caithfidh Breithiúna na nUaschúirteanna, ar gach ócáid... an éide seo a leanas, eadhon, cóta dubh agus veist de dhéanamh agus d'ábhar caighdeánach, mar a bhíodh á gcaitheamh ag Abhcóide Sinsir...

The Judges of the Superior Courts shall... wear the following costume, namely, a black coat and vest of uniform make and material of the kind worn by Senior Counsel...

Sainmhíniú/míniú Mudoch's: A senior barrister of professional eminence, called to the Inner Bar by the Chief Justice in the Supreme Court, on the approval of the government. A Senior Counsel (equivalent to a Queen's Counsel or QC in England) takes precedence over junior counsel in court and wears a silk gown (hence the phrase to take silk). It is unprofessional conduct for a barrister to apply to the government for admission to the Inner Bar unless that barrister has a bona fide intention to conduct his practice as a member of the Inner Bar and enjoys a status of professional eminence by virtue of his practice at the Bar: Code of Conduct for the Bar of Ireland (Mar 2006) r.11.2. Admission to the Inner Bar is confined to practising barristers (ibid r.11.1). A client is never required to

Mar shampla as na hAchtanna, féach 'Regulations under this section may ... specify different numbers of counsel (whether junior counsel or senior counsel or both) in relation to different classes ... of such actions', 'Féadfar le rialacháin faoin alt seo ... líonta éagsúla abhcóidí (cibé acu abhcóidí sóisir nó abhcóidí sinsir nó iad araon) a shonrú i ndáil le haicmí éagsúla ... de chaingne ... den sórt sin a dúradh' (Acht Uimh. 4 de 1988). — Micheál Ó Cearúil, Coiste Dlí

Figure 2: The Terminologue Extranet user interface showing the narrow screen layout.

4 Discussion

Terminologue is the result of a long evolution. The Gaois research group in Dublin City University has been developing terminology management tools since the 2000s, mostly to support the NTD. The NTD could be described as a *public-service terminology database:* it is government-funded and its purpose is to disseminate Irish-language terminology among the public. The NTD is also a very popular website, serving over a million searches every month (Měchura & Ó Raghallaigh 2010).

All this means that the NTD, and the software that supports it including Terminologue, has evolved somewhat in separation from the mainstream of terminology work worldwide, which – as it seems to us – is geared mainly towards the needs of corporations and translators. The NTD, on the other hand, targets the general public and its actual user base is wide and broad, encompassing more or less everybody who ever writes non-fiction texts in Irish.

In other words, Terminologue as a terminology management tool is a good fit for public-service terminology work (in any language) but may or may not be a good fit for corporate terminology work, depending on the exact requirements. Some of Terminologue's features exist specifically to support the requirements of public-service terminology work: an example is Terminologue's support for rich grammatical annotation of terms and for the sharing of terms between entries. On the other hand, some features typically found in other terminology software are absent in Terminologue, such as a clear separation between concept-level data, language-level data and term-level data, because a need for this never arises in (our experience of) public-service terminology work.

5 Conclusion and Future Work

In conclusion, we will outline some of the developments we have planned and some possible future directions. Among the developments, closest to production is work to improve the responsiveness and appearance of the interface on smaller screens. Currently users can set the interface to 'narrow screen' mode, but parts of the interface still do not respond well to a narrower screen. We are also looking into streamlining the interface translation process. In addition, users have requested that the public interface include an alphabetic list of all terms in a termbase, to complement the search and term cloud browse functionalities currently available.

With regard to entry structure, we are looking at a number of possibilities. For example, the possibility of including a 'term type' field has been suggested by a number of users and would be compatible with the TMF metamodel and TBX. We are also keen to introduce the option of labelling relationships between concepts. This would allow us to model and visualise conceptual ontologies. The ability to configure default values, to define mandatory fields, and to carry out automatic entry validation have all been mooted. Related to this, we are also looking at ways to better ensure data consistency, to simplify data indexing, and to optimise paging through large termbases.

Terminologue will continue to be actively developed over the next number of years, thanks to financial support from Foras na Gaeilge and institutional support from Dublin City University.

6 References

Měchura, M. B. (2006). Finding the Right Structure for Lexicographical Data: Experiences from a Terminology Project. In E. Corino, C. Marello, C. Onesti (eds.) *Proceedings of the 12th Euralex International Congress*, Torino, 6–9 September 2006. Torino: Edizioni dell'Orso, pp. 1:189–98.

Měchura, M. B. (2008). Giving them what they want: search strategies for electronic dictionaries. In E. Bernal, J. DeCesaris (eds.) *Proceedings of the 13th Euralex International Congress*, Barcelona, 15–19 July 2008. Barcelona: Universitat Pompeu Fabra, pp. 1295–1299.

Měchura, M. B. (2012). Léacslann: A Platform for Building Dictionary Writing Systems. In R. V. Fjeld, J. L. Torjusen (eds.) *Proceedings of the 15th Euralex International Congress*, Oslo, 7–11 August 2012. Oslo: University of Oslo, pp. 855–861.

Měchura, M. B. (2017). Introducing Lexonomy: An Open-Source Dictionary Writing and Publishing System. In *Proceedings of eLex 2017 Conference*, 19-21 September, Leiden, pp. 662–679.

Měchura, M. B. & Ó Raghallaigh, B. (2010). The Focal.ie National Terminology Database for Irish. In *Proceedings of the 14th Euralex International Congress*, 6–10 July 2010. Ljouwert/Leeuwarden, Netherlands.

Steurs, F., De Wachter, K. & De Malsche, E. (2015). Terminology tools. In F. Steurs, H. J. Kockaert (eds.) *Handbook of Terminology: Volume 1*. John Benjamins: Amsterdam, pp. 222–249.

Acknowledgements

This research was carried out with funding from Foras na Gaeilge.

Speech
Etymology
Idioms
Glossary
NLP
Lemma
Meaning
Dictionary
Word
Lexicon
Definition
Corpora
Pronounciation
Headword
amples
Entry
lexicology
Dictionary Use
Lexical Resources

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Index of Authors

Roman

EURALEX XIX

Congress of the
European Association
for Lexicography

Lexicography for inclusion

7-9 September 2021
Virtual

www.euralex2020.gr

Index

gmentation

Referenc

Form

Orthography

Ori

Definition

Sylla

Publishing

Graphemics

Pragmatics

Translate

Understand

Spelling

Context

Library

Ex

Spoken

Semantics

Order

ISBN 978-618-85138-2-2