EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr

Book of Abstracts

Edited by Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

The XIX EURALEX International Congress: Lexicography for inclusion

Edited by: Zoe Gavriilidou, Lydia Mitits, Spyros Kiosses

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

Published by: SynMorPhoSe Lab, Democritus University of Thrace

Komotini, Greece, 69100
e-edition

Publication is free of charge

Acknowledgements

We would like to thank all those who have made the XIX EURALEX XIX Virtual Congress 2021 possible, by contributing to the reviewing, to the logistics and by financially supporting the event.

SPONSORS

+Μόρ
Φως
Ση

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** (Chair) *(Democritus University of Thrace, Komotini)*

Maria **Mitsiaki** *(Democritus University of Thrace, Komotini)*

Tinatin **Margalitadze** *(Ivane Javakhishvili Tbilisi State University)*

Gilles-**Maurice de Schryver** *(Ghent University)*

Simon **Krek** *(University of Ljubljana, Center for Language Resources and Technologies)*

Annette **Klosa-Kueckelhaus** *(Leibniz-Institut für Deutsche Sprache)*

Tanara **Zingano Kuhn** *(Centre for General and Applied Linguistics Studies, University of Coimbra)*

George **Xydopoulos** *(University of Patras)*

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

BI- AND MULTILINGUAL LEXICOGRAPHY

HISTORICAL AND SCHOLARLY LEXICOGRAPHY AND ETYMOLOGY

LEXICOGRAPHY AND CORPUS LINGUISTICS

LEXICOGRAPHY AND LANGUAGE TECHNOLOGIES

LEXICOGRAPHY AND SEMANTIC THEORY

LEXICOGRAPHY FOR SPECIAL NEEDS

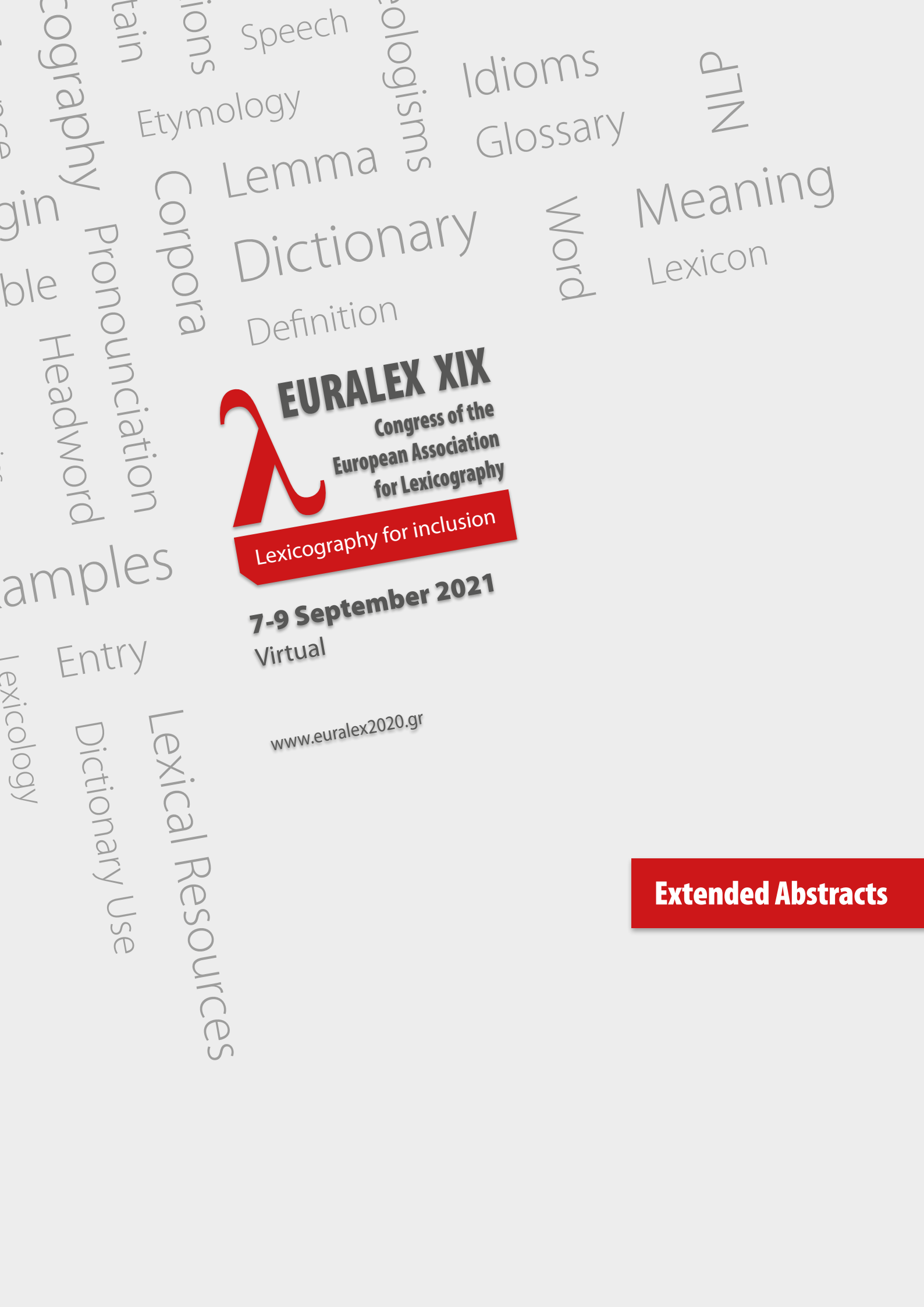
LEXICOGRAPHY FOR SPECIALISED LANGUAGES, TERMINOLOGY AND TERMINOGRAPHY

LEXICOLOGICAL ISSUES OF LEXICOGRAPHICAL RELEVANCE

PHRASEOLOGY AND COLLOCATION

REPORTS ON LEXICOGRAPHICAL AND LEXICOLOGICAL PROJECTS

THE DICTIONARY-MAKING PROCESS



Extended Abstracts

Extended Abstracts

Neology

Pescapalabras: Citizen science strategies applied to the detection and evaluation of neologisms

Nava Maroto[1], **Miguel Sánchez Ibáñez**[1]

[1] *Universidad Politécnica de Madrid, Spain*

Abstract

Crowdsourcing has long been established and tested as a valid methodology in lexicography (Čibej, Fišer and Kosem, 2015), and several possibilities of user collaboration have been soundly described, ranging from user-generated content to wikis and crowdsourcing (Rundell, 2016). Besides, the lexicographic filter as a method to detect neologisms has proved to be insufficient when used alone, since it is not capable of retrieving a significant amount of relevant new lexical units. It overlooks many important factors concerning the lexical updating of languages, such as the speakers' insights and preferences, the formal and semantic variation of the new units, or the influence news and current affairs have on the coinage of new words.

We firmly believe that lexicography should take advantage from the long-established citizen science experiences carried out in the field of the life sciences and others (Bonney et al, 2009) in order to redefine the way neologisms are retrieved and included in dictionaries. The methodological approach suggested by citizen science relies on lay people's collaboration in different research tasks, from mere data collection to more complex activities, such as lexicographic data evaluation and dictionary building.

Our proposal is an electronic platform called *Pescapalabras* (literally "word-fishing", in Spanish), where lay citizens are encouraged to contribute to with neologisms found in their everyday lives, as well as to rate proposals suggested by other users.

Our platform consists of different modules. First, a web page will serve as the basic interface of communication with volunteers. The webpage contains information about our past and current projects and dissemination activities carried out in libraries and education institutions in order to attract and train potential collaborators. From the web page, volunteers will be able to download a mobile application which, once installed on the user's device, can be used to capture new words (neologisms) identified in the press or other means of communication. Collaborators will suggest these neologisms, and afterwards linguistic experts will validate them and will assess their suitability to be included in a dictionary. This application will also contain a more ludic side, where games will be proposed as a means to suggest new names for new realities and to validate other users' suggestions. In this way, we aim at motivating our collaborators.

The validated neologisms suggested by citizens will be stored in a database where more linguistic data will be added by experts in linguistics and, if applicable, these proposals will be considered for their inclusion in a dictionary of neologisms, as well as analyzed in detail in order to detect trends and patterns in the lexical updating of the Spanish language. In order to decide whether a neologism is suitable for its inclusion in a dictionary, we will rely on the scoring system devised by Sánchez Ibáñez (2018), whose validity has also been contrasted with speakers' intuitions about neology (Sánchez Ibáñez and Maroto, in press).

In this workshop the advances in the platform design and testing will be presented, together with the protocols devised to control the whole process. The discussion of some issues raised by our proposal, such as the possibility of sharing the collected data as linguistic linked open data (LLOD), or the main characteristics that a neology dictionary should have in order to meet the users' needs, will also be discussed.

References

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V. and Shirk, J. 2009. "Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy." BioScience, 59(11), pp. 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Čibej, J., Fišer, D. and Kosem, I. 2015. The role of crowdsourcing in lexicography. In Proceedings of the eLex 2015 conference "Electronic lexicography in the 21st century: Linking lexical data in the digital age", Herstmonceux Castle, United Kingdom, August. 2015. (pp. 70–83).

Rundell, M. (2016). "Dictionaries and crowdsourcing, wikis and user-generated content." In *International Handbook of Modern Lexis and Lexicography*, ed. by Patrick Hanks and Gilles-Maurice de Schryver. Berlin: Springer-Verlag. 71.

Sánchez Ibáñez, M. 2018. "Definiendo 'en positivo' los neologismos formales: Hacia un análisis cuantitativo de la correlación entre sus características" *Pragmalingüística*, 26: 349-372.

Sánchez Ibáñez, M. and Maroto, N. In press. "Beyond timelines: challenges to combine theoretical premises and speakers' insights on the assessment, validation and inclusion of neologisms in dictionaries".

Identification of Neologisms in Japanese Corpora using Synthesis

James Breen[1], Timothy Baldwin[2], Francis Bond Nanyang[3]

[1] *Monash University Melbourne, Australia*
[2] *University of Melbourne Melbourne, Australia*
[3] *Technological University, Singapore*

Abstract

1. Introduction

This paper reports on part of a major study into the extraction of neologisms from Japanese corpora. In the study three main approaches were explored: a.analysis of morpheme sequences in Japanese texts to determine the presence of potential new or unrecorded terms. The processes included processing the texts with a morphological analyzer to produce sequences of tagged morphemes, tagging of the morphemes with features derived from combinatory data derived from large lexicons and corpora, and processing the tagged morphemes with rule-based and machine-learning-based chunkers to assemble candidate words and expressions. b analysis of language patterns which are often used in Japanese in association with new and emerging terms. These patterns are usually associated with discussions or explanations of new terms. (Breen et al., 2018) c.synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora. In this paper we report on the third of these, based on the synthesis of possible Japanese words. (Another component, covering compound verbs, has been reported separately. (Breen and Baldwin, 2009)) A central issue when dealing with neologisms in Japanese is the nature of the orthography, with its use of multiple scripts, primarily *kanji* (Chinese characters), e.g. 猫, 犬, 鳥, 牛, etc. of which approximately 2,500 are in common use and which are used mainly for nouns and the roots of verbs, adjectives, etc.; and the *hiragana* and *katakana* syllabaries, each of 46 symbols plus diacritics. A major issue is the absence of any indication of the boundaries between the syntactic elements in texts. Automated text-processing in Japanese usually relies on morphological analysis software such as *MeCab* (Kudo, 2008) which employ large morpheme lexicons such as *UniDic* (Den et al., 2007), however unrecorded terms will (by definition) usually not be found in these lexicons.

2. Resources

As the experimentation with synthesized terms takes the form of create-and-test, a key requirement is access to appropriate large-scale corpora to test for the presence and usage patterns of the terms.

The main accessible Japanese corpus for this type of testing is the Google *n*-gram Corpus (Kudo and Kazawa, 2007), based on approximately 20 billion text segments extracted from WWW pages, and is provided in the form of sets of 1-grams to 7-grams with counts of the numbers of occurrences. As the Google corpus only reports *n*-grams which occur 20 or more times a second *n*-gram corpus was assembled using the smaller Kyoto University WWW Corpus, containing about 500 million text segments.

3. Investigation Approach

Three types of synthesized term formation were investigated:

a. **Abbreviation/Clipping**. This is a very common and productive process in Japanese, wherein the (usually) leading character of each of the components of a composite are taken to form an abbreviated compound. An example of this is 学割 *gakuwari* "student discount" from the full compound 学生割引 *gakuseiwaribiki*.

b. **Affixation**. The addition of prefixes and suffixes, often written with a single *kanji*, is a very common morphological process in Japanese (Tsujimura, 2006). Vance (1991) describes 63 single-*kanji* affixes commonly employed. The process is very productive and the resulting terms are not usually lexicalized unless they have an idiomatic meaning or unusual reading.

c. **Compounding**. As in many languages, the formation of terms by combining two or more words or morphemes is very common. The components can be independent words, as in 秋 空 *akizora* "autumn sky" where both 秋 *aki* and 空 *sora* can be used independently, or bound morphemes as in 警告 *keikoku*

"warning" where neither component can be used alone. (See Tanaka (2002) and Baldwin and Tanaka (2004) for earlier work in this area.) Two types of synthesized compounds were investigated:
i. 2-*kanji* compounds, as in the examples above;
ii. composites formed by aggregating known 2-*kanji* compounds, for example combining 警告 (above) with 射撃 *shageki* "firing, shooting" forms a composite 警告射撃 *keikoku shageki* meaning "warning shot".
Synthesized terms which were not already recorded in a reference lexicon were considered if they occurred more tham 100 times in the corpus. The evaluation of such terms was carried out by examining their occurrences in a combination of syntactic contexts. Japanese uses a large number of particles which are typically written in the *hiragana* syllabary. Counts of occurrences were extracted for the terms encapsulated in 37 combinations of the following common pre/postpended particles.
pre: は (*wa*), が (*ga*), に (*ni*), の (*no*), な (*na*), て (*te*), や (*ya*)
post: を (*wo*), が (*ga*), に (*ni*), の (*no*), な (*na*), や (*ya*)
Two types of evaluation were carried out on the sets of counts:
a. a machine-learning analysis using support-vector machine (SVM) models trained the patterns of counts from a range of established terms (Chang and Lin, 2011);
b. a heuristic approach using rules based on the numbers of encapsulations.

4. Investigation Outcome

a. **Abbreviation/Clipping**. Of 33,000 synthesized abbreviations 7,900 were evaluated of which 162 were identified as potential new nouns (2.0%). Hand-checking a sample revealed that few were actually valid abbreviations. Most were other types of collocations.
b. **Affixation**. Initial testing of potential terms generated by this technique resulted in large numbers which were clearly in regular use. It quickly emerged, however, that they almost always had quite predictable meanings and were unlikely to be included in a dictionary. For example, the noun 衒学 *gengaku* "pedantry" can take suffixes such as the personalizing suffix 者 *sha* to form 衒学者 *gengakusha* "pedant" or the adjective-forming suffix 的 *teki* to form 衒 学的 *gengakuteki* "pedantic".
c. **Compounding**.
i. 2-*kanji* compounds. As there are over 6 million combinations of the most common 2,500 *kanji*, two samples each of 40,000 compounds were generated from two ranges of kanji from which were excluded kanji for numerics, common affixes, etc. The unlexicalized compounds were tested against the two corpora, resulting in 200-500 compounds being classified from each sample. Hand-checking selections of these revealed that most were valid terms, with about 60% being proper names. Examples of such terms include 移 弦 *igen* "string-crossing (violin, etc. technique)", 春苗 *shunbyō* "spring seedlings" (also a girl's name), and 母珠 *moshu* "large bead(s) in a Buddhist rosary". The precision of the technique, i.e. the proportion of classified terms which proved to be valid, was quite high.
ii. 4-*kanji* compounds. The potential numbers of 4-*kanji* compounds which can be generated from known 2-*kanji* compounds is very large, however few record sufficient counts in the *n*-gram corpora to be considered further. Several batches of 1 million compounds were generated, with 400-500 from each being accepted for further analysis, and 30-50 being flagged by the models as potential terms. Again hand-checking confirmed their validity, with terms such as 英国王室 *eikokuōshitsu* "British royal family", and 欧州遠征 *ōshūensei* "European campaign (esp. with sporting teams)" being identified.
A point to note is that many of the accepted terms have meanings which are readily apparent from the components, and are unlikley to be included in a general dictionary.
As mentioned earlier, the synthesized compounds were tested using both machine-learning and heuristic models. It was noted that in general the heuristic models performed more effectively than the machine-learning approach.

5. Conclusion

In this paper we describe the investigation of a neologism detection approach involving the synthesis of possible Japanese words by mimicking Japanese morphological processes, followed by testing for the presence of candidate words in Japanese corpora. Of the techniques tested: abbreviation, affixation and compounding, the latter showed particular promise, with the 2-*kanji* compound generation and classification resulting significant numbers of unrecorded terms.

References

Timothy Baldwin and Takaaki Tanaka. 2004. Translation by machine of compound nominals: Getting it right. In *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*.
James Breen and Timothy Baldwin. 2009. Corpus-based Extraction of Japanese Compound Verbs. In *Proceedings of the Australasian Language Technology Workshop (ALTW 2009)*.
James Breen, Timothy Baldwin, and Francis Bond. 2018. The Company They Keep: Extracting Japanese Neologisms Using Language Patterns. In *Proceedings of the Global Wordnet Conference*.
Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese). *Japanese Linguistics*, 22:101–123.
Taku Kudo. 2008. *MeCab: Yet Another Part-of-Speech and Morphological Analyzer*. http://mecab.sourceforge.net/.
Taku Kudo and Hideto Kazawa. 2007. Japanese Web N-gram Corpus version 1. http://www.ldc.upenn.edu/Catalog/docs/LDC2009T08/.
Takaaki Tanaka. 2002. Measuring the similarity between compound nouns in different languages using non-parallel corpora. In *Proceedings of Coling 2002*.
Natsuko Tsujimura. 2006. *An Introduction to Japanese Linguistics*. Blackwell, Oxford, UK, second edition.
Timothy J Vance. 1991. *Instant Vocabulary through Prefixes and Suffixes*. Kodansha International, Tokyo.

New verbs and dictionaries: a method for the automatic detection of verbal neology

Rogelio Nazar[1], Irene Renau[1], Ana Castro[1]

[1] *Pontificia Universidad Católica de Valparaíso*

Abstract

Introduction. One of the most productive ways for the creation of new words in a language is using the already existing morphological resources, such as roots and affixes, combining them in innovative ways (Cabré, 2015; Makri, 2012; Moon, 2008; Muhvic-Dimanovski, 2004; Rey, 2005; Schmidt, 2008). For example, the Spanish suffix *-ear* is commonly used to create new verbs, such as *googlear* ('to google'), *guasapear* ('to send a whatsapp message'), *rapear* ('to sing a rap'), etc. As in many other morphologically rich languages, to detect a verb form in Spanish is a complex task. In Spanish, there are regular and irregular verbs, and even in the case of the regular verbs there are around 50 different conjugated forms (RAE, 2009). In this proposal, we describe a method to detect Spanish verb forms, either regular or irregular, using a set of rules combined with corpus data. With this method, we can detect any verb in use in the corpus. In order to apply this system to the detection of neological verbs, we compare the list of detected verbs with the list of verbs in a dictionary.

Methodology. We implemented an algorithm for the automatic extraction of Spanish verbs from corpora based on a system of rules which exploit the typical morphology that verbs bear. We hand-coded a list of verb endings obtained from the *Nueva gramática de la lengua española* (RAE, 2009), using only those endings with a minimum length of three letters. We also included the list of enclitics (pronouns which can be attached to verbs) and prefixes. Having such a list, we extract all word forms having a minimum frequency (5 occurrences) in the corpus, which in this case is the EsTenTen (Kilgarriff & Renau, 2013). When there was a match between a word form and a verb ending, it was stored in a two-dimensional hash table for root and ending. If at the end of the process we found that the same root was combined with multiple endings (4, as a minimum), then it was considered a verb candidate, and as such it was later be submitted to further evaluations. The first was to find the infinitive form, which in Spanish can only end in *-ar*, *-er* or *-ir*. Thus we attached the newly found root to each of this endings and selected the most frequent combination. Once the infinitive form was found, we discarded all verbs that were already listed in the dictionary –we used the *Diccionario de la lengua española,* DLE (RAE, 2014) via EnclaveRAE, which allowed us to do complex searches. We also used the list of prefixes in this process. If a prefix was detected in an infinitive, and the verb was listed in the dictionary without the prefix, then the candidate was considered a neologism. The remaining candidates were then screened for the detection of typographic errors (e.g., the verb *avanzar* 'to move forward' can be misspelled as *abanzar*, which needs to be discarded). For this, we implemented a special case of a spell checking algorithm consisting of a combination of orthographic similarity measures and insertion/substitution rules. They both compared the verb candidates with those verbs in the dictionary. Similarity was calculated with a Jaccard index using two-letter sequences as features, and the rule-based system inserts and changes letters in the verbs in order to find a match with the dictionary. A further filter consisted of a distributional similarity method, based on the intuition that two words with high distributional similarity can be spelling variation of the same word. In this same process we also checked that the candidate showed the distributional behaviour of actual verbs, for which we identified the typical profile of co-occurrence of Spanish verbs. Thus, if a verb candidate survived all these filters and bore the distributional behaviour of a verb, then it was considered a neologism candidate.

Results. We obtained a list of 10,034 tokens with a root + a verb suffix, and with an infinitive in the corpus which discarded a possible homography with a noun or an adjective. Thus, this list was considered to be a list of verbs. Of this list, 5,685 (56,65%) were classified as already existing verbs, as they were found in DLE, while the others were considered candidates to new verbs and processed by a different algorithm. This algorithm filtered the units to discard typos. As a result, 774 forms (7,71%) remained as real verbs, and the rest (n = 4,911, 48,94%) where discarded as typos or orthographic mistakes. As this 774 forms were not found in DLE, they were considered neological verbs.

Table I shows some examples of the list of new verbs.

Type of neology mechanism (Cabré, 2015)	New verbs	English translation
Prefixation (adding a prefix to an existing verb)	*coorganizar, desactualizar, intercomunicar, precalcular, reasignar,* etc.	'to co-organize, to out-to-date, to inter-communicate, to pre-calculate, to reassign', etc.
Suffixation (adding a verb suffix to an existing noun)	*aperturar, chocolatear, cooperativizar, ficcionar, gelificar,* etc.	'to open, to add chocolate, to create an association, to turn into fiction, to turn into gel', etc.
Parasynthesis (adding a prefix and suffix to an existing noun)	*adinerar, desvirtualizar, enrutar,* etc.	'to make someone rich, to devirtualize, to put in a route', etc.
Loans (adding a Spanish verb suffix to a loan)	*customizar, draftear, loguear, resetear, spoilear,* etc.	'to customize, to make a draft, to log, to reset, to spoil', etc.

Table I. Examples of new verbs detected with the proposed method.

The list shows many new verbs that are used in everyday language, and that were not incorporated into the dictionary because they were not considered normative (such as *spoilear* 'to spoil', which has a Spanish equivalent, *estropear,* but *spoilear* is very common in Spanish today) or because they were not detected by the lexicographic team, considered infrequent or too specialised (such as *gelificar,* which is a perfectly normative verb, a term from cooking).

Conclusions. The proposed method is able to detect verbs of a morphologically rich language such as Spanish, and can differentiate the new verbs from the ones that a dictionary already has. It is a relatively simple method, but very useful for lexicographic teams. The proposed algorithm can be easily incorporated into a workflow for creating or updating a lexicographical database, so it can be used regularly to detect new verbs and contribute to a more dynamic and updated dictionary.

Keywords: corpus linguistics, Spanish, neology, verbs.

References

Cabré, T. (2015). Neology, a nex field in search of its scientific stabilization. *Caplletra*, (59), 125-136.
Kilgarriff, A., & Renau, I. (2013). esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95, 12-19.
Makri, J. (2012). Terminología y tipología de los procesos de la neología formal. Debate Terminológico. ISSN: 1813-1867, (08), 3-15.
Moon, R. (2008). Lexicography and linguistic creativity. *Lexikos*, 18(1).
Muhvić-Dimanovski, V. (2004). New Concepts and New Words–How Do Languages Cope With the Problem of Neology?. *Collegium Antropologicum,* 28(1), 139-146.
Real Academia Española. (2009). *Nueva gramática de la lengua española.* Madrid: Espasa Calpe.
Real Academia Española. (2014). Diccionario de la lengua española. 23.a ed. Madrid: Espasa.
Rey, A. (1976). Néologisme, un pseudo-concept?. *Cahiers de Lexicologie* 28(1), 3-17.
Schmid, H. J. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. *Anglia-Zeitschrift für englische Philologie,* 126(1), 1-36.

The presence of Brazilian Portuguese neologisms in dictionaries

Ieda Maria Alves[1], **Bruno Maroneze**[2]

[1] *Universidade de São Paulo, Brazil*
[2] *Universidade Federal da Grande Dourados, Brazil*

Abstract

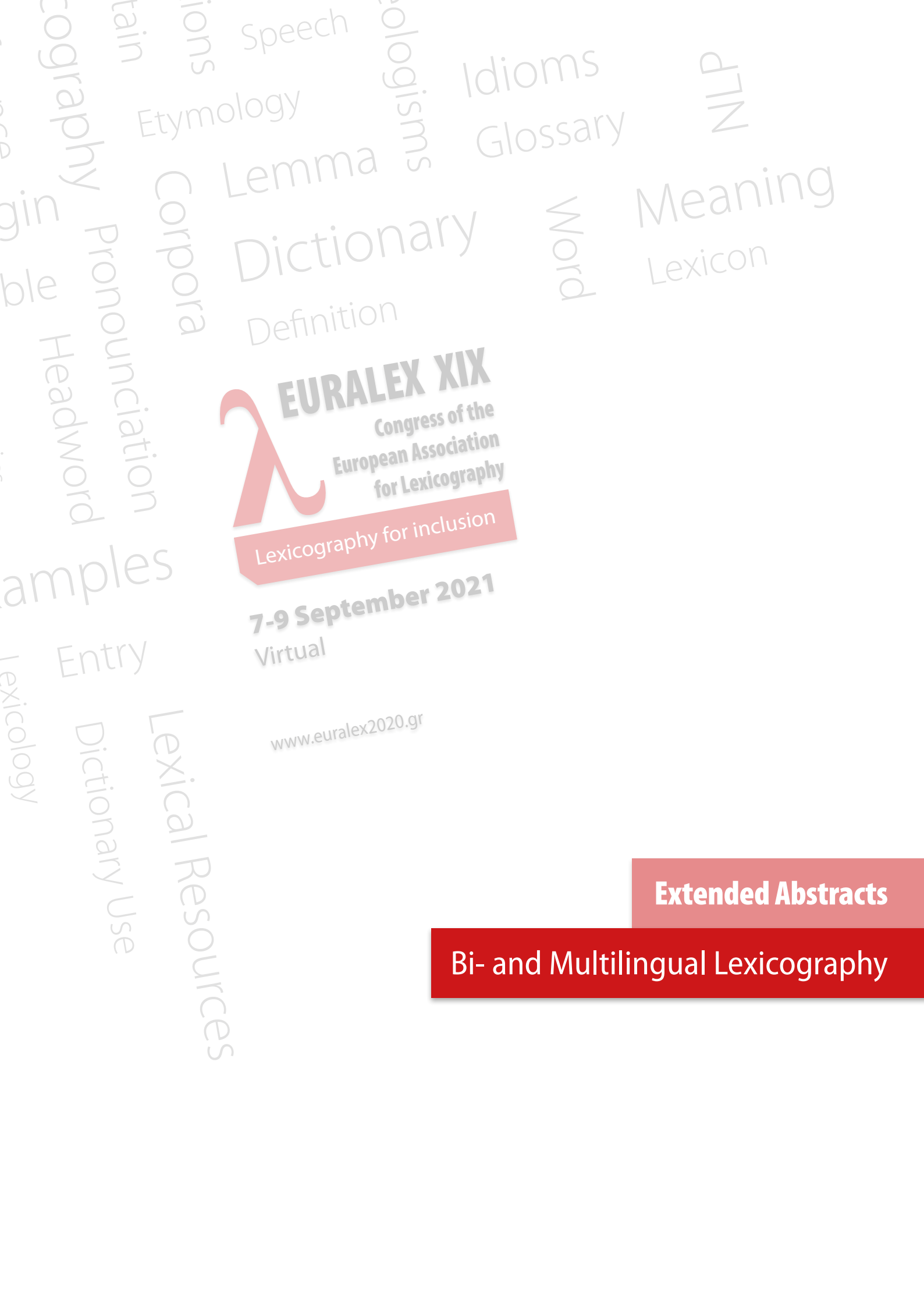
According to the data we have collected in the TermNeo Project (Observatory of Neologisms of Contemporary Brazilian Portuguese), collected in the highest-circulation Brazilian newspapers (Folha de S. Paulo, O Globo) and magazines (Veja, IstoÉ, Época), since 1993, vernacular neologisms are more productive in the development of the Portuguese language than neologisms that come from other languages (loanwords). Vernacular neologisms, formed by the processes of derivation, compounding and other processes (truncation, word blending and acronyms) correspond to 83% of the collected neological lexical units, with foreign words accounting for approximately 17% of neological formations. Among the vernacular neologisms, morphosyntactic ones, which include the derivation and compounding processes, are the most frequent, being distributed as follows: prefix derivatives (26%), suffix derivatives (10%), compounds (24%) and syntactic compounds (13%). The other collected neologisms represent the following processes: semantic neologisms (4%), and other processes such as formations with acronyms, truncations and blendings (6%).

These neologisms are differently included in dictionaries. In order to show that, we make a comparison between the online dictionaries Houaiss (considered the most complete dictionary of Brazilian Portuguese) and Aulete (an old dictionary that, in its recent online version, encourages the participation of users). The elements studied in this comparison are the morphological elements '-ódromo' and '-ômetro', which are old elements in the language, but have recently changed meaning.

The element '-ódromo' originally means a place for races (from 'hipódromo' – hippodrome, a place for horse races) but, recently, has begun to refer to a place where any action is performed. For example, 'camelódromo', a word that exists in Brazilian Portuguese since the 1980's, means a place where one can find 'camelôs' (street vendors). It is included in dictionary Aulete, but, although it is not a recently coined word, it is not included in dictionary Houaiss. With a similar meaning, 'sambódromo' is a place where one can watch samba parades during Carnival; this word also exists since the 1980's and, in this case, it is included in both dictionaries; only Houaiss includes an explanation of the structure of this word, and neither of the dictionaries registers the new meaning of the element '-ódromo'.

The element '-ômetro', with a meaning close to English '-meter', forms many names of measuring instruments (such as 'termômetro', thermometer), and is defined as such in dictionary Aulete (although not in Houaiss); but, more recently, it also forms popular words, such as 'achômetro' (from 'achar' – to have an opinion), which refers to an imaginary measure instrument that would make possible to someone to give an opinion; or 'chutômetro' (from 'chute' – a kick or, metaphorically, a guess), which is an imaginary instrument that would help people to take a guess. Both neologisms are in the dictionary Aulete, although the popular meaning of '-ômetro' is not described in the dictionary.

These examples show that Aulete, despite being a very traditional dictionary, includes many popular words, possibly because of the participation of users; but it does not explain their formation; instead, it presents only the traditional meanings of the morphological elements. Dictionary Houaiss, on the other hand, is more 'cautious' in the inclusion of neologisms, preferring not to register 'camelódromo', 'achômetro', 'chutômetro' and others, but presents the word structure of some of them, like 'sambódromo'. With this small study, we intend to show these and other differences between both dictionaries regarding the presentation of neologisms.



Extended Abstracts

Bi- and Multilingual Lexicography

The Perceived Impact of the *Oxford Bilingual School Dictionary: isiXhosa and English* on Students and Teachers in the Eastern Cape province, South Africa: an Evaluation Study

Megan Hall[1], Nontsikelelo Ntusikazi[2], Nomfundiso Mbali[3], Nomalungisa Ngondo[4], Phillip Louw[5]

[1] *Independent, Cape Town, South Africa*
[2] *Independent, Cape Town, South Africa*
[3] *Independent, Cape Town, South Africa*
[4] *Indigenous Languages Action Forum, Cape Town, South Africa*
[5] *Oxford University Press Southern Africa, Cape Town, South Africa*

Abstract

Most students at South African schools use a language other than their home language as their Language of Learning and Teaching. Primary school students' ability to read and understand in their Language of Learning and Teaching for grades 1 to 3 (ages 7 to 9) is known to be poor (Mullis, Martin, Foy and Hooper, 2017).

Results from later in the school system are similarly low, for example, in 2014, pass rates for the last grade 9 Annual National Assessment were 3% for Mathematics, 18% for First Additional Language, and 48% for Home Language, nationally (Department of Basic Education, 2014: 43–45).

Yet according to the UNESCO Institute for Statistics (2020), South Africa allocated 18% of government expenditure to education in 2016, a greater percentage than the United Kingdom (13.8%), Germany (10.9%), the Russian Federation (10.9%) or France (9.6%). Measured by their PIRLS 2016 (Mullis et al. 2017) and TIMSS 2015 Maths Grade 4 results (Mullis, Martin, Foy and Hooper, 2016), these countries have better educational outcomes than South Africa.

In this context, while acknowledging the very different histories and current situations of the countries mentioned, the general question arises: What learning and teaching support material is it worthwhile to invest in? More specifically, what perceived value can a bilingual dictionary yield?

In South Africa, bilingual dictionaries have been widely used to support the acquisition of some additional languages, e.g. Afrikaans-speakers learning English, English-speakers learning isiZulu (also called Zulu). However, this is not the case for all language-learning situations at schools. The development of dictionaries, whether monolingual or bilingual, for South African languages is uneven; where dictionaries do exist, they may not have been updated for some decades, they may not have been developed using modern lexicographic methods, and/or they may not be provided or available to school students.

The purpose of this research was to ascertain whether teachers felt that the Oxford Bilingual School Dictionary: isiXhosa and English (De Schryver and Reynolds, 2014) had an impact on themselves and their students, and, if it had an impact, whether teachers perceived the impact to be positive or negative.

A preliminary study was conducted in the Port Elizabeth educational district of the Eastern Cape (one of nine South African provinces), in March 2016. In 2018, the study was extended to include all educational districts in the Eastern Cape; this study was completed in February 2019.

Schools participating in the research (whether in the preliminary or extended stage) had ordered dictionaries through the Provincial Education Department in 2014; no copies were given gratis, nor was any associated dictionary training offered to participating schools. Although the Provincial Department subsequently reduced schools' original orders due to budget constraints, all participating schools had originally ordered a minimum of ten dictionaries and had received at least one copy at the end of 2014 or at the start of 2015. Participating schools had therefore had access to the dictionaries for more than

a year at the time of interview in the preliminary study, and, for the extended study, for four years (2015–early 2019).

All participating schools offered isiXhosa (also called Xhosa) as a Home Language and English as a First Additional Language. All participating teachers taught grade 7 students, grade 7 being the final year of primary school in South Africa. The Language of Learning and Teaching was English.

The permission of the Provincial Education Department for both stages of the research was sought and received, as was the permission of each school principal. The interview instrument was designed to elicit both positive and negative perceptions and was administered in isiXhosa, the home language of the teacher participants, by telephone. Almost all interviews were recorded. All the interviews were translated into English and transcribed.

The results from the preliminary and extended study may not be completely comparable, due to the length of access to the dictionary being different in the two cohorts. However, the methodology was the same and the cohorts were drawn from the same list of candidate schools, using the same criteria.

The scope of this study unfortunately did not allow the data about teachers' perceptions to be compared to test results from relevant schools. The primary reason for this is the Dept. of Basic Education not conducting national assessments at the grade 7 level as it does at the grade 12 level, resulting in testing being done by schools themselves alone. School-based test results are unavailable; they would also be unsuitable for comparison even within a province.

The research data was analysed from the lexicographic perspective of language reception vs language production.

The results of the preliminary and extended study, considered together, suggest that a large majority of participating teachers felt that the dictionary had positively impacted themselves and their students. According to participants, the positive impact on students encompassed improved language reception skills, as well as language production skills, across the home language and the first additional language. Participants also report feeling that using the dictionary supported a range of behaviours and attitudes that contribute to teaching and learning being more effective. Examples include increased self-confidence and willingness to use the additional language.

Finally, many teachers reported the dictionary's positive impact on content (or non-language) subjects.

The results relating to the perceived positive impact on the home language may be surprising since bilingual dictionaries are traditionally used to support an additional language rather than a home language, while monolingual dictionaries are used to support the development of the home language.

The perceived positive impacts on language skills in both Home Language and First Additional Language suggest that using this dictionary could support the additive bilingualism policy of the national Department of Basic Education.

Keywords: dictionary use, bilingual, user feedback, impact, user study, additive bilingualism, language reception, language production, language of learning and teaching

Twentieth-century English-Polish phraseological dictionaries: In search of a model

Mirosława Podhajecka

Institute of Linguistics, University of Opole, Poland

Abstract

In the light of European dictionary-making traditions, the history of English-Polish/ Polish-English lexicography is not extensive insofar as its modest beginnings may be traced back only to 1788, and the first major dictionary was published in the mid-nineteenth century. Neither has it been particularly rich. Although a sizable number of general-purpose bilingual dictionaries appeared during this period, most of them were pocket-size reference works produced by authors who were 'enlightened amateurs' rather than genuine specialists. The history of English-Polish phraseological dictionaries, which has attracted little attention thus far (see Podhajecka 2016), is even humbler. The first booklet that may be considered a phraseological dictionary, Adam Richter's *Polish Dictionary of English Idioms, Proverbs and Slangs* [sic], came out in Tel-Aviv in 1945 and never actually found its way into the hands of Polish learners of English, at least in Poland.

Until the early 1950s, only one high-quality phraseological dictionary, Mieczysław Kobylański's *Wybór idiomów angielskich* (1951), was launched onto the Polish market. This was by no means an impressive development, but, tellingly, the volume was compiled by an expert. A graduate in English from the Jagiellonian University, Kobylański knew several foreign languages, had both interest in and practical experience of foreign-language teaching, and was well aware of the complexity and fuzzy boundaries of English phraseology. Regrettably, for ideological reasons, the dictionary never went into another edition.

The situation changed significantly with the publication of two phraseological dictionaries: Piotr Borkowski's *An English-Polish Dictionary of Idioms and Phrases* (1963) and Roman Gajda's *Wybór idiomów angielskich* (1970). Although the former was originally issued in London, it came to be widely republished both in the West and in Poland. The latter, too, enjoyed unflagging popularity, which compelled Wiedza Powszechna, a state-owned publisher, continuously to issue subsequent editions well into the 1990s. The market demand for English handbooks and English-Polish/Polish-English dictionaries stemmed from the growing significance of English as a foreign language in Polish schools and universities.

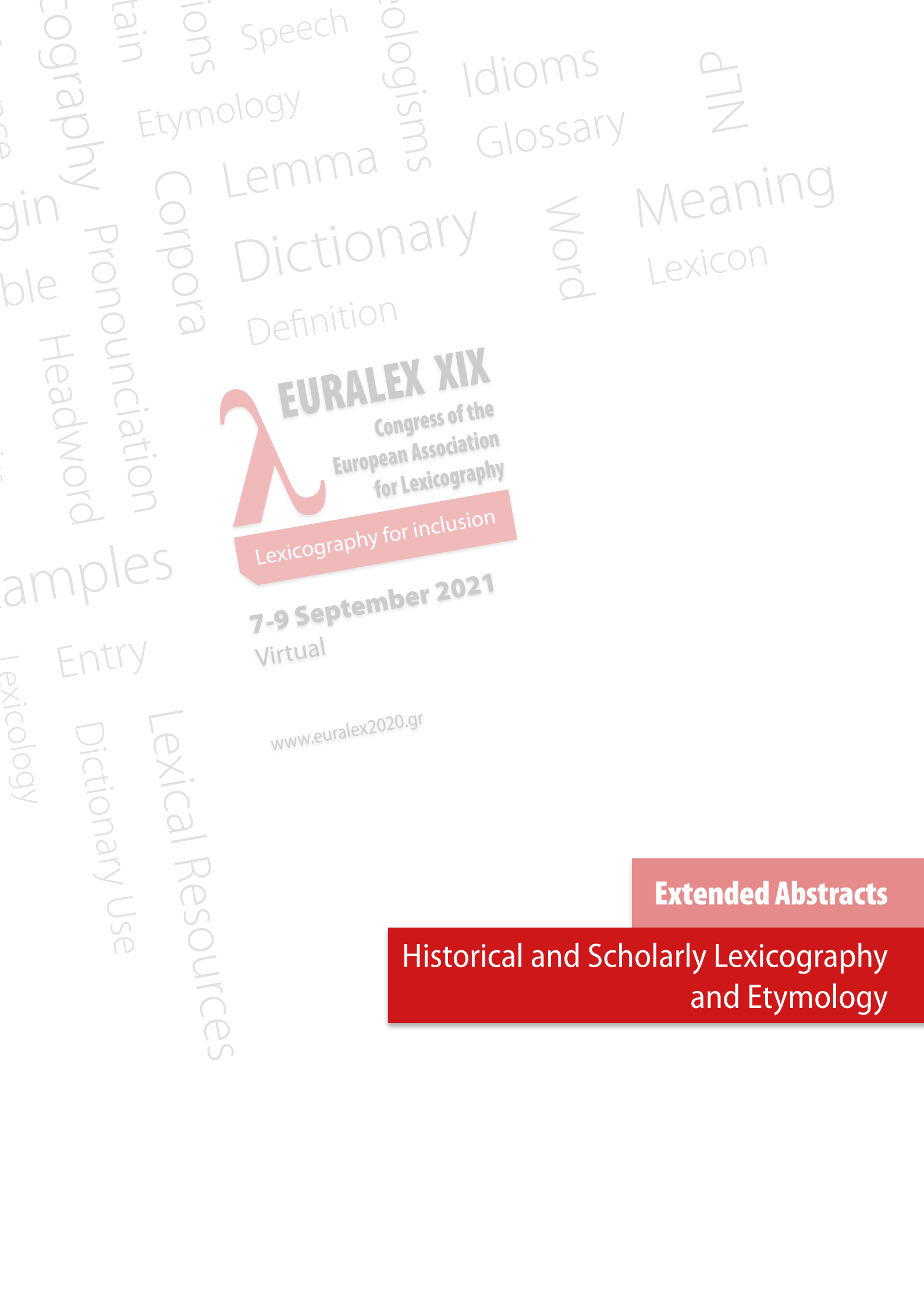
The aim of this paper is to analyze these two competing dictionaries from qualitative and quantitative perspectives. They seem to have shared several parallels. Firstly, both were of a similar size and length; secondly, both filled, at more or less the same time, an important market niche; thirdly, both were compiled by journalists, i.e., mere language enthusiasts; and lastly, both must have drawn on *The Kosciuszko Foundation Dictionary* (1959–1961) by Kazimierz Bulas, Francis Whitfield, and Lawrence Thomas, the most exhaustive dictionary of English and Polish available at that time (cf. Adamska-Sałaciak 2005, 2016). Upon closer scrutiny, however, one discovers marked differences not only in the contents of the dictionaries, but also in their authors' ideas of phraseology.

The analysis begins with the front matter. Borkowski is articulate regarding the motivations which prompted him to undertake the task. His agenda is complex. He tackles, on eleven pages of the Polish preface and three pages of the English foreword, phraseologisms typical of British and American English, the Bible and national literatures as sources of idiomatic expressions, differences between cultural equivalents, classifications of word combinations, and stylistic requirements. He pays special attention to explaining his arrangement of idioms, even though, as preliminary research indicates, he failed to treat them with absolute consistency. Gajda's three-page introduction is better suited for language learners. He describes briefly what word combinations he regards as idioms and how they are arranged in the dictionary, openly acknowledging the sources used in his compilation. Still, his discussion of *mieć węża w kieszeni* and *his money comes from him like drops of blood*, a Polish idiom and an English proverb respectively, seems somewhat out of place.

There are differences between the two dictionaries in terms of their macrostructures. While Borkowski resorts to idioms (in capital letters) in alphabetical order according to the main content word, which may in itself be a challenge to identify, Gajda uses headwords (in bold capital letters) and headphrases (in bold). Borkowski's wordlist includes approximately 4,000 word combinations, some of which are actually single words, and Gajda's has 257 headwords and over 1,000 headphrases.
The microstructures are even more divergent. Borkowski's entry covers exclusively the English idiom and its Polish equivalent. The use of capital letters for the headword, contrasted with unmarked typeface for the equivalent, is the only typographical device. The typical entry structure in Gajda's dictionary includes the headword, i.e. the main content word, and one or more headphrases. Each idiom is paired with one or more Polish counterparts and is followed by an English illustrative example, its Polish translation, and one or more translation or gap-filling exercises. As well as unmarked type, Gajda employs bold and italics.
The back matter in Borkowski's dictionary is composed of a single blank page for the user's handwritten notes and a list of abbreviations, including *hist* 'historical', *pol* 'political', and *sl* 'slang'. Gajda's back matter is more extensive, offering a key to exercises and an index of English headwords. We know that the financial restrictions in production of Borkowski's dictionary made it impossible for him to incorporate an index, but this flaw is compensated for by the Polish editions, even though they include a list of content words rather than idioms.
How can we evaluate the quality of both dictionaries? Their structural patterns are indicative of a search for an ideal model to present idiomatic expressions in dictionary form. Borkowski's strategy consisted in collecting as many word combinations as possible in the hope of providing the target user with a potentially comprehensive set, however precarious it may have turned out to be (cf. Murray 1996: 139). Gajda, by contrast, focused on the learning process entirely disregarded by his competitor. This explains why he provided Polish translations both for the idioms and the examples of usage, as well as exercises which aimed at consolidating the user's knowledge. What remains odd in this otherwise laudatory overview is Gajda's apparent, but unacknowledged, debt to Mieczysław Kobylański.

Bibliography

Adamska-Sałaciak, Arleta. 2005. "A lexicographical remake: The story of the *Kosciuszko Foundation (English-Polish) dictionary*." In Henrik Gottlieb, Jens Erik Mogensen, and Arne Zettersten (eds.), *Symposium on lexicography XI. Proceedings of the Eleventh International Symposiums on Lexicography May 2–4*, *2002*, *at the University of Copenhagen*. Tübingen: Max Niemeyer. 85–100.
Adamska-Sałaciak, Arleta. 2016. "Continuity and change in *The (New) Kosciuszko Foundation Dictionary*." *Studia Anglica Posnaniensia* 51(1): 83–98.
Borkowski, Piotr. 1963. *Angielsko-polski słownik idiomów i zwrotów / An English-Polish dictionary of idioms and phrases*. London: P. Borkowski.
Bulas, *The Kosciuszko Foundation Dictionary* (1959–1961) by Kazimierz Bulas, Francis Whitfield, and Lawrence Thomas
Gajda, Roman. 1970. *Wybór idiomów angielskich* [A selection of English idioms]. Warszawa: Wiedza Powszechna.
Kobylański, Mieczysław. 1951. *Wybór idiomów angielskich* [A selection of English idioms]. Warszawa: Państwowy Zakład Wydawnictw Szkolnych.
Murray, Laura. 1996. "Translation and transaction in American immigrant phrasebooks". *Canadian Review of American Studies* 26(2): 139–161.
Podhajecka, Mirosława. 2016. *A history of Polish-English/English-Polish bilingual dictionaries (1788-1947)*. Opole: Wydawnictwo Uniwersytetu Opolskiego.
Richter, Adam. 1945. *Polish dictionary of English idioms*, *proverbs and slangs* / *Słownik zwrotów angielskich.* Tel-Aviv: Wydawnictwo Książek Polskich "Świt".



Extended Abstracts

Historical and Scholarly Lexicography and Etymology

Making an ongoing historical dictionary accessible to an English-speaking audience

Tarrin Wills[1], **Simonetta Battista**[1], **Ellert Þór Johansson**[1]

[1] *Dictionary of Old Norse Prose, University of Copenhagen, Denmark*

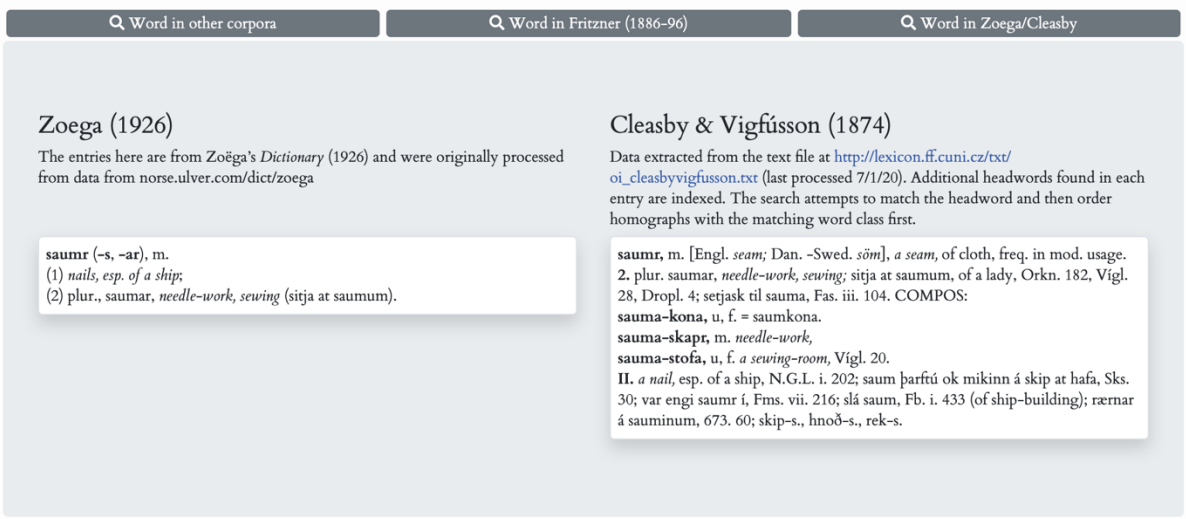
Abstract

The Dictionary of Old Norse Prose (ONP) is a long-term lexicographic project to semantically analyse an important corpus of medieval writings. The traditional audience for such projects consists of researchers and students of Old Norse language, who often have some knowledge of a modern Scandinavian language. The texts covered by the dictionary include many of general interest, such as the sagas of Icelanders, and texts of interest to fields such as history of religion, archaeology and comparative literature. This paper addresses the problem of how to make such a dictionary inclusive of a much larger audience, particularly those who do not have a modern Scandinavian language but wish to access the rich resources that ONP provides. We also assess the accuracy of one of the features implemented: user-requested machine translations using Google Translate.

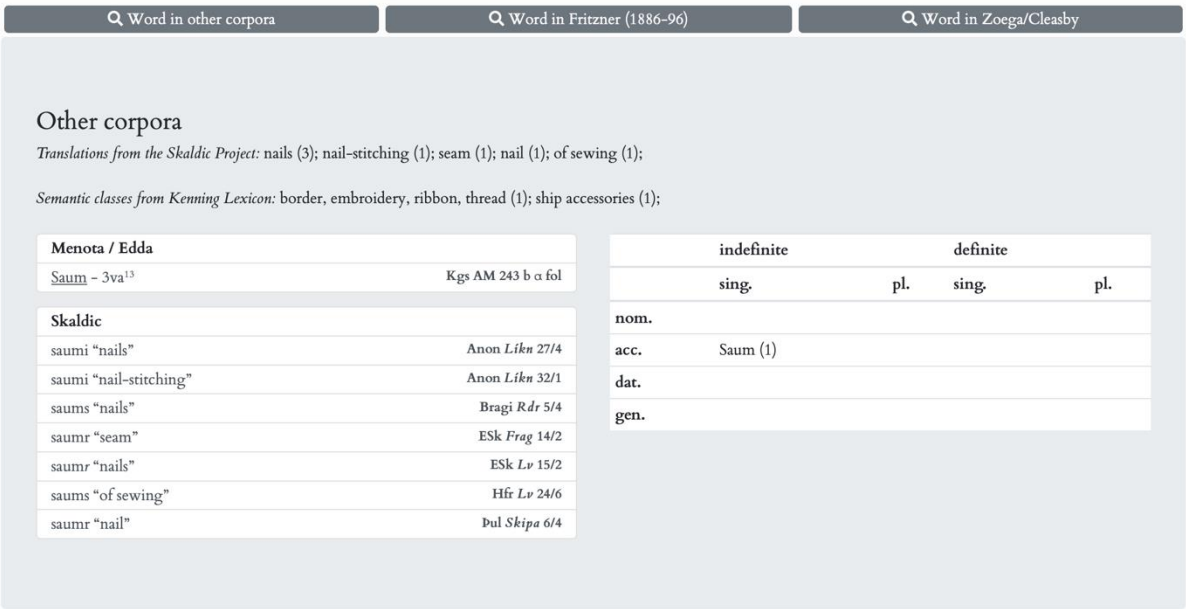
ONP is working on producing entries with both English and Danish as the target languages, but since the dictionary went digital-only (2004) the focus has been on Danish. Currently around 20% of the lexicon has definitions in English and around 50% in only Danish. The remainder is unedited at this stage. Specialist users who can read Scandinavian languages (such as Norwegian and Swedish) tend to be comfortable reading Danish, meaning that they can potentially understand a large proportion of ONP's definitions. However, according to analytics data, less than 20% of ONP's digital user base work in these languages. This suggests that a large proportion of the definitions in the dictionary are potentially not available to the overwhelming majority of users, thus excluding a large proportion of ONP's current and potential users.

The new web application (onp.ku.dk) developed for the project includes features designed to make the dictionary more inclusive for its users who do not have sufficient knowledge of Danish to use the large proportion of the dictionary with only Danish definitions. The techniques described here use external resources and fall broadly into two categories: supplementation (via querying of external resources such as dictionaries) and automatic translation (via Google Translate).

ONP's database includes detailed information about the occurrence of words in other dictionaries and glossaries. This information can be used to link words in ONP accurately to out-of-copyright digitised dictionaries, supplementing ONP where it is not edited, or where definitions are lacking in English. The most important of these dictionaries is Fritzner's Old Norse dictionary (1886-96), which has been processed so that in almost all cases a user can access the precise homograph for any given word in ONP. The definitions are in Norwegian, but we discuss later how these can be processed further. There also exist reasonable-quality digitisations of dictionaries by Cleasby and Vigfússon (1874) and Zoega (1926). These dictionaries have English as the target language and have been imported into a database that can be linked to ONP, giving users an English interpretation of almost all words in ONP, albeit a nineteenth-century interpretation.

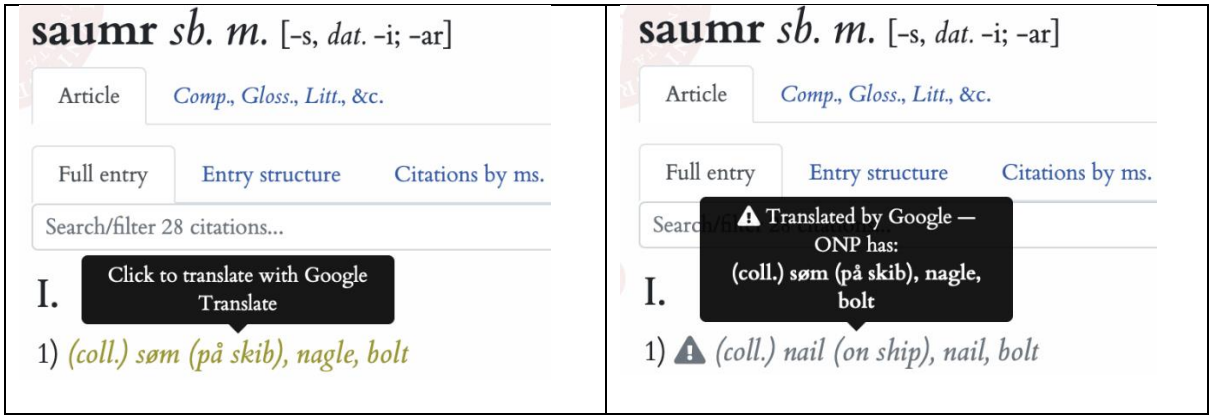


ONP also provides links to concordances in other corpora (Menota and the Skaldic Project) for each homograph, that is, linking to individual distinct entries rather than lemma strings. The Skaldic Project includes translations of every word in its context and a link to ONP's wordlist: at this stage there are 126,000 translated words representing 10,000 headwords in ONP. It therefore can provide contextual translations for headwords in ONP.



These external resources provide information for almost all words in ONP, including the large proportion that have not yet been edited, meaning that users can get information about the semantics and grammar of the lexicon while ONP continues its work.

The second general technique used by the web application is to integrate automatic translation via the Google Translate API. Where a definition exists only in Danish, the text of the definition in the web application becomes interactive so that when a user clicks or taps on the text, the application will substitute the translation via the API (the original text is saved in a 'tooltip'). The same applies to the Norwegian text in Fritzner's dictionary.



The technique here of machine translating specific definitions on demand means that user-requested translations can be logged and assessed for accuracy. There is a relatively small body of scholarly literature assessing the accuracy of Google Translate as a translation engine. For short phrases and terminology, and with Western European languages as the target, the level of accuracy is at least 74% as reported already in 2014 (Patil et al. 2014). The service's performance is continually improving.

The automatic translations of definitions requested by external users from Danish (ONP) and Norwegian (Fritzner) were logged in the period from 15-28 January 2020 (this data will be updated). These included 268 distinct pieces of Danish text (avg. 34 characters) and 90 distinct pieces of Norwegian text (avg. 40 characters) requested by approximately 60 external users. We have manually categorised on a scale as: 1. Accurate (including repetitive translations of synonyms), 2. Accurate but unidiomatic, 3. Partially inaccurate (but with inaccurate material easily identifiable in the context), and 4. Inaccurate (misleading or incorrect). Accuracy is assessed in the context of the original headword in Old Norse.

Preliminary data indicate around 90% of the machine translations fall into the first three categories, that is to say, in spite of repetitions and some errors, a user can interpret the generated English definition in the context of the web entry to get a sense of the semantics of the particular usage and without a great risk of them misunderstanding the particular sense of the word.

The service is therefore accurate enough to be very useful for users who do not have a grasp of Danish, but not nearly accurate enough to be used on its own for research purposes. These features provide information for a non-Danish-speaking audience in the majority of instances where no English language information is otherwise available, with only a small minority of instances creating the potential for error.

Keywords: Historical Lexicography, Machine Translation, Old Norse

Velum, un corpus textuel au service d'un dictionnaire : les problèmes d'une langue rare

Renaud Alexandre[1], **Bruno Bon**[1], **Krzysztof Nowak**[2]

[1] *Dictionary of Old Norse Prose, University of Copenhagen, Denmark*

Abstract

Il y a un siècle exactement, en 1920, le Dictionnaire européen du latin médiéval est lancé. Il doit fournir à la communauté internationale le premier instrument de travail scientifique, rendant compte de la diversité des usages de la langue latine médiévale, sur l'intégralité du continent européen, entre 800 et 1200. Avant de commencer la rédaction proprement dite du dictionnaire, il a naturellement fallu répertorier les sources et dépouiller les textes. Le latin médiéval présente des caractéristiques contradictoires, puisqu'il est à la fois clos et ouvert. Clos car plus aucun texte dans cette langue ne peut être écrit aujourd'hui ; ouvert car de nouveaux textes sont constamment découverts. On peut estimer à 40 % la proportion de textes médiévaux édités, ce qui rend nécessaire une mise à jour constante du répertoire correspondant : d'abord publié en 1957 avec le premier fascicule du dictionnaire, il a été republié dans des versions augmentées en 1973 (sous le titre d'*Index scriptorum novus mediae latinitatis*) et 2005, et il est désormais constamment mis à jour depuis sous une forme numérique (https://www.glossaria.eu/scriptores). La masse considérable de documents en jeu (10 000 références) interdisait tout dépouillement exhaustif : il est impossible de relever toutes les occurrences, et il serait impossible de les traiter. Le fichier du *Novum Glossarium* compte environ 2 millions de fiches, résultat des dépouillements réalisés depuis un siècle. Malheureusement, le dépouillement sélectif, par nécessité, a présenté de nombreux inconvénients, notamment des biais inévitables dans la sélection des occurrences. En résulte une surreprésentation des termes ou des sens perçus comme étranges ou inhabituels, et une sous-représentation des mots les plus fréquents ; c'est une des raisons qui ont conduit le *Thesaurus linguae latinae*, confronté au même problème, à repousser la publication de la lettre N, riche en conjonctions. Si ces biais ne sont pas corrigés, le dictionnaire produit par construction une image déformée du vocable étudié et du monde dont il est issu. De manière contre-intuitive, l'arrivée des premières bases de données textuelles en latin médiéval à la fin des années 1990 n'a pas amélioré la situation. La considérable augmentation des données disponibles n'a fait qu'accentuer les problèmes existants ou en créer de nouveaux. La surreprésentation des textes littéraires par rapport aux textes diplomatiques (dont le rapport est de 90%-10% dans les bases de données, et de 50%-50% dans les établissements de conservation) a ajouté un nouveau biais ; la faiblesse des modalités d'interrogation – encore valable aujourd'hui – ne permet pas une exploitation *a posteriori* des résultats d'une requête. Néanmoins, une équipe de deux personnes chargée de rédiger un instrument aussi ambitieux ne pouvait faire l'impasse sur les possibilités qu'ouvrait le développement du traitement automatique des langues, même si le latin médiéval relève des langues rares et sans débouché commercial.

Dans la mesure où le dictionnaire s'appuie sur un corpus de textes dont une partie est déjà disponible de manière dispersée, nous avons décidé de construire un corpus du latin médiéval comptant 100 millions de mots, assorti de trois corpus de cinq millions de mots chacun, pour effectuer des comparaisons (Antiquité, époque patristique et époque scolastique) ; c'est l'objectif du projet Velum (Visualisation, exploration et liaison de ressources innovantes pour le latin médiéval). Directement conçu pour permettre la rédaction du *Novum Glossarium Mediae Latinitatis*, ce corpus est caractérisé par une annotation qui doit permettre de repérer et simplifier la sélection des occurrences : métadonnées de genre (pour les domaines linguistiques), de lieu (pour la variation géographique), de date (pour l'amplitude chronologique des attestations), de *part-of-speech* (pour les remarques syntaxiques) et de lemme (pour les caractères morphologiques) ; l'ensemble de ces annotations est absent des collections de textes disponibles. Enfin, la nécessité de disposer d'un corpus représentatif, pour que le dictionnaire qui en est issu soit également représentatif, nous amène à créer un ensemble tout à fait différent des collections déjà existantes.

La première année du projet Velum a été consacrée à l'analyse de notre répertoire, l'*Index scriptorum*, pour tenter de dégager les critères de représentativité (répartition de nos sources dans le temps et l'espace, et distribution dans les différents genres). La moitié des sources repérées était déjà disponible dans des formats numériques (TXT et XML), en particulier les textes issus du projet openMGH (sélection de volumes des *Monumenta Germaniae Historica*, disponibles au format XML-TEI sous licence CC BY 4.0) : pour chaque sigle du dictionnaire dont l'œuvre a été numérisée dans openMGH, nous avons créé un fichier XML dans lequel nous avons inséré les métadonnées pertinentes inscrites dans l'*Index scriptorum*. Ces fichiers ont rejoint une base de test sous NoSketch Engine, qui nous permet d'ores et déjà d'enrichir considérablement le prochain fascicule du dictionnaire.
La poursuite de la constitution du corpus est nettement plus délicate, puisqu'il s'agit de traiter des textes non disponibles au format numérique. Même si nous avons pu limiter considérablement le travail de numérisation proprement dit grâce aux nombreuses banques d'images numériques disponibles (Gallica, archive.org, etc.), le traitement de fichiers PDF nécessite énormément de travail pour en extraire des textes exploitables. Nous en avons automatisé une partie, mais nous nous sommes heurtés, à chaque étape, à des difficultés nouvelles. Ainsi, l'extraction des images des PDF (avec pdfimages) est différente selon la composition du document. Ensuite, les images doivent être améliorées avec ScanTailor (redressement, binarisation). Enfin, on peut procéder à l'OCR (avec Tesseract), qu'il faut ensuite relire (avec PoCoTo), en corrigeant les erreurs récurrentes. Enfin, dans un dernier temps, le texte doit être dissocié du paratexte (titre courant, numéro de page, notes de bas de page), avec Transkribus. Ces différentes opérations sont en cours.

Keywords: sémantique, latin médiéval, lexicographie électronique, linguistique de corpus.

References:

F. Blatt, Y. Lefèvre, J. Monfrin, F. Dolbeau and A. Guerreau-Jalabert (1957-2015). Novum Glossarium Mediae Latinitatis, Copenhague, Bruxelles, Genève.
B. Bon (2015). "Histoire et perspectives du Novum Glossarium Mediae Latinitatis". Archivum Latinitatis Medii Aevi, 73, p. 297-308.
R. Theron, L. Fontanillo, (2013). "Diachronic-Information Visualization in Historical Dictionaries", Information Visualization 14 (2), p. 111-136.
B. Bon (2009-2011). "Outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (1-3)". Bulletin du centre d'études médiévales d'Auxerre, 13-15.
D. Jurafsky, J. H. Martin, (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Upper Saddle River.
Union Académique Internationale and B. Bon (1973-2005). Index scriptorum novus Mediae Latinitatis, Hafniae, Genève.

De VerbaAlpina à APPI : les outils en partage

Esther Baiwir[1], Pascale Renders[1]

[1] *Université de Lille*

Abstract

Le projet APPI *(Atlas pan-picard informatisé),* qui a débuté en 2018, vise à mettre en réseau toutes les données atlantographiques décrivant le parler picard. Ce dialecte a pour particularité de s'étendre sur deux pays, la France et la Belgique, dans lesquels les dialectologues ont rassemblé des collections de matériaux selon des méthodes différentes : l'une purement atlantographique (du côté français), l'autre intermédiaire entre diatopie synchronique et lexicographie historique (du côté belge).

La constitution du corpus se base sur les 660 faits de langue répertoriés dans les volumes 1 et 2 de l'*Atlas linguistique picard* (ALPic), auxquels sont adjoints les matériaux picards de l'*Atlas linguistique de la France* (ALF) et de l'*Atlas linguistique de la Wallonie* (ALW) présentant les mêmes faits de langue. La définition de la macrostructure de l'APPI, toujours en cours, nécessite d'analyser sémantiquement et morphologiquement les matériaux des divers atlas afin de statuer sur leur équivalence.

La seconde étape consiste en une numérisation des matériaux ainsi identifiés. La variabilité des systèmes graphiques nécessite de se tourner vers une transcription en alphabet phonétique international (API), tout en gardant dans le système une trace de la graphie originale.

L'enrichissement des matériaux est envisagé de deux manières : par l'adjonction d'une typisation, qu'elle soit phonétique ou morphologique, et par l'étymologisation des données et leur intégration dans l'une des familles lexicales décrites dans le FEW ; *a minima*, il s'agit d'ajouter aux données dialectales une « référence FEW » (Baiwir/Renders 2019).

D'un point de vue technique, le projet APPI avait pour ambition de créer un nouvel outil en langage XML, assorti d'une interface de consultation et de cartographie. Toutefois, une collaboration fructueuse avec l'équipe munichoise du projet VerbaAlpina (https://www.verba-alpina.gwi.uni-muenchen.de/) a mené à une extension de leur modèle informatique, moyennant divers ajustements apportés à leurs outils.

APPI et VerbaAlpina ne partagent pas leur objet d'étude ; notre projet possède donc un site Web spécifique (https://anr-appi.univ-lille.fr), utilisant un plugin fourni par le projet VerbaAlpina dans une page intitulée *plan interactif*. Ce plugin permet l'affichage des données encodées sous une forme cartographique. La page se divise en deux zones: la carte proprement dite et l'interface d'interrogation. Cette dernière permet de choisir les critères de regroupement des données (conceptuel, étymologique ou morpholexical) et permet également d'afficher des données non langagières, par exemple les points d'enquête de chaque atlas.

Les outils partagés ne consistent pas uniquement en un système de tables et un outil de consultation / visualisation. Le module d'encodage, appelé *Transcription tool*, a également été importé. Il contient trois parties: une zone affichant les cartes originelles des atlas à encoder (en mode image), une zone affichant les règles de transcription et une zone destinée à l'encodage proprement dit. Les différents signes graphiques et agrégats de diacritiques sont encodés grâce à un système d'encodage linéaire appelé *betacode* (https://www.verba-alpina.gwi.uni-muenchen.de/fr/?page_id=13&db=192&letter=B), permettant une ergonomie maximale dans le travail le plus fastidieux. L'encodage se fait en respectant les graphies originelles ; ensuite, un module de transformation du *betacode* en API s'applique. Sur la carte interactive, chaque forme s'affiche par défaut en API lorsque cette transcription est disponible. La forme originale de l'atlas (non API) reste toutefois accessible.

Le bilan tout provisoire de cette coopération est de livrer à la communauté des chercheurs un outil fonctionnel, le site de l'atlas APPI. Dans celui-ci, tout l'ALPic est disponible en mode image, de même que divers documents et aides à la lecture. La ressource APPI proprement dite se présente sous la forme d'un onglet « macrostructure », intégrant 200 notions et rassemblant en mode image les cartes des trois atlas pour ces notions, et d'un onglet « plan interactif », livrant actuellement une vingtaine de cartes complètement accessibles et typisées pour les trois atlas. D'autres, en cours, viendront prochainement enrichir le corpus. Diverses pistes de réflexion ont pu être ouvertes grâce de ce

chantier, qu'elles soient linguistiques (par exemple Kaisin à paraître) ou techniques (voir Brasseur/Baiwir 2019).
Enfin, nous rappelons la dimension collaborative de la construction des outils, grâce au modèle que constitue, sur le plan technique et informatique, le projet VerbaAlpina. Evidemment, l'économie en temps et en moyens a été très importante pour notre projet, mais nous voulons croire que celui-ci a également permis à l'équipe munichoise de porter un regard neuf sur leur outil et de l'améliorer, afin entre autres de le rendre utilisable dans d'autres cadres. Cet exemple d'un partage des outils nous semble être l'un des enseignements majeurs du projet APPI.

Keywords: dialectology, atlas, digitization

References

ALF = Gilliéron, Jules & Edmond Edmont (1902-1910). *Atlas linguistique de la France*, 1920 cartes, Paris, Champion.
ALPic = Carton, Fernand & Maurice Lebègue (1989-1998). *Atlas linguistique et ethnographique picard*, 2 vol. Editions du CNRS.
ALW = Remacle, Louis, Legros, Élisée, Lechanteur, Jean, Counet, Marie-Thérèse, Boutier, Marie-Guy, Baiwir, Esther (1953-). *Atlas linguistique de la Wallonie*, Liège, Université de Liège (10 vol.).
Baiwir, Esther & Pascale Renders, « Numérisation des données dialectales d'oïl : le projet APPI comme laboratoire », publication en ligne dans le cadre du projet APPI (Atlas pan-picard informatisé, sous la direction d'Esther Baiwir), avril 2019, 9 p. (https://appi.univ-lille.fr/data/medias/baiwirrenders2019).
Brasseur, Patrice & Esther Baiwir, « *L'ALN : actualités et perspectives», communication orale lors de la journée d'étude intitulée* Quel dialogue numérique entre les atlas linguistiques galloromans ?, Université de Lille, 26 septembre 2019.
FEW = Wartburg, Walther von *et al.* (1922–2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vol., Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden.
Kaisin, Cécile (à paraître), « L'article défini : entre facteurs d'unité et de césure, un élément emblématique de la langue picarde ?», *Les dialectes de Wallonie* 37.
VerbaAlpina = Krefeld, Thomas | Lücke, Stephan (Hrsgg.) (2014–): VerbaAlpina. Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit, München, online, http://dx.doi.org/10.5282/verba-alpina

eFontes. The Electronic Corpus of Polish Medieval Latin

Krzysztof Nowak

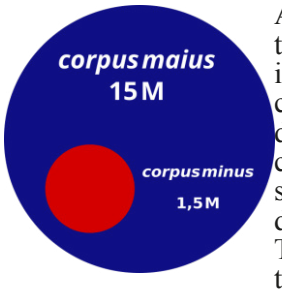
Department of Medieval Latin, Polish Academy of Sciences, Kraków, Poland

Abstract

The Electronic Corpus of Polish Medieval Latin *eFontes* aims to account for the use of the Latin language on the Polish territory during the Middle Ages. Medieval Latin was widely employed all across the Europe from the 8th century onward, however, in Poland its use can be dated back to the beginnings of the 11th century when the foundation of the Polish Kingdom was laid (Stotz 2002). As in Western countries, Latin will be here in permanent contact with the vernacular, contributing to the emergence of complex linguistic situation, that of *diglossia* (Van Acker 2010). As a prestige language, Latin was employed in almost every domain of written official communication from lay and church administration to religious writing to academic instruction to poetry and historiography.

The *eFontes* is a general corpus that aspires to reflect this variety of uses. The texts to be included have been chosen primarily for their communicative function and genre to which they belong, but corpus representativeness is controlled also in regard to the place and time of text composition. Although formally synchronic, the corpus covers the years 1000-1550, which makes for over five centuries of language use. In that, it reflects the historical dynamics of the written production: from the humble beginnings when texts were composed mostly by foreign authors at the court of Polish kings to its use in almost every domain of daily life. The upper chronological limits have been conceived to provide researchers with means to investigate the language used at the outset of Renaissance, for example, to empirically validate the linguistic revolution that is believed to have taken place at that moment. As far as the geographical frame is concerned, the corpus reflects the relative instability of the Polish territory which was subject to major changes during the Middle Ages. The wide definition of what counts as 'Polish territory' that has been adopted in the corpus should allow the research on interference between Latin and vernaculars.

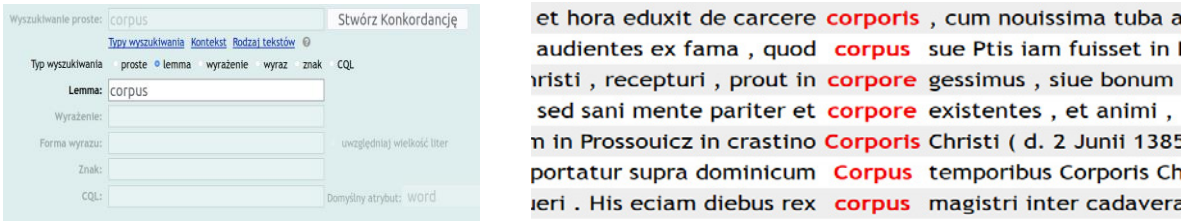
Representativeness, although crucial for corpus design, is also a notoriously difficult goal to achieve. It is even more the case in historical corpora since the relation of historical texts preserved to our times to the language use they are supposed to represent is complex (Hernández Campoy and Schilling 2014). One obvious reason for that is that they often they would come to us due only to chance. Another one is, that, apart from the number of manuscripts, medievalists do not have at their disposal tools that would help in assessing text popularity which is a major factor in corpus design. As for the texts themselves, they are often available only in late copies or have been conserved in fragmented or mutilated form so they can be hardly believed to reflect the actual language use. Finally, for many texts we are simply unable to establish the exact date of creation or author which renders metadata description and, consequently, corpus compilation even more difficult.



All these problems are reflected in a relatively small size of the prototype of the corpus which at the moment contains ca. 5M tokens. By 2023, however, it is expected to be substantially increased to reach 15M tokens. To ensure better control over its balance and representativeness, the corpus is now also being re-designed. Within a large and roughly balanced corpus (*corpus maius*) a strictly controlled sub-corpus (*corpus minus*) is separated that includes mostly text samples. The design that is known from major corpora projects reflects also different interests of its users, both linguists and historians.

The workflow of the project starts with selecting text editions. Once scanned, the image files are OCR-ed with *tesseract* and roughly proofread with *PoCoTo*. Next, a simple algorithm is applied to page regions as recognized by the OCR program in order to, first, separate Latin from non-Latin text and, second, to detect such structural elements as page numbers or marginalia. Once the automatic processing is completed, texts are handed over to annotators who proofread them, ensure that only the Medieval Latin text is preserved in the file and perform a basic content annotation.

Thus prepared, texts are next lemmatised and tagged with part-of-speech labels with TreeTagger (Schmid 1994). In its current state, the annotation is, however, to be used carefully since the project have relied on the tagset and the tagging model that were trained on medieval Latin texts significantly less varied than those gathered in the *eFontes* corpus. This, in turn, results in large number of lemmatisation errors which impact the validity of conclusions one may draw from the corpus at this point. This is going to change diametrically since in the next two years manual annotation of a text sample is planned that will be basis for subsequent training of the tagger.



The *eFontes* corpus is currently published through the NoSketchEngine interface (https://scriptores.pl/fontes). It gives scholars access to basic tools of corpus linguistics, such as CQL queries, collocation lists, concordances etc. In the same time, the corpus query tools does not, obviously, cater for every research scenario. For this reason, the corpus tool will be closely linked to other tools currently developed by the team. On the one hand, the platform *Editiones* allows users to consult editions of the texts included in the corpus (https://scriptores.pl/editiones). On the other hand, more advanced statistics will be available through the R-based *Voces* app (Nowak 2017).

Keywords: corpus linguistics, historical linguistics, Latin

References

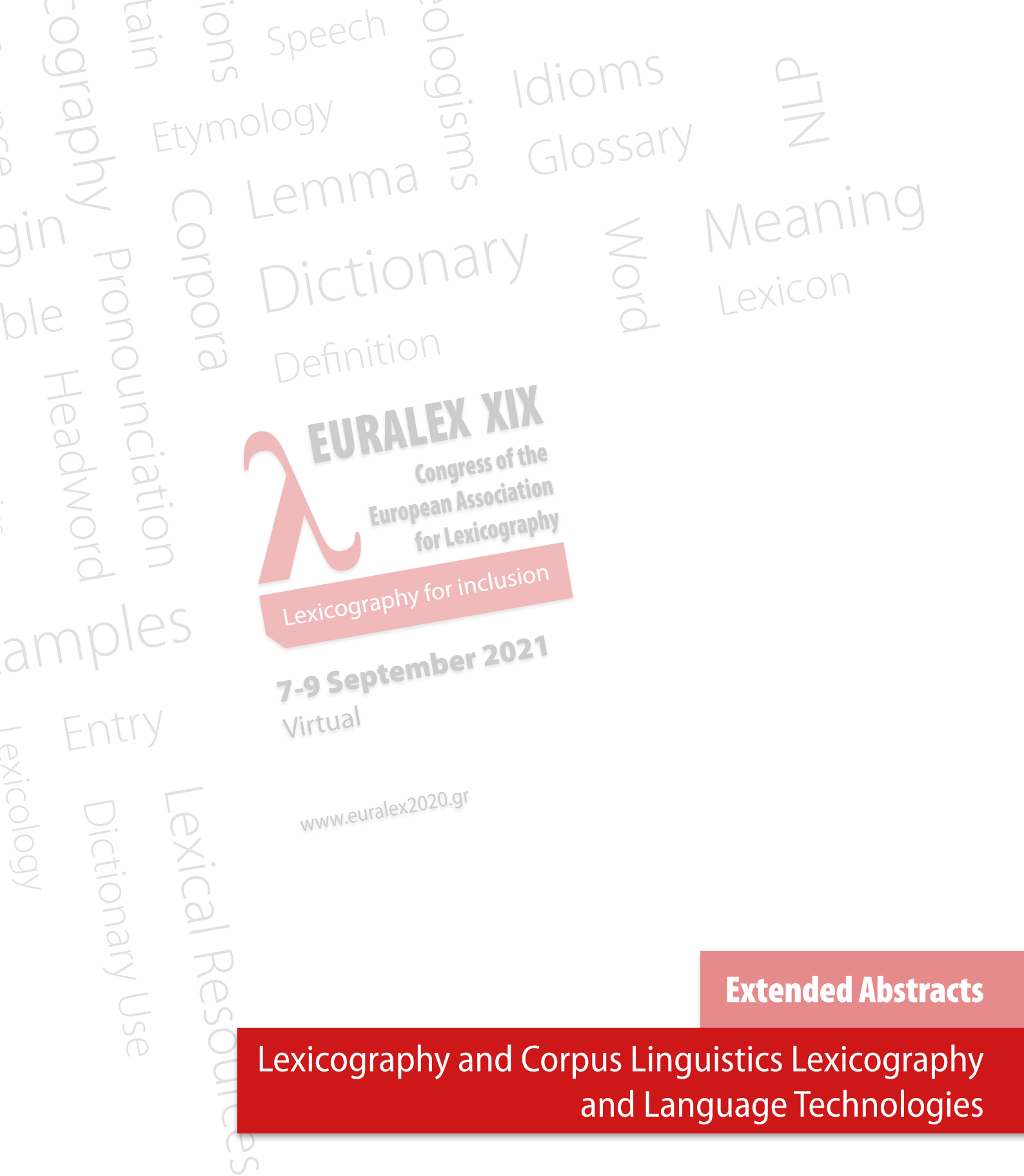
Hernández Campoy, Juan Manuel, and Natalie Schilling. 2014. 'The Application of the Quantitative Paradigm to Historical Sociolinguistics: Problems with the Generalizability Principle'. In *The Handbook of Historical Sociolinguistics*, edited by Juan Manuel Hernández Campoy and J. Camilo Conde-Silvestre, 63–79. Chichester: Wiley Blackwell.

Nowak, Krzysztof. 2017. '*Voces*. An R-Based Dashboard for Lexical Semantics'. In *Digital Humanities 2017. Conference Abstracts*. Montréal: McGill University & Université de Montréal.

Schmid, Helmut. 1994. 'Probabilistic Part-of-Speech Tagging Using Decision Trees'. In *Proceedings of International Conference on New Methods in Language Processing*, 12: 44–49. Manchester, UK.

Stotz, Peter. 2002. *Handbuch zur lateinischen Sprache des Mittelalters: Einleitung, Lexikologische Praxis, Wörter und Sachen, Lehnwortgut*. Vol. 1. München: C.H. Beck.

Van Acker, Marieke. 2010. 'La Transition Latin / Langues Romanes et La Notion de «diglossie»'. *Zeitschrift Für Romanische Philologie* 126 (1): 1–38.



Extended Abstracts

Lexicography and Corpus Linguistics Lexicography and Language Technologies

By the way, do Dictionaries Deal with Online Communication? On the Use of Meta-Communicative Connectors in CMC Communication and their Representation in Lexicographic Resources for German

Andrea Abel

Institute for Applied Linguistics, Eurac Research, Bolzano/Bozen, Italy

Abstract

Introduction & Research Question: Online writing (CMC) is steadily increasing[1]. Partly, specific writing conventions are developing. CMC communication plays an increasing role in educational contexts, too. It is even used in high-stakes tests, such as school leaving examinations[2]. Therefore, the question arises which language resources teachers but also students can rely on in case of doubts. Mapping actual language use is one of the main functions of lexicography. Lexicography is evolving and increasingly considers oral language next to written language use (Eichinger, 2005; Fiehler, 2015; cf. Davies, 2017). In our paper, we investigate the question whether and how interaction-oriented online writing (Storrer, 2013) is represented in dictionaries.

Data & Method: The investigation started with a larger study on the use of German connectors in traditional vs. online writing (cf. Abel and Glaznieks, 2020). In the study, selected connectors (relying on Breindl et al. 2014) where analyzed comparing different corpora:

	corpus	tokens
corpora of interaction-oriented online writing	Wikipedia article discussions	376,478
	Wikipedia user discussions	377,373
	Facebook	373,383
corpora of text-oriented writing	newspaper texts	376,678
	student texts	376,184

In this paper we will address the research question in form of a case study, focusing on two meta-communicative connectors, i.e. *übrigens* (*by the way*) and *das heißt* (*that is to say, i.e.*), considering the Facebook corpus only. The usages where checked with their descriptions in two online general German dictionaries: the "DUDEN online" (Duden Online, no date) and the "DWDS" (Digital Dictionary of the German Language, DWDS, no date) . In addition, we consulted the specialized dictionaries of German particles by Métrich et al. (2009) and Helbig (1994).[3]

Some Results: This section reports some results of the study. We first introduce the main functions of the selected connectors as described in Breindl et al. (2014), then we present the different usages found in our Facebook corpus. Finally, we discuss the representations of the particular usages in the dictionaries mentioned before.

[1] see https://de.statista.com/statistik/daten/studie/475072/umfrage/taegliche-nutzungdauer-von-sozialen-medien/ (16.01.2020)

[2] see e.g. https://www.srdp.at/ (16.01.2020)

[3] On the difficulties of classifying the part of speech (connector, particle, discourse marker, adverb) see e.g. Breindl et al. 2014

Example 1: *übrigens*
Übrigens is used for discourse organization. Usually it refers to a side information, often in form of a parenthesis. It is connected with an (even abrupt) change of subject. There are no restrictions with regard to the syntactic position (Breindl, Volodina and Waßner, 2014).
What is striking in the Facebook corpus is that *übrigens* is used relatively often – although altogether rarely – in pre-prefield position (for a definition see Pasch *et al.*, 2003), especially in comparison with our newspaper and student texts[4]. In our Facebook corpus we detected particular usages, i.e. functions, in the pre-prefield position:

- Signaling a change of subject. The change of subject is less important than the distance between the connector and the action of reference. Thus, the action of reference can be located in a distant part of the interaction or even outside the current interaction. A necessary prerequisite for mutual understanding is always the activation of a common previous knowledge.
- Signaling an attempt to steer the topic of a communication back to a previous one. By doing so, the user tries to bridge a parenthesis, while the connector normally serves to insert a parenthesis in a discourse (see e.g. the definition in Breindl, Volodina and Waßner, 2014).
- Signaling the start of a conversation in an initial post, as in oral conversations (for oral conversations cf. Duden, 2016).

Example 2: *das heißt*
The main function of *das heißt* is to provide a reformulation of the so called "external connect", i.e. of a linguistic expression that does not immediately follow the connector but is linked to it. In addition, it can be used to specify an expression (*genauer gesagt – more precisely*), to generalize an expression (*allgemeiner gesagt – more general*) or to correct an expression (*besser gesagt – or rather*).
Again, we found particular functions in our Facebook corpus:

- Establishing interactional coherence when an external reformulation is used to ensure understanding. In this case it is not the writer who reformulates his or her own expression but the interlocutor (in the sense of *i.e./so you are telling me that …*).
- Self-initiating a self-correction after a slip of the pen, that we know from conversational linguistics. In that case, however, the correction is referred to a slip of the tongue (cf. Pfeiffer, 2015).

To answer the research question, we can summarize the results as follows:

- Differences at a diaphasic and diamesic level are not consistently considered in the dictionaries selected for the study. None of the dictionaries mentions online writing at all. Thus, particular functions of CMC communication as presented before are not represented in the resources. The DUDEN online and the DWDS exclusively present the well-established functions (see Breindl, Volodina and Waßner, 2014). The two specialized dictionaries contain much more detailed descriptions (including e.g. references to oral vs. written language) and lexicographic examples. With regard to CMC communication we have to keep in mind that the specialized dictionaries – or rather, their sources – originated in the 1980ies and 1990ies.
- Differences connected with the syntactic position of the connectors are not considered in any of the dictionaries. Thus, there is no clue about e.g. the particular function that *übrigens* may have in the pre-prefield position (for an analysis of its role in the middle field of a sentence in oral conversations see Egbert, 2003).

Discussion & Outlook: In our study we detected particular uses of two connectors in a Facebook corpus that differ from the descriptions in well-established reference works. However, we would need a larger data base to verify whether our findings represent individual or peculiar cases in our corpus or whether such usages have already become part of everyday language. More generally, large social media corpora for the German language covering different CMC genres and including relevant metadata as well as

[4] *übrigens* in pre-prefield position: newspaper corpus: 5 occurrences out of a total of 41, student corpus: 1 out of 3, Wikipedia AD corpus: 5 out of 118, Wikipedia UD corpus: 28 out of 121, Facebook corpus: 13 out of 55

complete interactions (cf. Imo, 2017) would be a great asset not only for practical lexicography but also for research in applied linguistics.

References

1. Abel, A. and Glaznieks, A. (2020) 'Textqualität in sozialen Medien', in Marx, K., Lobin, H., and Schmidt, A. (eds) *Deutsch in sozialen Medien. Interaktiv – multimodal – vielfältig. Berlin/Boston:* Berlin/Boston: de Gruyter.
2. Breindl, E., Volodina, A. and Waßner, U. H. (2014) *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*. Berlin/Boston: de Gruyter.
3. Davies, W. (2017) 'Gymnasiallehrkräfte in Nordrhein-Westfalen als SprachnormvermittlerInnen und Sprachnormautoritäten', in Davies, W. et al. (eds) *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke*. Tübingen: Narr Francke Attempto, pp. 123–146.
4. Duden (2016) *Die Grammatik: unentbehrlich für richtiges Deutsch*. Berlin: Dudenverlag.
5. Duden Online (no date) 'Duden'. Berlin: Bibliographisches Institut/Dudenverlag. Available at: www.duden.de (Accessed: 20 January 2020).
6. DWDS (no date) 'Das Digitale Wörterbuch der deutschen Sprache.' Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. Available at: www.dwds.de (Accessed: 20 January 2020).
7. Egbert, M. (2003) 'Die interaktionelle Relevanz einer gemeinsamen Vorgeschichte: Zur Bedeutung und Funktion von "übrigens" in deutschen Alltagsgesprächen', *Zeitschrift für Sprachwissenschaft*. Walter de Gruyter, Berlin/New York, 22(2), pp. 189–212.
8. Eichinger, L. M. (2005) 'Standardnorm, Sprachkultur und die Veränderung der normativen Erwartungen', in Eichinger, L. M. and Kallmeyer, W. (eds) *Standardvariation: Wie viel Variation verträgt die deutsche Sprache?* Berlin/Boston: de Gruyter, pp. 363–381.
9. Fiehler, R. (2015) 'Grammatikschreibung für gesprochene Sprache', *Sprachtheorie und germanistische Linguistik*, 25(1), pp. 3–20.
10. Helbig, G. (1994) *Lexikon deutscher Partikeln*. 3., durchg. Leipzig [u.a.: Langenscheidt Verl. Enzyklopädie.
11. Imo, W. (2017) 'Interaktionale Linguistik und die qualitative Erforschung computervermittelter Kommunikation', in Beißwenger, M. (ed.) *Empirische Erforschung internetbasierter Kommunikation*. Berlin/Boston: de Gruyter, pp. 81–108.
12. Métrich, R. *et al.* (2009) *Dictionary of German Particles: Unter Berücksichtigung ihrer französischen Äquivalente*. Berlin/Boston: De Gruyter.
13. Pasch, R. *et al.* (2003) *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer*. Berlin: De Gruyter.
14. Pfeiffer, M. (2015) *Selbstreparaturen im Deutschen: Syntaktische und Interaktionale Analysen*. Berlin/Boston: de Gruyter.
15. Storrer, A. (2013) 'Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia', in Feilke, H., Köster, J., and Steinmetz, M. (eds) *Textkompetenzen für die Sekundarstufe II*. Stuttgart: Fillibach bei Klett, pp. 277–304.

Keywords: CMC Communication, Connectors, General Dictionaries, Specialized Dictionaries

How to Rank Word Senses According to Frequency

Dominika Kováříková[1], **Václav Cvrček**[1]

[1] *Institute of the Czech National Corpus, Charles University, Prague, Czech Republic*

Abstract

1 Introduction: Language corpora are linked to dictionaries from the very beginning of their existence. The extent of language material in corpora led to a great simplification and improvement of a dictionary making process but there is a downside to the volume of the current linguistic evidence: a need to understand the specificity of large data analysis and to find reliable criteria for classification of language phenomena.

The most common lexicographic day-to-day task is identification and description of relevant word senses. Frequency is one of the main characteristics that should be taken into account in non-marginal word sense identification and in ranking word senses (see also Kilgrariff 1997; Lopukhina et al. 2016), even though other factors such as semantic, paradigmatical, morphological or pragmatic criteria (Svensén 2009; Atkins & Rundell 2008) might prevail in the lexicographic description of the word, especially when the frequency criterion is not decisive.

2 Online Tool for Lexicographers: The task of identification and description of word senses can be now facilitated by a free online tool designed for lexicographers. The tool is fast, simple and easy to use, so it does not make the dictionary making process any more difficult. It is helpful in assessing: 1. how many senses of the relevant word are sufficiently attested in the corpus data sample, and 2. how the senses should be ranked based on their frequency in the sample. The tool is a module of a larger application **Calc: Corpus Calculator** (Cvrček 2019; https://www.korpus.cz/calc/) which has been recently released by the Czech National Corpus (CNC).

Calc provides quick support to corpus users when calculating basic statistical tasks commonly encountered in research. It is currently divided into seven modules, each reflecting specific research problems: from simple tasks such as adequate interpretation of word frequency in the corpus or comparing two frequencies between corpora to more complex ones such as establishing the N-gram correspondence for contrastive research on parallel corpora or calculating an index of lexical richness insensitive to the text length. The emphasis is on simple intuitive and interactive interface which is based on statistical software R (R Core Team 2018) specifically the package Shiny for building web apps "straight from R". Unlike similar tools available elsewhere, Calc is task-oriented which means that the app is structured according to the research questions rather than statistical methods.

The module **"Many features – 1 sample"** is inspired by lexicographic work on the Academic Dictionary of Contemporary Czech (Kochová & Opavská 2016) that is currently under way. However, it can be used for various research tasks other than lexicographic (e.g. comparing grammatical structures). It helps evaluate groups of features (typically word senses) identified within a random sample of a concordance. The information about the group frequencies (i.e. frequencies of individual senses identified manually within the concordance), the sample size and the size of the source concordance can be entered directly or it can be extracted from the URL produced by CNC concordancer KonText (Machálek 2014) during the process of sample analysis.

From this input, it is possible to derive the confidence intervals for each word sense using the bootstrapping technique and then evaluate how reliable the information about the existence of the individual groups actually is. Furthermore, it provides information about ranking of the groups (always taking into account the sample size). At the same time, the module calculates the probability of at least one occurrence from a marginal group (e.g. of a minor importance) in the examined sample.

The confidence intervals of group frequencies provide a more adequate and reliable estimate of the true frequency of senses taking into account the pervasive variability among random samples or even

different corpora. These pieces of information help lexicographers to decide which senses are actually attested in the sample and how justified is to order them by their raw frequencies within the sample. The optimum course of action is as follows: 1. establish a corpus sample, 2. identify the senses, 3. enter their frequencies into the lexicographic module of Calc, 4. interpret: what is attested, how to rank the senses, 5. in case of overlapping confidence intervals add more data and repeat the process.

3 Analysis of Word Senses: The reliability and usefulness of the presented tool was tested on several polysemous words found in the British National Corpus (BNC) and the Czech National Corpus (SYN2015): *book*, *cook* and *sweet*, and Czech words *bílý* ('white′), *běh* ('run′, noun) and *automaticky* ('automatically′). To each of the English words, word senses were assigned according to the Collins Online Dictionary, for Czech words we used Academic Dictionary of Contemporary Czech. Frequencies of the word senses in a corpus sample of 200 lines were entered to the lexicographic module in Calc. Figure 1 and 2 show frequencies of word senses of *sweet* and *cook*, respectively. Figure 2 also shows how the reliability of results increases in longer corpus sample. The results were compared to word sense ranking in Collins Online Dictionary, in online Oxford Learner's Dictionaries, and in online Academic Dictionary of Contemporary Czech for Czech words.

4 Results: For five out of the six analyzed words, the most frequent word sense appears as the first one in the entry (the exception was the word *běh* 'run'). Some of the further senses were unattested in the corpus sample: their frequency was either not decisive or they were not present at all, for example *book of matches*, *sweet as a* 'dessert' or *bílý* ('white') used in reference to the White Army. On the other hand, several left out word senses (collocations) that were found in the corpus sample were significantly more frequent than those listed, e.g. *sweet spot*, *sweet nothings*, *bílé pečivo* ('white bread', the opposite for 'whole grain bread').

Based on the detailed analysis of several words, we can conclude that the decision about the presence or ranking of word senses should be based on adequate statistical interpretation of sample analysis. This is facilitated by easy-to-use and fast lexicographic module "Many features – 1 sample" within the Calc application.

Figure 1. Chart for the word ***sweet*** from the lexicographic module of Calc application. Word senses 1 to 13 are listed under the headword *sweet* in the Collins Dictionary, 14 to 21 are other collocations found in BNC. The blue bars represent the word senses sufficiently attested in the corpus sample of 200 words. The orange word senses are not sufficiently attested in the sample. The black lines are confidence intervals: overlapping confidence intervals indicate that the frequency proportions could change in a different corpus sample, and larger sample might produce more reliable results.

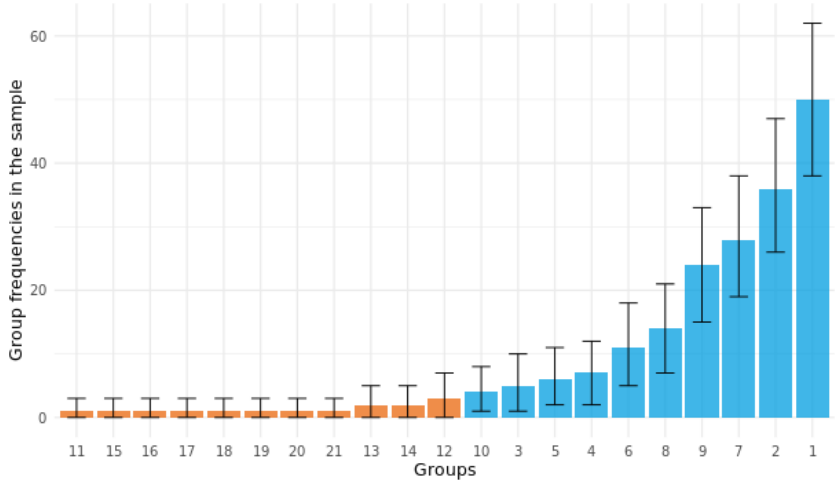
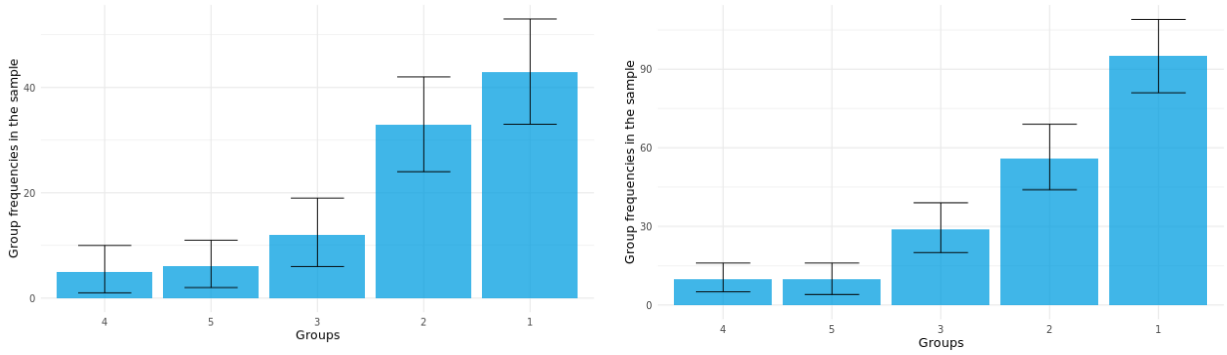


Figure 2. Chart for the word ***cook*** from the lexicographic module of Calc. The blue bars list the word senses sufficiently attested in the corpus sample of 100 (left) and 200 (right) words. There are no orange bars which means that all listed word senses are sufficiently attested in the sample. The black lines indicate confidence intervals. Overlapping confidence intervals on the left imply that the frequency proportions could change in a different corpus sample and more data need to be analyzed. Larger sample (right) provides more reliable results.



Keywords: lexicography, word sense frequency, word sense ranking, corpus-based application

References

Academic Dictionary of Contemporary Czech (2020). Available at http://www.slovnikcestiny.cz, last accessed 16/01/2020.

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

Collins Online Dictionary (2020). Available at www.collinsdictionary.com, last accessed 15/01/2020.

Cvrček, V. (2019). *Calc: Corpus Calculator* (Version 1) [Computer software]. Praha: Institute of the Czech National Corpus, FF UK. Available at https://www.korpus.cz/calc/

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In Sojka, P., Kopecek, I., Pala, K., eds. *Text, Speech, Dialogue. Lecture Notes in Artificial Intelligence* Vol. 3206. Springer Verlag: 103–112.

Kochová, P., Opavská, Z., eds. (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Praha: Ústav pro jazyk český.

Křen, M. et al. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Praha: Institute of the Czech National Corpus, FF UK. Available at http://www.korpus.cz.

Lopukhina, A., Lopukhin, K., Iomdin, B., Nosyrev, G. (2016). The Taming of the Polysemy: Automated Word Sense Frequency Estimation for Lexicographic Purposes. *Proceedings of the XVII. Euralex International Congress*, pp. 249–256.

Machálek, T. (2014). *KonText – aplikace pro práci s jazykovými korpusy* [Cs]. Praha: Institute of the Czech National Corpus, FF UK. Available at https://kontext.korpus.cz

Oxford Learner's Dictionaries (2020). Available at https://www.oxfordlearnersdictionaries.com/, last accessed 16/01/2020.

R Core Team (2018). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Available at https://www.R-project.org/

Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge, Melbourne, New York: Cambridge University Press.

The British National Corpus, version 2 (BNC World) (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Praha: Institute of the Czech National Corpus, FF UK. Available at http://www.korpus.cz.

Merging Collocations from Russian Dictionaries with Corpus Data into a Unified Database

Maria Khokhlova[1], **Jelena Kallas**[2]

[1] *Department of Mathematical Linguistics, St Petersburg State University, St Petersburg, Russia*
[2] *Institute of the Estonian Language, Tallinn, Estonia*

Abstract

The paper deals with the Russian collocations database created within the project on studying collocability in 2018 [deleted for anonymity]. The given lexical database can be used for language learning applications and other research tasks. We discuss the part of the database including three printed Russian dictionaries ([DRL 1981-1984; Deribas 1983; Borisova 1995]) and one electronic dictionary [Kustova 2008]. The aim of the study is threefold. First, to analyze how much the dictionaries overlap and give the same results. Second, to investigate to what extent the collocations that are presented in the dictionaries can be rated by users as significant ones. To a certain degree we can say that the latter deals with the evaluation of lexicographic work from users' viewpoint. Third, we analyze possibilities for the enrichment of the data by means of methods used in automated lexicography for collocation extraction. According to [Kallas et. al 2019: 24] automatic collocation extraction was implemented by 12.7% out of 150 lexicographers participated in the survey of lexicographers needs. As a source we will use the Russian Web corpus available in Sketch Engine and that would enable to have a balance between dictionary and corpus collocations.

When merging dictionary data from different resources it is not only the question of a unified lexicographic format that should be discussed but also of data relevance. Describing collocations a lexicographer needs to select ones based on their representativeness in corpora, coverage in dictionaries and also appropriateness for language users and their purposes. Lexicographers put much more attention to a user in order to accommodate user needs. There is a range of works on the topic [Dziemianko 2010; Heid, Prinsloo, Bothma 2013; Lew, Schryver 2014] but, to our best knowledge, such an analysis was not provided for Russian dictionaries and there is still a need in up-to-date tools that follow modern trends in lexicography.

In our investigation we interpret collocability of a lexical unit extremely widely and understand it as the ability to connect with other lexical units. According to Teliya's [Teliya 1996] classification: idioms (rabochaja loshad' "working horse"), phraseological units (teljachij vostorg "foolish enthusiasm"), stock phrases (vsego khoroshego "all the best"), cliche (minutu vnimanija "minute of attention").

Altogether the examined part of the database contains 11,210 phraseological units extracted from [DRL 1981-1984] that were marked with a special diamond '◊' symbol. The total amount of the headwords that had phraseological units in their entries was 5,955 (about 7.4% from the entire word list of the dictionary). The initial preprocessing of [Borisova 1995] involved OCR procedure and resulted in a total sum of 3,058 collocation candidates. Additional 232 collocations were added to the database from the quotations. The dictionary [Deribas 1983] comprises more than 3,770 collocations with perfective and imperfective verb pairs and 383 with only imperfective verb forms. After processing the dictionary data, excluding prepositional phrases and representing phrases with perfective and imperfective verbal forms as different pairs we had a list of 7,923 items. We confined ourselves to noun phrases (about 7,000 collocations) from [Kustova 2008]. In total the mentioned dictionaries comprised about 29,000 verbal and noun collocations.

The second phase was to investigate the overlap between the dictionaries and to merge collocations. The merged list of the dictionaries [DRL 1981-1984; Kustova 2008] had 10,205 positions and only 59 of them were described in both dictionaries (less than 1% of the whole list). It can be explained by the fact that [Kustova 2008] describes collocations with denoting high intensity whereas [DRL 1981-1984] aims at a comprehensive representation of the lexicon and also puts emphasis on phraseological units. We also made a deeper analysis of the merged collocation list to understand what groups of lexis (and keywords) tend to form the nucleus and what can be seen as specific ones for a certain dictionary.

Second, the paper presents the results of the survey conducted in November 2019 - January 2020. The purpose was to evaluate the extracted collocations and to investigate the needs of Russian language learners and user perception. The first group involved 87 students studying computational linguistics. The first task assumed lists of keywords and the respondents were asked to write collocates for them. As the second task they were given collocation candidates from the merged list and had to mark the ones that should be presented in dictionaries or called collocations. The results show that about 63% of phrases were selected as ones to be described in a dictionary and 89% of phrases were labeled as collocations. The rest of the list might be seen as terms or free word combinations. The second group consisted of 33 foreign students studying Russian (level A2/B1). The respondents were asked whether they use Russian monolingual dictionaries and if yes, whether they consult printed or electronic ones, what kind of data they search for. Also they were given random samples of collocation candidates and had to, first, tick the known ones and second, to fill the gaps suggesting collocates to a certain number of keywords.

One can observe the following tendency, i.e. subjects more often note as relevant collocations those that were presented in the electronic dictionary rather set phrases presented in printed dictionaries. That is, paper dictionaries either give low-frequency or obsolete examples that are not perceived by native speakers (especially of the young generation) as significant ones. A user can be overwhelmed by the data presented in various dictionary entries. And when merging the collocation lists special attention should be paid to filtering out the results and their representation. As the next step, we are going to apply corpus-based approach and compare collected data with the data extracted from the new Russian Web corpus. Comparing dictionary and corpus collocations will allow to evaluate and to mark them accordingly that can be true for collocation candidates missing in corpora or hapax legomena, e.g. possible obsolete lexis or specialized language. In future filtered collocations will be integrated into Sõnaveeb [Tavast et al 2018].

References

Borisova, E. G. (1995b). A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords [Slovo v tekste. Slovar' kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Moscow.

Deribas, V. M., (1983). Verb-Noun Collocations in Russian. Moscow: Russian language [Ustojchivyje glagol'no-imennyje slovosochetanija russkogo jazyka]. Moskva: Russkij jazyk.

Dictionary of the Russian Language [Slovar' russkogo jazyka v 4 tomakh], (1981–1984). Yevgen'yeva, A. P. (ed.-in-chief). Vol. 1–4, 2nd edition, revised and supplemented. Moscow: Russkij jazyk.

Dziemianko, A. (2010). Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. International Journal of Lexicography, Volume 23, Issue 3. P. 257-273.

Heid, U., Prinsloo, D., Bothma, Th. (2013). Dictionary and corpus data in a common portal: state of the art and requirements for the future. Lexicographica - International Annual for Lexicography / Internationales Jahrbuch für Lexikographie. Volume 28. P. 269-291.

Kallas, J., Koeva, S., Langemets, M., Tiberius, C., Kosem, I. (2019) Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs. In Kosem, I., Zingano Kuhn, T., Correia, M.,

Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.) 2019. 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. P. 519-536.
Kallas, J., Koppel, K., Langemets, M., Männiko, K., Nurk, T., Viks, Ü. Developer OÜ TripleDev: Martin Laubre, Raigo Ukkivi, Arvi Tavast, Sander Lastovets, Sander Rautam. Eesti Keele Instituut. Sõnaveeb 2019. Available at http://www.sonaveeb.ee (accessed February 2, 2020).
Tavast, A., Langemets, M., Kallas, J., Kristina, K. (2018). Unified data modelling for presenting lexical data: The Case of EKILEX. In Proceedings of the XVIII EURALEX International Congress. EURALEX: Lexicography in Global Contexts. Eds. Čibej, J., Gorjanc, V., Kosem, I. and Krek, S. Ljubljana, 17–21 July 2018. Ljubljana: Ljubljana University Press, Faculty of Arts, 749−761.
Kustova, G. I. (2008). Dictionary of Russian Idiomatic Expressions [Slovar' russkoj idiomatiki. Sochetanija slov so znachenijem vysokoj stepeni]. Available at http://dict.ruslang.ru (accessed February 2, 2020)
Lew, R. Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution. International Journal of Lexicography. Volume 27, Issue 4. P. 341–359.
Teliya, V. N. (1996). Russian Phraseology: Semantic, Pragmatic and Cultural Aspects [Russkaja frazeologija: semanticheskij, pragmaticheskij i lingvokul'torologicheskij aspekty]. Moscow.

Keywords: Collocations, Russian language, paper dictionaries, lexicographic database, language users

Accessing the lexicon of Gronings with WoordWaark

Wilbert Heeringa[1], **Goffe Jensma**[1], **Anna Pot**[1]

[1] *Centre Groningen Language & Culture, University of Groningen, Groningen, The Netherlands*

Abstract

The regional Groningen language is spoken in the furthest northeast of the Netherlands by approximately 310,000 speakers. It belongs to the Dutch part of Low Saxon which is recognized according to Part II of the European Charter for Regional or Minority Languages. This recognition implies the facilitation and/or encouragement of the use of the language in speech and writing in public and private life.

In May 2018 we started developing WoordWaark which means literally 'word work'. This web app gives the Groningen community access to the lexicon of the language. Additionally, the app offers the speakers the opportunity to help build the inventory of the lexemes on which the program is based. This makes the language database highly dynamical, and strengthens the involvement of the Groningen community.

In WoordWaark users have access to the Groningen lexicon in three different ways: 1) via a search in multiple traditional dictionaries at once, 2) via a corpus search, 3) via a clickable map by which the lexica of individual locations are accessed.

1. Search multiple dictionaries at once

Currently two dictionaries are available in WoordWaark. An interface that makes it possible to search in multiple dictonaries at once needs to be convenient arranged. Existing apps that offer searching in multiple dictonaries are Taalweb Frysk (https://taalweb.frl/) and OneLook (https://www.onelook.com/) . Both apps served as examples when designing the interface in WoordWaark.

We aim to add more dictionaries. We also plan to enrich one of the dictionaries with audio samples by which users can hear the pronunciation of the lemma, thus obtaining a talking dictionary.

2. Corpus search

Users can search via word tokens that are found in a corpus. On entering a word token, the program displays the sentences in which the word appears. The search can be narrowed by entering a region, location, author and/or time interval.

Currently, the corpus consists of only 20, 750-word tokens in 33 articles written by 12 authors found in two journals and one newspaper. We plan to extend the corpus by including text of more than 850 books - novels, poetry collections and informative books - that appeared in a period 1822-2016.

The digitalization of the corpus includes several challenges. First, the sources appeared in a time interval of almost 200 years, and it is known that the dialect changed in this period of time, for example as the result of strong provincialism (in the first half of the 19th century) and later by the influence of Standard Dutch. Therefore, variation due to diachronic pronunciation changes is expected to be found in the corpus. Second, the spelling of words in Gronings changed. Third, the Groningen regional language shows geographical dialect variation.

We will build a search interface so that when searching for a particular word token, other (diachronic/spelling/regional) variants are found as well. For example, when a user enters *twai* 'two', variants such as *twai*, *twaie*, *twij*, *twije*, *twei*, *tweie*, *twie*, *twee* are also retrieved from the corpus. This requires that all variants should be linked with a standard form, for example *twij*.

If different conjugations (such as diminutives and plurals) exist for a word token, the user may wish to obtain all (non-)conjugated forms when entering one of them. Therefore, conjugated word tokens should be linked to their lemma.

3. Search lexica of individual locations

Users can access the lexicon by clicking on a location in a map. By clicking on a location, the user gets access to a part of the lexicon of the Groningen dialect spoken in that location. In WoordWaark two such maps are found.

Map 1: Question and answer

When clicking on a location in the map the user gets access to the results of large-scale questionnaires that were conducted between 2012 and 2015. The results of the questionnaire of the location that was chosen by the user are presented as a small dictionary. Users can search this dictionary by entering words in Gronings, Dutch, Frisian and English.
Dictionaries usually contain words of a language spoken in one or more countries or in a particular region. However, our approach where each place has its own unique dictionary does more justice to the fact that the Groningen dialect differs from location to location.
When a user finds a word lacking in the dictionary of the location where his or her dialect is spoken, s/he can add the word via the donate tab, where the user is also asked to pronounce the word while this is recorded. After having submitted the word with its translation the data is immediately added to the dictionary and accessible via the clickable map.
We also aim to include material collected by the app Stemmen, which means 'voices'. The idea behind this app is taken from Adrian Leemann. In 2013 he developed the *Dialäkt Äpp* in order to collect Swiss German dialect material. Using this app, users indicate which variants they use for each of *x* words. Next, the application guesses their local dialect. The Groningen version of this app was developed by Nanna Hilton (supervisor) and Daniel Wanitsch (programmer). Stemmen is an example of citizen science, i.e. scientific research conducted, in whole or in part, by amateur (or nonprofessional) scientists. The Stemmen app is available in WoordWaark.
Map 2: From old to young
When clicking on a location in the map the user gets access to more archaic and/or eccentric words or phrases that elderly speakers from 17 villages wanted to pass on to the next generation. As for the local dictionaries, this data is available per location as well. For each word a recording is provided in which the word or phrase is discussed by the speaker and an interviewer.
The app is developed in the R language using the Shiny framework. Additionally, JavaScript, CSS and HTML are used. WoordWaark is available at: https://woordwaark.nl/. The app works best on a device with landscape orientation with a minimum screen size of 360 by 640 points (e.g. Samsung Galaxy SIII).

Keywords: corpus, dialects, lexicography, lexicon, minority languages, regional languages

Multiword Expression identification and extraction from corpora: the case of Slovenian

Simon Krek[1,2], Polona Gantar[2], Iztok Kosem[1,2], Cyprian Laskowski[2], Janez Brank[1]

[1] *Jožef Stefan Institute, Artificial Intelligence Laboratory, Ljubljana, Slovenia*
[2] *University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia*

Abstract

In the abstract, we present work on building a machine-readable multiword expressions (MWE) lexicon of Slovenian from existing language resources, and identification of MWEs included in the lexicon in the latest version of Gigafida corpus, as well as extraction of examples from the corpus to be included in the final version of the MWE lexicon.

Language resources used

Slovene Lexical Database[1] (SLD) was created between 2008 and 2012 and represents a comprehensive syntactic and semantic description of a selected set of Slovene words. The database is structured as a network of interrelated semantic and syntactic information about a particular word. Semantic level represents the top level in the hierarchy with the lexical unit as its core element. This includes all senses of the headword, multi-word expressions and phraseological units. The syntactic level of description includes syntactic structures and corresponding collocations. All the higher types of information are confirmed by a selection of corpus examples. Multi-word expressions and phraseological units are treated independently from particular senses of the headword and have their own internal structure which requires the same types of information as single-word entries or senses. The resource is available under CC BY-NC-SA license.

Dictionary of Slovene Phrasemes[2] (DSP) includes the description and explanation of 13,125 Slovenian phrasemes. The use of phrasemes is represented by citations from lexical files built for the Dictionary of the Slovenian Standard Language as well as from the Nova beseda corpus and other resources. The entries also contain etymological information and equivalents in other languages. The resource is available under CC BY-NC-SA license.

Gigafida 2.0[3] is the upgraded version of the Gigafida corpus published in 2019. It includes 1,2 billion words and unlike its predecessor, it is a corpus of standard Slovene. Other improvements include the removal of duplicate texts and text fragments, an improved automatic linguistic tagging, and several new features in the interface design. New (automatic) annotation levels include: syntactic parsing according to JOS[4] and UD[5] standards, named entities and semantic role labels.

Creation of MWE lexicon, syntactic parsing and description of syntactic structures

The process of MWE lexicon creation started with the extraction of MWE headwords from both existing lexical resources, SLD and DSP. The extracted headwords were syntactically parsed with the same parser as was used for parsing of Gigafida corpus. Based on the result, the first version of the MWE lexicon was created which included information about internal (syntactic) composition of MWEs, together with morphological information about particular components. We provide an example with syntactic structure ID, seven components, lemma & morpho-syntactic descriptions (msd) and single-word lexicon IDs (Slo. vrteti se kot mačka okrog vrele kaše, /to spin around the boiled porridge like a cat/ Eng. “to be irresolute, to beat about the bush”)

```
<lexical_unit id="ssf.3762" syntactic_structure_ID="1001179">
    <!-- vrteti se kot mačka okrog vrele kaše -->
    <components>
        <component cid="1">
            <lexeme lemma="vrteti" msd="Vmpn" sloleks_ID="">vrteti</lexeme>
```

[1] http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza
[2] https://www.clarin.si/repository/xmlui/handle/11356/1129
[3] https://viri.cjvt.si/gigafida/System/About
[4] http://nl.ijs.si/jos/index-en.html
[5] https://universaldependencies.org/

```
            </component>
            <component cid="2">
                <lexeme lemma="se" msd="Px------y" sloleks_ID="">se</lexeme>
            </component>
            <component cid="3">
                <lexeme lemma="kot" msd="Cs" sloleks_ID="">kot</lexeme>
            </component>
            <component cid="4">
                <lexeme lemma="mačka" msd="Ncfsn" sloleks_ID="">mačka</lexeme>
            </component>
            <component cid="5">
                <lexeme lemma="okrog" msd="Sg" sloleks_ID="">okrog</lexeme>
            </component>
            <component cid="6">
                <lexeme lemma="vrel" msd="Ncfsg" sloleks_ID="">vrele</lexeme>
            </component>
            <component cid="7">
                <lexeme lemma="kaša" msd="Ncfsg" sloleks_ID="">kaše</lexeme>
            </component>
        </components>
    </lexical_unit>
```

Syntactic structure ID 1001179 is formally encoded as a combination of seven components with particular syntactic links and labels, together with underlying morphological features – we provide the formal description below using the JOS annotation system:

```
<syntactic_structure id="1001179">
    <system type="JOS">
        <components>
            <component cid="1" type="core" />
            <component cid="2" type="core" />
            <component cid="3" type="core" />
            <component cid="4" type="core" />
            <component cid="5" type="core" />
            <component cid="6" type="core" />
            <component cid="7" type="core" />
        </components>
        <dependencies>
            <dependency from="#" label="#" to="1" />
            <dependency from="1" label="del" to="2" />
            <dependency from="4" label="vez" to="3" />
            <dependency from="1" label="tri" to="4" />
            <dependency from="6" label="dol" to="5" />
            <dependency from="4" label="dol" to="6" />
            <dependency from="6" label="dol" to="7" />
        </dependencies>
        <definition>
            <component cid="1">
                <restriction type="morphology">
                    <feature POS="verb" />
                    <feature type="main" />
                </restriction>
            </component>
            <component cid="2">
                <restriction type="morphology">
                    <feature POS="pronoun" />
                    <feature type="reflexive" />
                </restriction>
            </component>
            <component cid="3">
                <restriction type="morphology">
                    <feature POS="conjunction" />
                </restriction>
            </component>
            <component cid="4">
                <restriction type="morphology">
                    <feature POS="noun" />
                    <feature case="nominative" />
                </restriction>
            </component>
            <component cid="5">
```

```
                <restriction type="morphology">
                    <feature POS="adposition" />
                </restriction>
            </component>
            <component cid="6">
                <restriction type="morphology">
                    <feature POS="noun" />
                    <feature case="genitive" />
                </restriction>
            </component>
            <component cid="7">
                <restriction type="morphology">
                    <feature POS="noun" />
                    <feature case="genitive" />
                </restriction>
            </component>
        </definition>
    </system>
</syntactic_structure>
```

These two new resources – parsed lexicon and the collection of syntactic structures – provided the possibility to identify MWE in the corpus parsed with the same (syntactic) annotation system.

Multiword expressions identification and extraction procedure

An algorithm was devised that takes into account all lexical (lemma), morphological (msd) and syntactic (dependencies, lables) features included in the lexicon in individual entries, and identifies sentences with these elements in the corpus. However, as MWEs notoriously exhibit variation and flexibility, the algorithm also allows for syntactic slots that are required in syntactic terms but are not fixed in lexical terms. The algorithm requires that at least two content words (nouns, verbs, adjectives, adverbs) are present in the identified sentence, regardless of the number of components, in any combination. In the case of the MWE presented above, the algorithm produced 29 sentences, with variants in the verbal ("vrteti se" – to spin) and nominal ("mačka" – cat) slot. The columns include: (1) MWE ID, (2) syntactic structure ID, (3) corpus example ID, (4) syntactic slots filled with expected lexical elements, (5) positions of components in the sentence, (8) word forms of identified components and (8) example sentence. We present an example below:

```
(1) ssf.11812
(2) 1000557
(3) GF0215532.131.1
(4) |   CCaCaC  |
(5) 1   3       4       5       6       7       |
(8) Hodijo      kot     mačke   okrog   vrele   kaše    |
(9) Hodijo torej kot mačke okrog vrele kaše, zaključujem beležko.
```

In the full paper we will present the resulting lexicon in more detail, with numbers and detailed description of the procedure.

Keywords: Multiword expressions, machine readable lexicon, syntactic parsing, multiword expression extraction from corpora.

Semi-Automatic Detection of Multi-Word Latinisms in Large Corpora

Vladimír Benko[1,2], **Katarína Rausová**[1], **Michal Škrabal**[3]

[1] *Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia*
[2] *UNESCO Chair in Plurilingual and Multicultural Communication, Comenius University in Bratislava, Slovakia*
[3] *Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic*

Keywords: multi-word Latinisms, corpus-driven approach, semi-automatic procedure.

Introduction. As in many other European languages, Latin has played very important role in both Czech and Slovak since the early Middle Ages (starting with the Cyril and Methodius's mission in the 2nd half of the 9th century). This influence reached its height during the period of humanism, whereas in modern history (19th and 20th centuries) it gradually declined. Yet even today, Latinisms are an integral part of both languages. In addition to the single-word Latinisms that also often enriched the Czech/Slovak lexicon through other languages (German, possibly also Romance languages, and in Slovak, through Czech), there are also multi-word Latinisms on which we focus primarily in our paper. They represent a specific group of stable collocations with higher stylistic value (bookish expressions) and are used mainly in professional and journalistic texts, often with a special effect. Lists of Latinisms are mostly provided by dictionaries of catch phrases and proverbs or general Latin-Czech/Slovak dictionaries (Kuťáková et al., SSČ, NASCS, ASSČ for Czech, and Čermák & Čermáková, SSJ, SSSJ for Slovak). However, up to this point no one has attempted their automatic detection in corpora in order to obtain an inventory of truly live phrases of Latin origin in contemporary Czech and Slovak, with their real frequency distribution.

The Procedure. The current work has been inspired by our similar project carried out earlier this year targeted at identification of Latin phrases (written in Latin script) in a large Russian corpus (Benko & Rausová, 2020). The idea was based on using a Latin morphosyntactic analyser and extracting expressions containing continuous sequences of words recognized by it – *TreeTagger* (Schmid, 1994) is a suitable tool here as has has a Latin language model with reasonable coverage of Latin lexicon available, and it also provides for explicit indication of out-of-vocabulary words (*OOVs*) encountered during analysis. Encouraged by the success, we wanted to apply the methodology to Czech and Slovak corpora. The "Russian" pipeline consisted of steps as follows:

1. Take the tokenized corpus, delete annotation (if any)
2. Run a Latin *TreeTagger* on it
3. Delete tokens tagged as <unknown> (leaving empty lines)
4. Delete tokens containing numbers and punctuation
5. Delete annotation (tags and lemmas)
6. Merge multiple empty lines
7. Change newline characters after consecutive non-empty lines to spaces (i.e., putting multi-word expressions to single lines)
8. Produce a frequency list

Our original intention was to use the very same procedure for Czech and Slovak languages as well. The results of it applied to one Slovak sentence[1] is shown in Table 1. As it can be seen, besides one "correct" expression "*corpus delicti*", numerous "false homographs", i.e., Slovak words identified as Latin are present in the output. We may not care about the single-word items, but one multi-word "Latin expression" can be seen there. Due to high frequency of false homographs in the corpus (compared to low frequency of the real Latin phrases), analysis of the raw list was not feasible, and a supplementary procedure to get rid of apparently non-Latin items was necessary. We decided to use the language identification approach suggested by Vít Suchomel (2019) based on summing the normalized frequency of each word (in logarithmic scale) of the expression based on (lowercased) frequency lists derived from large web corpora. In our case, the respective frequency lists were extracted from the Aranea family of web corpora (Benko, 2016), and, inspired by the Suchomel's experiment, we opted for 8 languages (Latin, Slovak, Czech, English, French, Spanish, Italian, and German), i.e., those that are most likely to appear in Czech and Slovak texts.

Example of this secondary language identification is shown in Tables 2 and 3. It can be seen that the "*corpus delicti*" expression was correctly classified as being Latin, whereas "*a do*" was misidentified as Czech. In our case, however, we did not care much about correct classification of languages other than Latin. Moreover, in this particular case, "*a do*" is a valid expression both in Czech and Slovak, and the correct classification is not possible

[1] http://www.kvhbeskydy.sk/index.php/2012/12/03/hra-s-kostami-mrtvych-vojakov

without a larger context being considered.

The Results. Due to space restrictions imposed on this abstract, we provide comprehensive summaries for Slovak only, where the new version of the 4 Gigatoken *Araneum Slovacum Maximum* web corpus was used for our computation. The primary identification phase produced a frequency list of 1,040,126 multi-word expressions, out of which 446,311 had a non-hapax frequency. The secondary classification yielded to results shown in Table 4. The blue example expressions are in fact misclassified Slovak phrases (in some cases with missing diacritics), which is relevant in the first row only (i.e., decreasing the precision), while the red ones should have been in fact classified as Latin (i.e., affecting the recall). Results of the preliminary manual classification of the first 1,000 items identified as Latin by the automated procedure are summarized in Table 5.

The Czech results, based on the 2.25 Gigatoken *Araneum Bohemicum Novum MMXVII* show quite similar tendencies, and will be provided in the full version of our paper.

Conclusions and Further Work. Though some issues – both computational and linguistic – still persist (we'll also discuss them in the longer version of the text), the two-phase identification procedure proved to be surprisingly successful. We believe that it can be used by lexicographers not only for identification of multi-word Latinisms (in corpora of almost any Language), but also for detection of foreign expressions coming from languages other than Latin. Within the context of Czech and Slovak, multi-word English ("*Colours of Ostrava*"), French ("*crème de la crème*"), and German ("*Wien bleibt Wien*") expressions, often containing proper nouns, seem to appear in corpus texts quite often. We definitely plan to apply the described methodology to other Czech and Slovak corpora, so that frequency distributions by text types and other corpus metadata available could be obtained. We also plan to publish the normalized frequency lists derived from all Aranea corpora to provide for replicability of our research.

References

Benko, V. (2016). Two Years of Aranea: Increasing Counts and Tuning the Pipeline. In Proceedings of the Ninth International Conference on Language Re-sources and Evaluation (LREC 2016). Portorož: European Language Resources Association (ELRA), 2016, pp. 4245–4248. ISBN 978-2-9517408-9-1.

Benko, V. & Rausová, K. (2020). Data-Driven Approach to Identification of Latin Phrases in Russian Web-Crawled Corpora. Submitted to CompLing 2020 Conference, St. Petersburg, Russia. 20–21 June 2020.

Čermák, J. & Čermáková, K. (2016). Slovník latinkých citátov. Bratislava: Ikar.

Karlíková, H. (2017): Latinismy v českém lexiku. In: P. Karlík, M. Nekula, J. Pleskalová (eds.), CzechEncy – Nový encyklopedický slovník češtiny. URL: https://www.czechency.org/slovnik/LATINISMY V ČESKÉM LEXIKU (Accessed at: 31. 1. 2020)

Kuťáková, E., Marek, V. & Zachová, J. (2002). Moudrost věků. Lexikon latinských výroků, přísloví a rčení. Leda: Praha 2002.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Suchomel, V. (2019). Discriminating Between Similar Languages Using Large Web Corpora. In A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 129–135. Brno: Tribun EU.

Dictionaries

ASSČ: Akademický slovník současné češtiny. Ústav pro jazyk český AV ČR: Praha. Available online. URL: http://www.slovnikcestiny.cz.

NASCS (2005). Kraus, J. et al.: Nový akademický slovník cizích slov A–Ž. Praha: Academia.

SČFI (2009). Čermák, F. et al.: Slovník české frazeologie a idiomatiky [vol. 1–4]. Praha: Leda.

SSČ (1994). Slovník spisovné češtiny pro školu a veřejnost. Praha: Academia.

SSJ. (1959–1968). Slovník slovenského jazyka. Bratislava: Vydavateľstvo SAV. Available online. URL: http://www.juls.savba.sk/ssj_peciar.html.

SSSJ (2006–2016): Slovník súčasného slovenského jazyka I, II & III. Bratislava: VEDA. Available online. URL: http://www.juls.savba.sk/pub_sssj.html.

Table 1. Steps 2, 3, and 4+5 of the primary language identification

Word	Tag	Lemma	Word	Tag	Lemma	Word
Treba	N:voc	<unknown>				
dodať	N:voc	<unknown>				

,	PUN	,	,	PUN	,	
že	ADJ	<unknown>				
do	V:IND	do	do	V:IND	do	do
prílohy	ADJ:NUM	<unknown>				
listu	N:dat	<unknown>				
už	ADJ	<unknown>				
úradníci	V:IND	<unknown>				
tento	V:PTC:abl	tendo|teneo	tento	V:PTC:abl	tendo|teneo	tento
zoznam	N:acc	<unknown>				
samozrejme	V:IND	<unknown>				
nepriložili	N:dat	<unknown>				
a	PREP	a	a	PREP	a	a
do	V:IND	do	do	V:IND	do	do
dnešného	N:dat	<unknown>				
dňa	V:PTC	<unknown>				
nám	V:IND	<unknown>				
ani	ADJ	anus	ani	ADJ	anus	ani
žurnalistom	N:acc	<unknown>				
tento	V:IND	tento	tento	V:IND	tento	tento
"	PUN	"	"	PUN	"	
corpus	N:nom	corpus	corpus	N:nom	corpus	corpus
delicti	N:gen	delictum	delicti	N:gen	delictum	delicti
"	PUN	"	"	PUN	"	
neukázali	ADJ	<unknown>				
.	SENT	.	.	SENT	.	

Table 2. Secondary language identification: "*corpus delicti*"

	la	**sk**	**cs**	**en**	**fr**	**es**	**it**	**de**
corpus	2.6013	0.0000	0.0000	0.7134	0.9839	0.6983	0.6762	0.0000
delicti	0.7660	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	3.3673	0.0000	0.0000	0.7134	0.9839	0.6983	0.6762	0.0000

Table 3. Secondary language identification: "*a do*"

	la	**sk**	**cs**	**en**	**fr**	**es**	**it**	**de**
a	3.7136	4.4429	4.4486	4.2598	3.7696	4.2349	4.1105	2.2192
do	1.4912	3.6983	3.7110	3.1957	1.0768	1.9010	1.5569	1.1399
	5.2048	8.1412	**8.1596**	7.4555	4.8464	6.1359	5.6674	3.3591

Table 4. Secondary language identification: Summary

Lang	**Count**	**%**	**Examples**
la	66,964	15.00	*de facto, in vitro, si iste, in memoriam*
sk	176,534	39.55	*na tento, ale tento proces, syra na, inde na tele*
cs	77,355	17.33	*ale ani, do USA, na sever, doma proti*
en	45,546	10.20	*transfer do, Tesco Mobile, Creative Suite, in Paris*
fr	11,820	2.65	*et al, si vie, Credit Suisse, de jure, Lynx lynx*
es	20,393	4.57	*Maria de, si mas, no da, Carmen Electra*
it	44,590	9.99	*ti ma, nano SIM, nato mi, Opera Mobile*
de	3,109	0.70	*an der, in an, die Presse, video hier*
	446,311	100.00	

Table 5. Manual analysis of the first 1,000 "Latin" items

Language	**Count**	**%**	**Example**
Latin	582	58.2	*ad hoc, status quo, in situ, pro forma, Homo sapiens*
Partially Latin	23	2.3	*boli de facto, Dome Quo Vadis, na ad hoc, Cena Pro Urbe*
Slovak	146	14.6	*si iste, idem si, tu super, mater mala, post ministra*
English	8	0.8	*Super Heroes, Super Series*

Other languages	12	1.2	*Kyrie eleison, in ter no, nece biti*
Personal name	54	5.5	*Victor Hugo, Monte Chirsto, Julius Caesar*
Other proper name	175	17.5	*Canon EOS, PS Vita, AC Sparta, Extra plus*
	1,000	100.0	

Million-Click Dictionary: Tools and Methods for Automatic Dictionary Drafting and Post-Editing

Miloš Jakubíček[1,2], **Vojtěch Kovář**[1,2], **Pavel Rychlý**[1,2]

[1] *Lexical Computing, Brno, Czech Republic*
[2] *Masaryk University, Brno, Czech Republic*

Abstract

In this paper we report on recent findings in automatic dictionary drafting and post-editing based on two ongoing lexicographic projects, an Urdu-English-Korean dictionary and a Lao-English-Korean dictionary. We describe the basic workflow used for automatic dictionary drafting and discuss some associated methodological challenges we were facing together with solutions that we applied.

Keywords: Urdu, Lao, Korean, English, automatic dictionary drafting, post-editing, Sketch Engine, Lexonomy

1. Introduction

Contributions of natural language processing and corpus linguistics have helped lexicographers automating many parts of the dictionary building process. Recent efforts therefore focus on generating a whole dictionary draft automatically, and having it post-edited afterwards by lexicographers, roughly in the same way as translators boost their work with machine translation [3].

In this paper we illustrate this process on the example of two bilingual dictionaries: from Urdu to English and Korean and from Lao to English and Korean. These dictionaries have been drafted fully automatically and later partially post-edited. We describe the structure of the dictionaries, tools and methods used for drafting the entries and discuss management and methodological issues of the workflow we have used.

The dictionaries have been drafted from web corpora we have built and loaded into Sketch Engine [1], a corpus query system with advanced analytic functions that were used to generate the automatic draft. The post-editing phase has been carried out in Lexonomy [2], a lightweight open-source dictionary writing system.

2. Sketch Engine

Sketch Engine is a leading corpus management system hosting several hundreds of corpora for (as of January 2020) over 100 languages. It offers many functionalities useful for lexicographers to carry out different parts of the dictionary building, such as devising a headword list, finding good dictionary examples, generating collocation candidates or thesaurus items. All these functions have been used independently by lexicographers in many dictionary projects. In 2017 a single function combining these features has been presented under the name One-Click Dictionary:[1] it builds a complete dictionary draft and exports it into Lexonomy.

3. Lexonomy

Lexonomy is a cloud-based open-source dictionary writing and online dictionary publishing system (see more in [3]) which is highly scalable and can adapt to large dictionary projects as well as small lexicographic works such as editing and online publishing of domain-specific glossaries, wordlists or terminology resources. Lexonomy allows editing from scratch but also accepts automatically generated dictionary drafts. Lexonomy is designed to interact with Sketch Engine and its corpora in two ways: accepting genenrated content ("push model") or interrogating corpora and retrieving different types of results ("pull model").

Push model

The push model refers to the initial dictionary draft generation. The process starts in Sketch Engine and requires that the user selects the corpus that should be used as the source data for the dictionary. Then the user decides how the dictionary headword list should be generated. Whether the dictionary headwords should be selected based on frequency using the wordlist tool or whether the headwords

[1] https://www.youtube.com/watch?v=TaC8sTFWkqs

should be selected from the terminology contained in the corpus using the Keywords & Terms tool. Then the user configures which parts of the dictionary entry should be generated (collocations, example sentences, synonyms, frequency information etc.). Sketch Engine then analyses the corpus and generates the required number of dictionary headwords with the required content and pushes, or exports, the automatically generated dictionary draft into Lexonomy where it is ready for further editing and for publication online.

Pull model

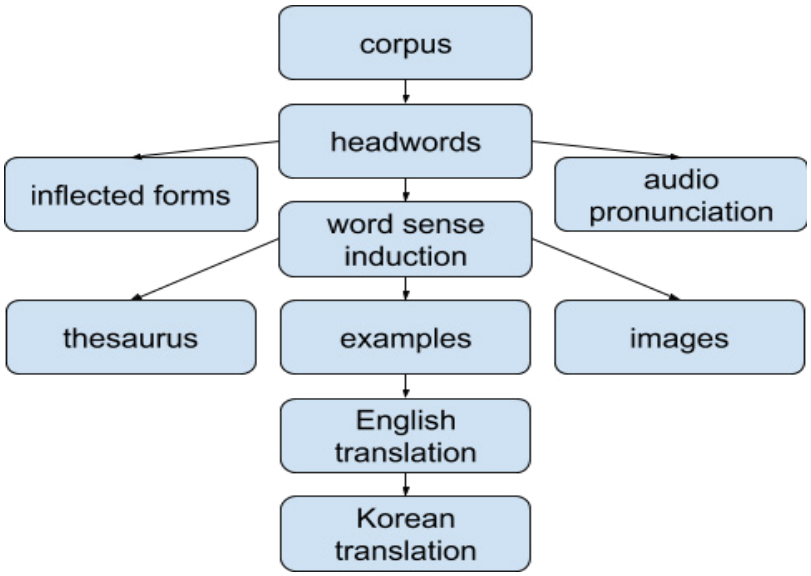
The pull model is associated with the process of post-editing the dictionary draft in Lexonomy. When the user, the dictionary editor, works with the automatically generated content, it might become necessary to check the source corpus or it might be necessary to generate additional information for the dictionary entry, for example, more collocations might be needed or different example sentences might be required. This is when the pull model comes in. Lexonomy is designed to communicate with Sketch Engine. A dictionary in Lexonomy can be linked to a specific corpus in Sketch Engine so that additional data can be pulled from the corpus if needed.

4. From One-Click Dictionary to Million-Click Dictionary

In the One-Click Dictionary approach, the whole dictionary draft is generated all at once. While that is useful in cases where the draft is not going to be post-edited, in the opposite case the post-editing can much more efficient if carried out step-by-step, so that errors in the automatic generation do not propagate. In this paper we argue that such a step-by-step post-editing of individual entry parts is much more time-efficient but also creates new technological and managerial issues rising from the

Figure 1: Post-editing workflow.

repetitive back-and-forth between the post-editing phase and new content generation from the corpus.



The basic workflow is described in Figure 1. Each step assumes that its parent task has been completed, clearly some of the tasks can be edited in parallel or split into a large number of batches. Key issues that we address in the paper is how to ensure data consistency and transparent backing of the underlying corpus evidence throughout the whole post-editing procedure. The reason for this is that any of the post-editing steps may result into revisions of the entry at different levels or even of the corpus material, in cases where the editor challenges the automatic corpus annotations such as part-of-speech tagging or lemmatization.

In this paper we describe our efforts on automating the management of the post-editing phase so that it would not require manual intervention between the individual post-editing tasks. We also discuss the overall efficiency of the process, based on the two dictionary projects in Urdu and Lao, each comprising 45,000 entries, out of which 15,000 have been manually post-edited.

References

[1] KILGARRIFF, Adam, et al. The Sketch Engine: ten years on. Lexicography, 2014, 1.1: 7-36.
[2] MĚCHURA, Michal. Introducing Lexonomy: an open-source dictionary writing and publishing system. In: Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference. 2017. p. 19-21.
[3] JAKUBICEK, Milos, et al. Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. In: The XVIII EURALEX International Congress. p. 65.

A sentiment lexicon for Albanian

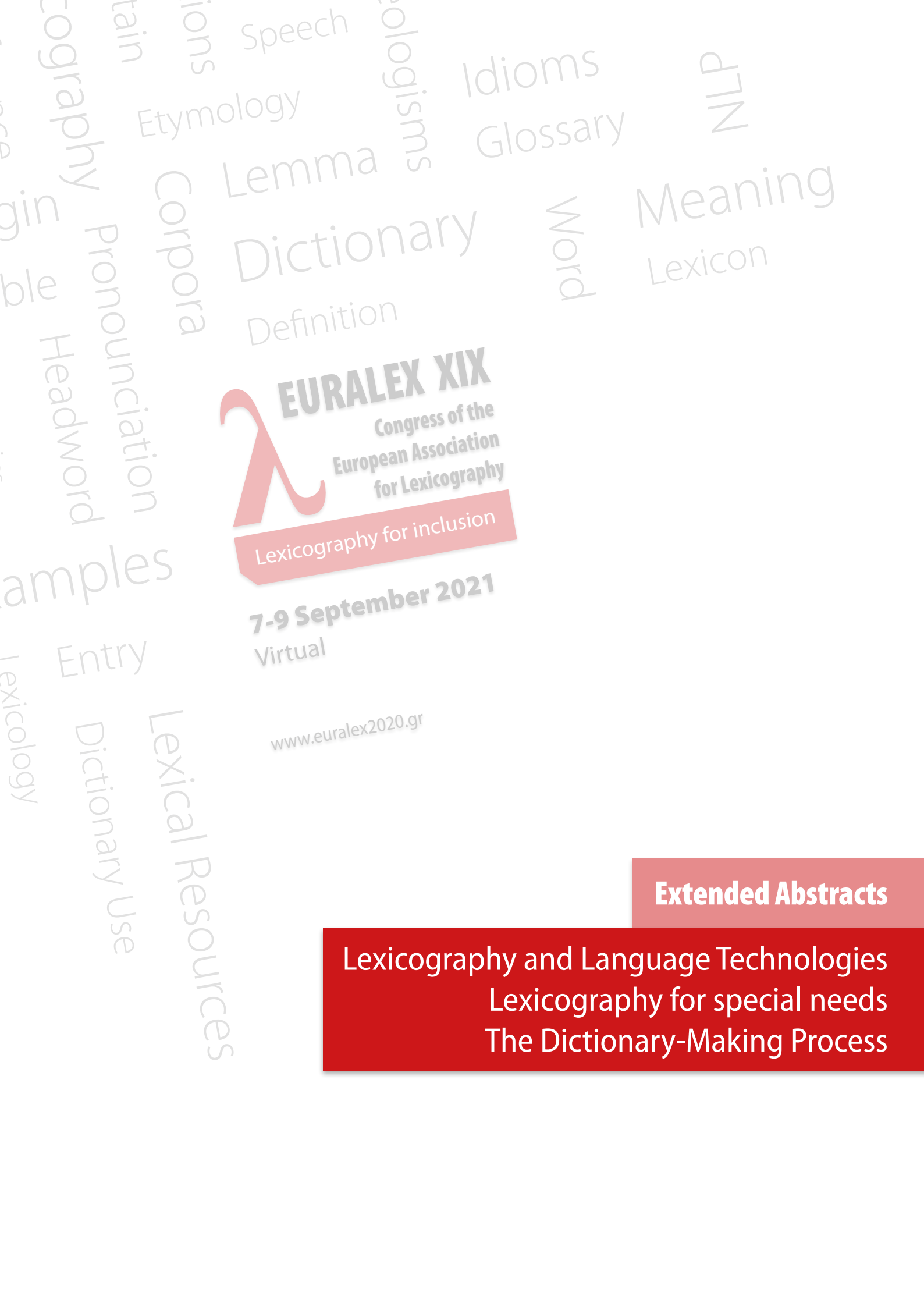
Besim Kabashi[1,2]

[1] *Corpus and Computational Linguistics, University of Erlangen-Nuremberg, Germany*
[2] *Albanology, University of Munich, Germany*

Abstract

Sentiment analysis or opinion mining is needed in many areas of the Natural Language Processing. A sentiment lexicon usually serves as basic knowledge resource for those tasks.
For creating the sentiment lexicon for the standard Albanian language, we used a list of lexical entries of an Albanian lexicon for natural language processing, first presented at EURALEX 2018 (Kabashi, 2018). Treated from another point of view, we extended the information of the mentioned lexicon with the sentiment knowledge. We did this for a selected amount of lexical entries, for ca. 10.000 of them.
As initial step the first 10.000 entries of the frequent lemmatized word list, extracted from the AlCo (Kabashi, 2017), an Albanian language corpus, are selected. The sentiment information, as a combination of polarity and intensity, was added manually to the lexical entries, i.e. expert way. Parallelly, for the lexical entries the sentiment information is generated by using data driven methods of natural language processing applied on the AlCo data, i.e. the machine learning way. In the first method, the sentiment information contains one polarity value as positive, negative or neutral and one of the scores strong or weak, i.e. -2, -1, 0, 1 and 2. At the moment we are looking forward to having another one or two expert(s) for the manual annotating of the lexical entries. This would help to improve the quality of the data annotated manually by the experts. In the machine learning way we generate new sentiment values from time to time after adding new texts to the corpus. The lexicon created automatically by the machine learning methods helps us in several cases to review entries annotated manually.

Keywords: Albanian Language, Sentiment Lexicon, Sentiment Analysis, Machine Learning, Natural Language Processing



Extended Abstracts

Lexicography and Language Technologies
Lexicography for special needs
The Dictionary-Making Process

Graded scientific definitions for school learners: A challenge for pedagogical specialized lexicography

Maria Mitsiaki[1], **Ioannis Lefkos**[2]

[1] *Department of Greek Philology, Democritus University of Thrace, Komotini, Greece*
[2] *Department of Educational & Social Policy, University of Macedonia, Thessaloniki, Greece*

Abstract

Lexicographic theory and empirical research has extensively focused on the structure, content, types, and limitations of dictionary definitions (a.o. Zgusta, 1971; Landau, 2001; Wierzbicka, 1985; Hanks, 1987; Mel'čuk, 1988; Laufer, 1993; Atkins & Rundell, 2008; Polguère, 2012; Sałaciak, 2012; Mel'čuk & Polguère, 2018). All theoretical postulates and empirical findings converge to a single point: definition-writing is the most difficult task at the entry-building stage and one of the most contentious aspects in compiling a dictionary. It is no wonder that good lexicographers are still being judged by their definition-writing abilities (Fontanelle, 2008).

The major difficulties in the definition-writing process are traced in deciding about content and form (Atkins & Rundell, 2008). Firstly, a tacit consent shared by the members of the lexicographic community is that definitions are static "one-size-fits-all" entities, i.e. only one is given per sense of a lemma, urging lexicographers to agonize over finding the ones that are intended to please everyone (Nielsen, 2011). Secondly, lexicographers have to select between several defining conventions and formats with unavoidable pitfalls. For instance, the unchallenged classical Aristotelian definition type, extremely famous in specialized lexicography, may suffer from direct or indirect circularity and obscurity (Sałaciak, 2012), which can be partially solved by adhering to a controlled defining vocabulary or resorting to alternative definitional formats, such as full-sentence definitions (Rundell 2006), extensional definitions (Geeraerts, 2003), and single-clause *when*-definitions (Lew & Dziemianko, 2012). This is the case especially for pedagogical lexicography and for monolingual dictionaries addressed to non-native speakers. It goes without saying that such alternative defining models bear their own disadvantages, i.e. length, reading load, overspecification, reduced precision, etc. (Rundell, 2006).

However, distinctions are blurred when it comes for definition-writing in the frame of pedagogical specialized lexicography. Despite the numerous studies examining definition wordings adapted to suit the specific needs of specific users, not much has been done to shed light on the structure of term definitions targeted to young learners who are engaged in the school scientific disciplines. In this case, the so-believed clear-cut distinction *genus-differentiae for terms with hierarchical organization* vs. *alternative defining models for everyday lexical items* (to promote pedagogical value and user-friendliness) is questionable, since pedagogically-aimed specialized dictionaries for school subjects can and should enter the school community at an early educational level. In fact, this linkage is even more imperative from a functional view (Halliday, 2004), since scientific disciplines are abundant in interlocking definitions, i.e. 'networks' where the definition of a concept is intertwined or presupposes that of another concept, leading to a strikingly complex semantic structure.

From the previous discussion it becomes evident that the most effective solution to handle with the inherent shortcomings of the definition enterprise is to establish a sound relationship between the definition and the needs and skills of its potential users (Atkins & Rundell, 2008). With all the benefits and promises of the electronic medium, multifunctional and multimodal definition models can arise, being targeted at users with differentiated needs. Within this perspective, a multifunctional dictionary may have one definition type for each function and use (Nielsen 2011).

It is this gap in the literature that this paper is intended to fill by laying the emphasis on the definition-writing methodology followed in ELEFYS, a web-based Greek Illustrated Science Dictionary for School (Mitsiaki & Lefkos, 2018) with multiple definitions of graded difficulty, addressing native and non-native primary school learners (www.elefys.gr). As a joint effort and a product of interdisciplinary collaboration between experts in the areas of applied linguistics and science education, ELEFYS aims at bridging the gap between pedagogical and specialized lexicography. Our

scope is two-fold: (a) to record, analyze, and compare the defining techniques and schemata of fundamental Science (Physics) concepts, i.e. *force, energy, current,* etc. as they appear in pedagogical material, i.e. primary and secondary school textbook corpora and (print/electronic) dictionaries for school, and general-use electronic dictionaries, so as to measure their linguistic and conceptual/cognitive load, and (b) to propose a model of graded definitions for scientific terms that complies with the learner's productive and receptive needs, but also gradually develops scientific academic literacy.

The methodological steps followed are outlined below.

Indexing and classification. Definitions of 20 science terms were searched for in school textbook corpora and 15 pedagogical specialized (e.g. *Oxford Primary Illustrated Science Dictionary*, 2013) or general-use dictionaries (e.g. *Dictionary of Standard Modern Greek*, 1998). The definitions were recorded and classified by their format.

Functional linguistic analysis. The indexed and classified definitions were analyzed as to the characteristics of scientific language they exhibit, proposed by Halliday (2004) and Anastassiadis-Symeonidis et al. (2014): (1) interlocking nature, (2) technical taxonomies, (3) special expressions, (4) lexical density, (5) syntactic ambiguity (passive constructions, extensive subordination, etc.), (6) nominalization, and (7) semantic discontinuity. These data analyzed according to these features were quantified, so as to capture tendencies.

Model construction. A graded defining model was devised. According to this model, gradation is achieved both in linguistic and conceptual terms, since definitions are ranked in a continuum of difficulty:

1. Level-A: contextual and functional key definitions, simplified language, avoidance of circularity/ obscurity, high pedagogical value, reduced scientific accuracy.
2. Level-B: use of low level academic wording and vocabulary, simple syntactic structure, limited term interlocking.
3. Level-C: academic language, interconnectedness of scientific concepts, higher level language structure and conceptual density.

Such a graded defining model follows a ranking from the simplest definition (suggested for a basic observation/understanding of the phenomenon) to the most complex (with academic wording). Gradation serves the different needs of learners in accordance with their age, cognitive state and linguistic competence in L1 and L2 and is flagged by a key image for contextual definitions and by one to three stars for conventional definitions in academic wording, so that difficulty in content or form can be marked. Finally, it scaffolds the integrated development of students' scientific (Osborne, 2002) and academic literacy (Bailey, 2007), and can prove to be a useful tool for CLIL and content-based approaches.

Keywords: graded definitions, Greek Illustrated Science Dictionary for School, pedagogical specialized lexicography

References

Adamska-Sałaciak, A. (2012). Dictionary definitions: problems and solutions. Studia Linguistica Universitatis Iagellonicae Cracoviensis, 129 (4), 323-339.

Anastassiadis-Symeonidis, A., Vletsi, E., Mitsiaki, M., Oikonomou, S., & Aleksandri, K. (2014). Seeing the world from a different perspective: light, colours. A Science and Language teaching approach to foreign students [in Greek]. In A. Psaltou−Joycey, E. Agathopoulou, & M. Mattheoudaki (Eds.), Proceedings of the 15th International Conference of Applied Linguistics, "Cross-curricular Approaches to Language Education." Thessaloniki: Greek Applied Linguistics Association, 244-272.

Atkins, B. S., & Rundell, M. (2008). The Oxford guide to practical lexicography. New York: OUP

Bailey, A. (2007). The Language Demands of School: Putting Academic English to the Test. Yale University Press, New Haven, CT.

Dictionary of Standard Modern Greek. (1998). Institute of Modern Greek Studies, Manolis Triantaphyllidis Foundation. [in Greek] Available at
http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/

Geeraerts, D. (2003). Meaning and Definition. In Van Sterkenburg, P. (Ed.), A Practical Guide to Lexicography. Amsterdam: John Benjamins, 83-93.

Halliday, M. A. K. (2004). The Language of Science. Continuum, London.

Hanks, P. (1987). Definitions and explanations. In Sinclair, J.M. (ed.), Looking Up. London: Collins, 116–136.
Landau, S. I. 2001. Dictionaries. The Art and Craft of Lexicography. Cambridge: Cambridge University Press.
Laufer, B. (1993). The effect of dictionary definitions and examples on the use and comprehension of new L2 words. Cahiers de Lexicologie 63:2, 131-142.
Lew, R., & Dziemianko, A. (2012). Single-clause when-definitions: Take three. In R. V. Fjeld & J. M. Torjusen (Eds.), Proceedings of the 15th EURALEX International Congress. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 997–1002.
Mel'čuk, I., & Polguère, A. (2018). Theory and Practice of Lexicographic Definition. Journal of Cognitive Science 19 (4), 439-492.
Mel'čuk, I. 1988. Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria. International Journal of Lexicography 1(3), 165–188.
Mitsiaki, M. & Lefkos, I. (2018). ELeFyS: A Greek Illustrated Science Dictionary for School. Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 373-385.
Nielsen, S. 2011. Function- and user-related definitions in online dictionaries. In Kartashkova, F. I. (ed.), Ivanovskaya leksikografi scheskaya shkola: traditsii i innovatsii [Ivanovo School of Lexicography: Traditions and Innovations]: A Festschrift in Honour of Professor Olga Karpova. Ivanovo: Ivanovo State University, 197-219.
Osborne, J. (2002). Science Without Literacy: A ship without a sail? Cambridge Journal of Education, 32 (2), 203-2018.
Oxford Primary Illustrated Science Dictionary (2013). Oxford University Press.
Polguère, A. 2012. Propriétés sémantiques et combinatoires des quasi-prédicats sémantiques. Scolia 26, 131–152.
Rundell, M. (2006). More Than One Way to Skin a Cat: Why Full-Sentence Definitions Have Not Been Universally Adopted. In Corino, E., Marello, C. and Onesti, C. (eds.), Atti Del XII Congresso Di Lessicografia, Torino, 6-9 Settembre 2006. Allessandria: Edizioni dell'Orso, 323-337.
Wierzbicka, A. 1985. Lexicography and Conceptual Analysis. Ann Arbor, MI: Karoma.
Zgusta, L. (1971). Manual of Lexicography. (Janua Linguarum. Series Maior, 39.) The Hague: Mouton.

App testing for dictionary design

Valeria Caruso[1], Roberta Presta[2], Flavia De Simone[2], Bich Ngoc Pham[3]

[1] Department of Literary, Linguistic and Cultural Studies, University of Naples 'L'Orientale', Naples, Italy
[2] Scienza Nuova Research Centre, University Suor Orsola Benincasa, Naples, Italy
[3] Department of Italian Studies, Hanoi University, Hanoi, Vietnam

Abstract

Structuring the design space and optimizing data visualization should be a key concern when dictionaries migrate to mobile applications. Small screens and consultation 'on the go' should not prevent users from easily accessing exhaustive data to carry out effective searches.

Surprisingly, however, research on the design process of dictionary layout is very limited in number (e.g. Müller-Spitzer et al. 2014; Müller-Spitzer & Koplenig 2014; Storjohan 2018) albeit popular reference works in metalexicography acknowledge design as "the overall principles that govern the production of efficient reference works" (Hartmann & James 1998) in terms of targeting: i) users' skills and needs, ii) features of dictionary contents and iii) the way these contents are presented and arranged. All these features are on focus in an ongoing design project aimed at compiling a dictionary app of Italian idioms for foreign learners (Caruso et al. 2019). This contribution presents the experimental phase of the app development.

The restricted macrostructure of the dictionary, where only idioms are listed, simplifies decisions on the basic dictionary layout because only one type of article is displayed. At the same time a comprehensive description of idioms is provided, because the boiled down semantic explanations given in general language dictionaries are too scant to support foreign speakers, as a preliminary investigation has also shown (Caruso 2016). For this reason, using the OWID Sprichwörterbuch (Steyer & Ďurčo, 2013) as a model, fifteen different data types have been chosen as microstructural items and inflection tables have been added to better support prospective users in production tasks.

With the aim in mind of developing a user-friendly tool for learning, the app design started as a joint collaboration between Lexicographers and Human Factors specialists (HF), who are experts in charge of design processes "concerned with ways in which both hardware and software components […] can enhance human–system interaction" (ISO 9241-210:2010). These professionals pursue the users' involvement in ideational protocols and plan iterative design cycles before a product is released, as indicated by the Human-centred design for interactive system framework (Cooley 1989), a core approach in the ergonomics field.

In the current project, Lexicographers and HF specialists defined activities for a co-design workshop held with prospective users, aimed at empathizing with them (Platter et al., 2014). Role-playing activities and brainstorming sessions helped in defining goals and needs of all the stakeholders involved -both Lexicographers, HF specialists and Learners- and were used to sketch several dictionary prototypes (Caruso et al. 2019) eligible for testing with real users.

In the current phase of the project, two of these prototypes, characterized by opposing features, were tested with prospective users. One, called prototype 1 (P1, see fig.1), has a straightforward layout listing all items in different search zones (Wiegand et al., 2013) of the same view and displays labels of the data types provided. Users are therefore faced with a rather traditional, scrollable access to lexicographic information. The second prototype (P2, fig.2) has a composite layout structure displaying basic information on top of the page (i.e. meaning, one usage example and frequent lexical substitutes), while other information is accessible by clicking on the labels of the corresponding datatypes, which are listed as separated rows of a table view. Additional indicators group together the items that are better suited for reception and production tasks: *Informazioni per la comprensione* ('Information for understanding', fig.2) and *Informazioni per la produzione* ('Information for production'). P2 combines i) lexicographers' wish to orienteering users' actions in task completion -as is also suggested by Lexicographic Function Theory, ii) ergonomics expectations concerning the users' ability in constructing a mental model of what they need for different activities, and iii) users' expectations about

the type of information they are more interested in, i.e. idiom origins which are displayed on top of the article. All these features considered, P1 is a representative of existing app dictionaries, while P2 is an innovative tool, based on lexicographic and ergonomic assumptions. Both prototypes are compiled with the same lexicographic data.

Fig. 1. Dictionary Prototype 1 (P1) *Fig. 2. Dictionary Prototype 2 (P2)*

Prototypes were tested with 32 Vietnamese students with an intermediate proficiency level in the Italian language (B2 of CEFR). Students were asked to solve fill-in exercises with missing idioms by using the prototypes as their information source.
To assess which tool is better suited for supporting foreign learners, prototypes were tested along different dimensions:

1. learning efficacy, measured in terms of correct answers provided using each prototype;
2. usability, i.e. the ease of use by the target users in a specified learning context, assessed using the System Usability Scale questionnaire (Brooke, 1996);
3. Acceptance, i.e., the degree according to which users believe that using the tool would enhance their learning performance, esteemed with the Davis' Technology Acceptance Model questionnaire (Davis, 1989).

Besides, users were observed during task completion for detecting the data types that were more consulted as well as design flaws causing problems in the interactions or manifestations of frustration.
Students were divided in 4 groups of 8 people each. After 3 minutes of free exploration of the app, each group was assigned a gap-filling exercise: task 1 contained 7 different idioms and other 7 were administered in task 2 group.
Each students' group used both prototypes (P1 and P2) to do the gap-filling exercise with the following distribution:
Students' group A: used P1 for task 1, had a 20 minutes break and then used P2 for task 2;
Students' group B: used P1 for task 2, had a 20 minutes break and then used P2 for task 1;

Students' group C: used P2 for task 1, had a 20 minutes break and then used P1 for task 2;
Students' group D: used P2 for task 2, had a 20 minutes break and then used P1 for task 1.
The experimental design is suited to minimize any order effects in using the two prototypes and to prevent fatigue and task disengagement (Hopstaken et al., 2015) on behalf of the students.
Results from the testing sessions will be discussed in detail together with improved design solutions of the idiom dictionary app.
Keywords: Dictionary design; Dictionary apps; Prototyping; Phraseology; Research on Dictionary Use, Human-Computer Interaction; Reports on Lexicographical and Lexicological Projects; The Dictionary-Making Process.

References:

Brooke, J. 1996. "SUS: a 'quick and dirty' usability scale". In Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland, A. L. (eds.). Usability Evaluation in Industry. London: Taylor and Francis.

Caruso, V., 2016, "Dizionari elettronici e apprendimento delle espressioni idiomatiche: monitoraggio dei bisogni e prospettive future", in Bianchi, F. & Leone, P., (a cura di), *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*, Milano: Studi AItLA, pp. 173-189 (ISBN 978-88-97657-12-5, ISBN edizione digitale: 978-88-97657-13-2).

Caruso, V., Balbi, B., Monti, J., Presta, R., 2019, "How Can App Design Improve Lexicographic Outcomes? Examples from an Italian Idiom Dictionary", in Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*. 1-3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ.

Cooley, Mike (1989). "Human-centred Systems". Designing Human-centred Technology. The Springer Series on Artificial Intelligence and Society, 133–143.

Fred D. Davis, 1989. "Perceived usefulness, perceived ease of use, and user acceptance of information technology". MIS quarterly: 319-340.

Hartmann, R.R.K., Gregory, J. 1998. *Dictionary of Lexicography*. London: Routledge.

Hopstaken, J.F., van der Linden, D., Bakker, A.B., Kompier, M.A.J. 2015. "A multifaceted investigation of the link between mental fatigue and task disengagement". Psychophysiol, 52: 305-315.

Müller-Spitzer, C., Koplenig, A., 2014. „Questions of design". In Müller-Spitzer, C. (Hrsg.): *Using Online Dictionaries*. Berlin/ Boston: de Gruyter, Lexicographica: Series Maior, 189–204.

Müller-Spitzer, C., Michaelis, F., Koplenig, A. 2014. „Evaluation of a new web design for the dictionary portal OWID". In Müller-Spitzer, C. (Hrsg.): *Using Online Dictionaries*. Berlin/ Boston: de Gruyter, Lexicographica: Series Maior, 207-228.

Steyer, K. & Ďurčo, P. 2013. „Ein korpusbasiertes Beschreibungsmodell für die elektronische Sprichwortlexikografie". In Banayoun, J. M., Kübler, N., Zouogbo, J. P. (eds.): *Parémiologie, Proverbes et formes voisines*. Sainte Gemme, Band 3, 219-250.

Storjohann, P. 2018. "Commonly Confused Words in Contrastive and Dynamic Dictionary Entries". In Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (Esd.): Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts. Ljubljana: Ljubljana University Press, 187-197.

Wiegand, H. E., Beer, S., Gouws, R. H. 2013. "Textual structures in printed dictionaries: An overview". In: Gouws, R. H., Heid, U., Schweickard, W., Wiegand, H. E. (eds.): Dictionaries. An International Encyclopedia of Lexicography: Supplementary volume: Recent Developments with Special Focus on Computational Lexicography. Berlin: de Gruyter, 31–73.

Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian

Iztok Kosem[1,2], Simon Krek[2], Polona Gantar[1]

[1] *Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*
[2] *Jožef Stefan Institute, Ljubljana, Slovenia*

Abstract

Finding a way to manage large interconnected datasets, and not viewing different lexicographic resources as isolated units, has become a very important challenge in modern lexicography. Namely, one quickly faces considerable duplication of work if projects with different lexicographic focus, such as general dictionaries, collocations dictionaries, thesauri and so forth, are compiled completely separately, by different teams, even if several parts of microstructure are shared. Of course, target users can be different, requiring different definitions, examples etc.; however, this should not be used as an excuse why a common database of integrated resources has not been considered.

There is already evidence about the topicality of such approach to data modelling across Europe, as several institutions have presented plans of data consolidation, for example for Estonian (Tavast et al. 2018), German (Geyken 2019), Polish (Żmigrodzki 2018), and Dutch (Colman 2016). Large databases of a variety of data on a language have also been a topic at a workshop on the Future of Academic Lexicography, held in Leiden in November 2019. A further boost to this trend will be provided by tools produced as part of the ELEXIS infrastructure, which will enable linking of lexicographic resources intralingually and cross-lingually.

In this paper, we present the case of the Digital Dictionary Database for Slovenian, which aims to become a one-for-all database for the Slovenian language, to be used for both in the compilation of language resources and natural language processing tasks. The plans for the database have been described in detail in Klemenc et al. (2017). At the time of writing, the database contained data from two resources: Slovene morphological lexicon and the Collocations Dictionary of Modern Slovene. But with several other projects ongoing, the database modelling is continuous and the model is therefore regularly being updated to cater for different types of data, such as synonyms (from the Thesaurus of Modern Slovene), translations (from the bilingual dictionaries, currently Slovenian-Hungarian dictionary), and corpus examples (authentic or modified ones). Consequently, we can present only current status of the model, knowing that it will soon undergo more changes. First, we will look at the database model, focussing on the most challenging concepts, for example lexeme and lexical unit, and connections between different constituent parts of the model. Relatedly, the schema used for exporting data from the database into external tools such as dictionary writing systems is examined; several decisions that ensure compatibility between data models of different lexicographic resources had to be made, and we will discuss both advantages and shortcomings.

The second part of the paper looks in more detail at the ontology of semantic types that we devised for our lemmas, and also collocations; we believe that we will later be able to apply these semantic types to multiword units and patterns (or different arguments in patterns). Semantic types are closely related to our database model, as they represent a type of information we will record to enable linking our data with other languages. We have tested different ontologies, for example Wordnet, Corpus Pattern Analysis (CPA), Lexiconet, Framenet, Estonian ontology, and Simple-Clips ontology, and established that none of them completely met our purpose. For example, Wordnet lexicographer files were too broad, while CPA categories and especially Lexiconet and Framenet categories were in places too fine-grained and/or too overlapping. Still, it did not make sense to start completely from scratch, so we took Wordnet categories and tried to divide them to the extent where each subcategory had enough

representatives to legitimize its inclusion. At the time of writing, the ontology is still being finalised, but it is known that there are 19 top-level categories which are almost overlapping with Wordnet categories – this is not a coincidence as it was our aim to make linking our database with databases of other languages as straightforward as possible. More significant changes can be observed at subcategory levels, where in many categories one can find many similarities with Lexiconet, but without the repetitiveness of possible categories (e.g. Lexiconet groups substances by source, by function and by natural state, whereas our categorization follows a more source-function division).
Further plans include analysing the relationship between our semantic types and semantic concepts of words, in order to determine whether direct mapping is possible, or to what extent. We also plan to examine to what degree do methods such as word sense induction succeed in grouping collocations that belong to the same semantic type. We will also speak briefly about the potential of semantic types for dictionary users, as such information, which might not even be made visible, can facilitate many useful searches in the dictionary, without necessitating any changes to the key parts of the dictionary microstructure.
Bringing it all together in the end, we aim to demonstrate how all this, i.e. semantic types and interconnected databases, can be efficiently used to improve everyday lexicographic workflow, and can facilitate the compilation of new resources, and linking with other existing foreign resources.

Keywords: digital dictionary database, Slovenian, collocations, semantic types, semantic concepts

References

Arhar Holdt, Špela; Čibej, Jaka; Dobrovoljc, Kaja; Gantar, Apolonija; Gorjanc, Vojko; Klemenc, Bojan; Kosem, Iztok; Krek, Simon; Laskowski, Cyprian; Robnik Šikonja, Marko. (2018). Thesaurus of Modern Slovene: By the Community for the Community. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 401-410. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Colman, Lut. (2016). Sustainable lexicography: where to go from here with the ANW (Algemeen Nederlands Woordenboek, an online general language dictionary of contemporary Dutch)? *International Journal of Lexicography*, 29/2, pp. 139-155.

Geyken, Alexander. (2019). The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. Iztok Kosem & Tanara Zingano Kuhn (eds.) *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography. Book of abstracts*. Lexical Computing CZ s.r.o., Brno, Czech Republic.

Klemenc, Bojan; Robnik-Šikonja, Marko; Fürst, Luka; Bohak, Ciril; Krek, Simon. (2017). Technological Design of a State-of-the-art Digital Dictionary. Gorjanc, Vojko, Gantar, Polona, Kosem, Iztok, Krek, Simon (eds). *Dictionary of modern Slovene: problems and solutions.* 1st ed. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 10-22.

Kosem, Iztok; Krek, Simon; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka; Laskowski, Cyprian Adam. (2018). Collocations dictionary of modern Slovene. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 989-997. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 749-761.

Tiberius, Carole; Niestadt, Jan. (2010). The ANW: an online Dutch Dictionary. Anne Dykstra & Tanneke Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Ljouwert: Fryske Akademy/Afûk, 181 (abstract).

Żmigrodzki, Piotr. (2018). Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish. Jaka Čibej, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 209-219. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/download/118/211/2973-1?inline=1

ELEXIS tools for lexicographers (demo)

Iztok Kosem

Jožef Stefan Institute, Ljubljana, Slovenia

Abstract

In this demo, we present a variety of tools and workflows that have been developed in the European Lexicographic Infrastructure (ELEXIS), a Horizon 2020 project. The project has now entered its second half and infrastructure is slowly being set up so it is important that the community gets familiar with its parts, and gets an opportunity to include their own data into various services, and test the tools developed.

The tools include mobile apps such as Game of words and CrossTheWord, which are games with a purpose and support the integration of crowdsourcing techniques into lexicographic workflow. In Games of words, at the time of writing, Dutch, English, Estonian, Slovene and Portuguese are being supported, with more languages expected by the time of the conference. All games use collocational data as input, and partners have prepared datasets that are directly useful for the lexicographic projects. Similarly, CrossTheWord contains lexical data, at the moment from Babelnet, and is designed in a way where the players help clean the data by playing puzzles and selecting synonyms.

In addition to demoing individual tools, we will show their role in the lexicographic workflow. For example, participants will be able to see the way the corpus data gets imported into mobile games, the way it is monitored, and how the crowdsourced data gets exported and imported into a dictionary writing system such as Lexonomy. We believe that such a demo is necessary to show the full potential of crowdsourcing in lexicography, given that until now such attempts have been isolated and separated from the lexicographic process.

Another tool with a different purpose is ELEXIFINDER and its aim is to enable searches through lexicographic bibliography. By the time of the conference, we expect major improvements to the tool, including more content, improved ontology for paper categorization, and a dedicated page with instructions for researchers on how to contribute by uploading individual papers or entire sets.

The participants will be able to test other tools made available by the infrastructure, however some of them are part of other presentations and will be thus given less attention at this particular demo.

Keywords: ELEXIS, tools, games, crowdsourcing, lexicographic bibliography

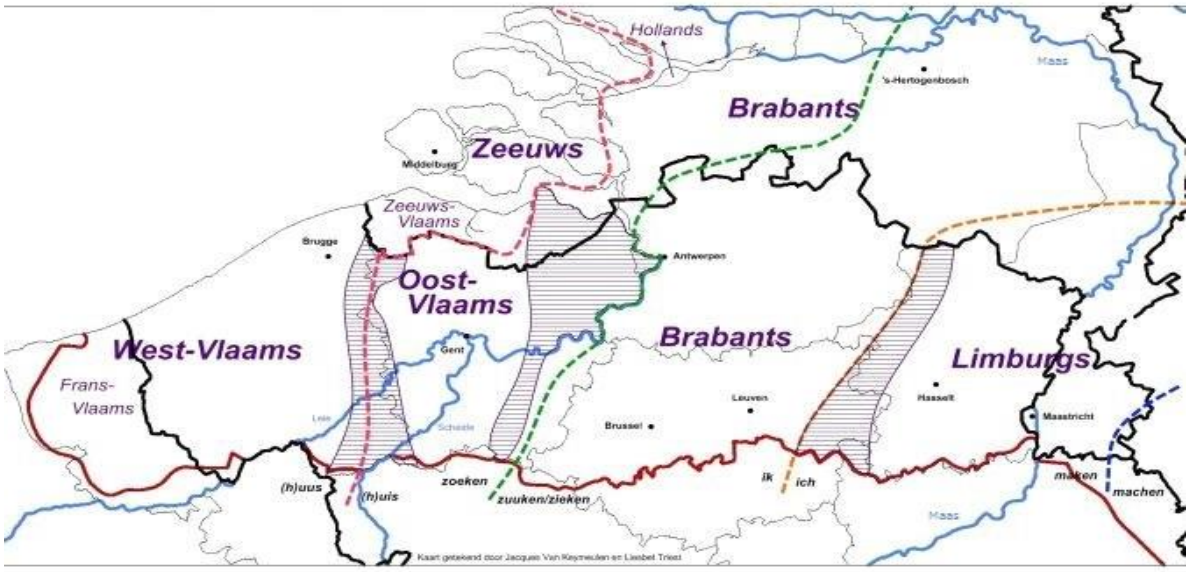
The Dictionary portal of the Southern Dutch Dialects

Veronique De Tier[1], Katrien Depuydt[1], Jesse de Does[1], Tanneke Schoonheim[1], Jacques Van Keymeulen[2], Sally Chambers[2]

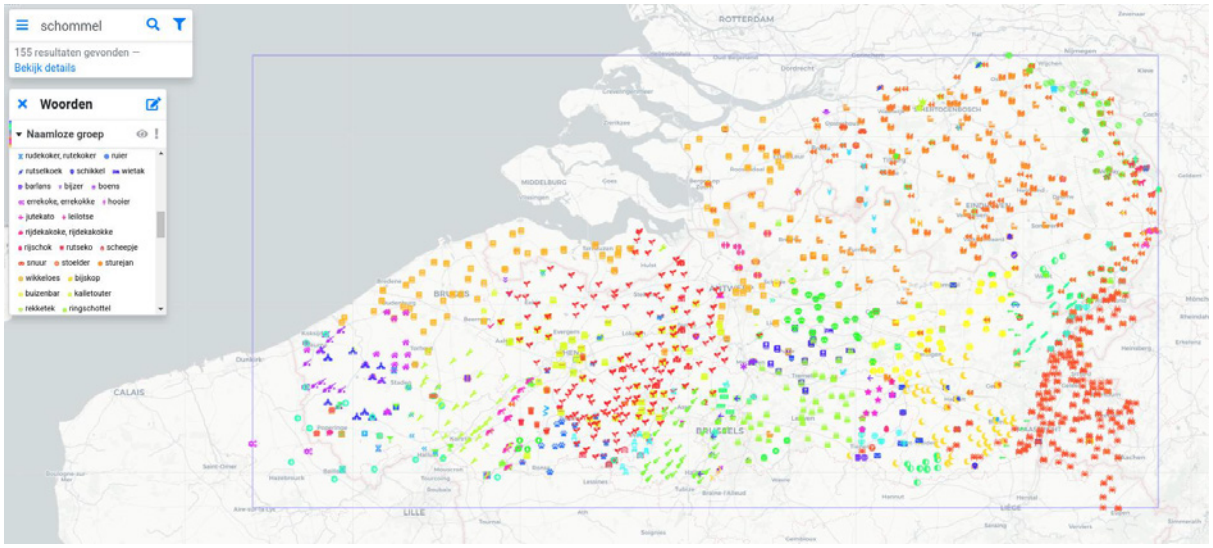
[1] *Instituut voor de Nederlandse Taal (Dutch Language Institute), Leiden.*
[2] *Universiteit Gent (Ghent University).*

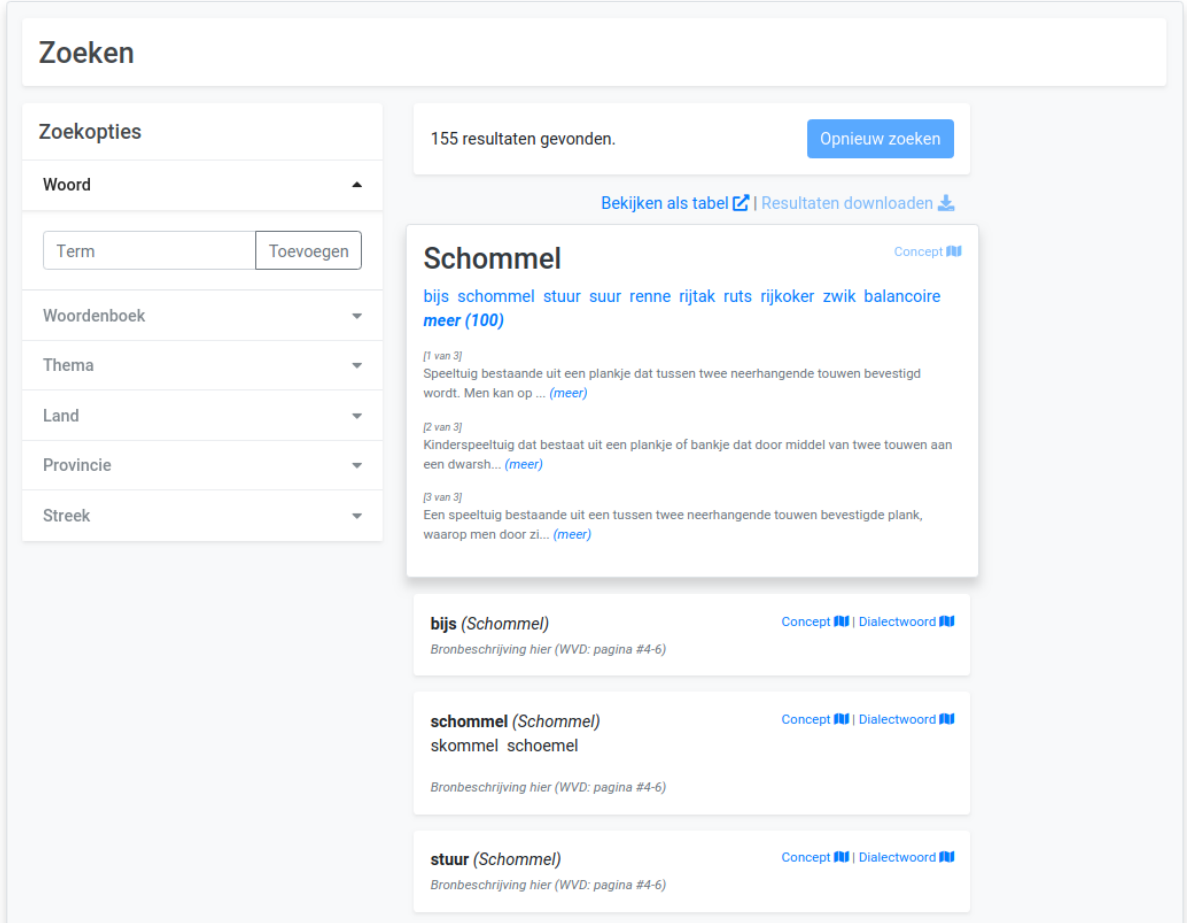
Abstract

The southern Dutch dialect area consists of four dialect groups, partly found in the Netherlands, in Belgium and in a small part of France: the Flemish, Brabantic, Limburgian and the Zeeland dialects. Each of these dialects has been described in a separate dictionary. The Flemish, Brabantic and Limburgian dialect dictionaries are more or less similarly constructed onomasiological dialect dictionaries (thematically arranged), whereas for the Zeeland dialect, it is a semasiological dictionary (alphabetically arranged). In 2016 an infrastructure project started to combine these dictionaries in one dictionary portal: the database of the Southern Dutch Dialects (DSDD). This project was initiated by Ghent University and undertaken in close collaboration with the Dutch Language Institute (Leiden).



The aim of the project was to work with a pilot dataset of 1500 concepts. Principles, strategies and tools for data alignment were developed and a prototype for online application, the dictionary portal. As far as the design of the portal was concerned, inspiration was found in the online versions of the individual dialect dictionaries (e-WVD, e-WBD, e-WLD), in similar projects like Verba Alpina (Switzerland, Italy, Germany) and Regionalsprache.de (REDE - Germany), and in dedicated workshops with partners in the field of geography/cartography. The portal application will be demonstrated here.

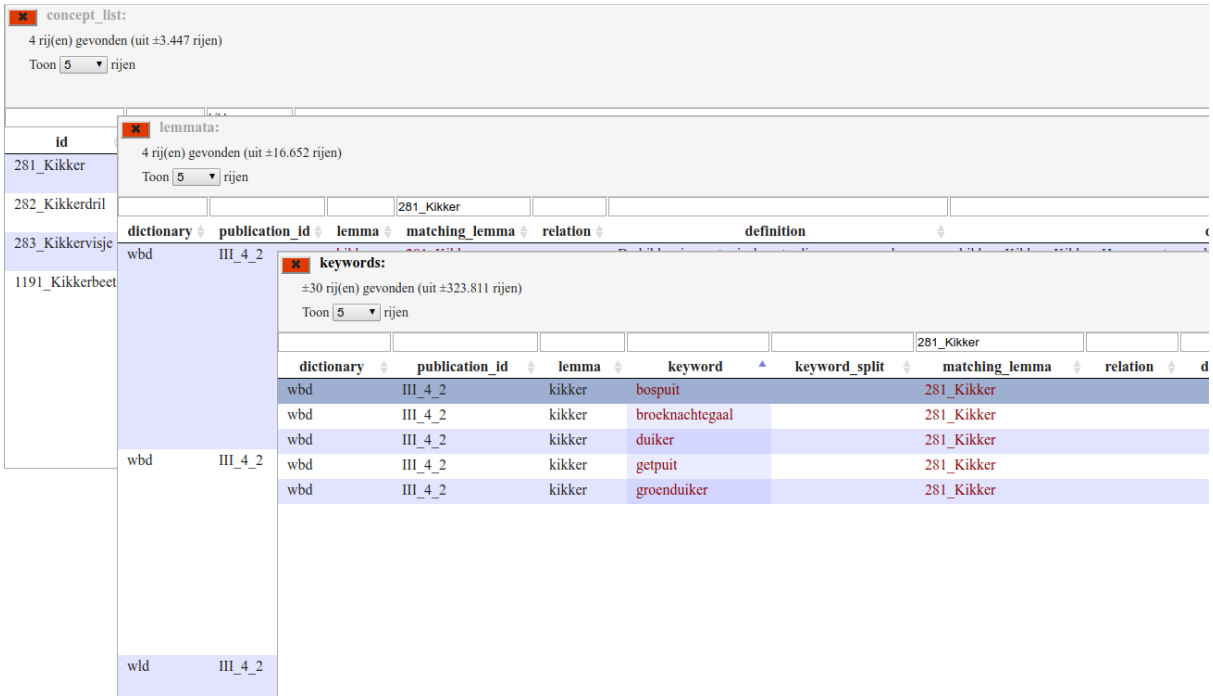




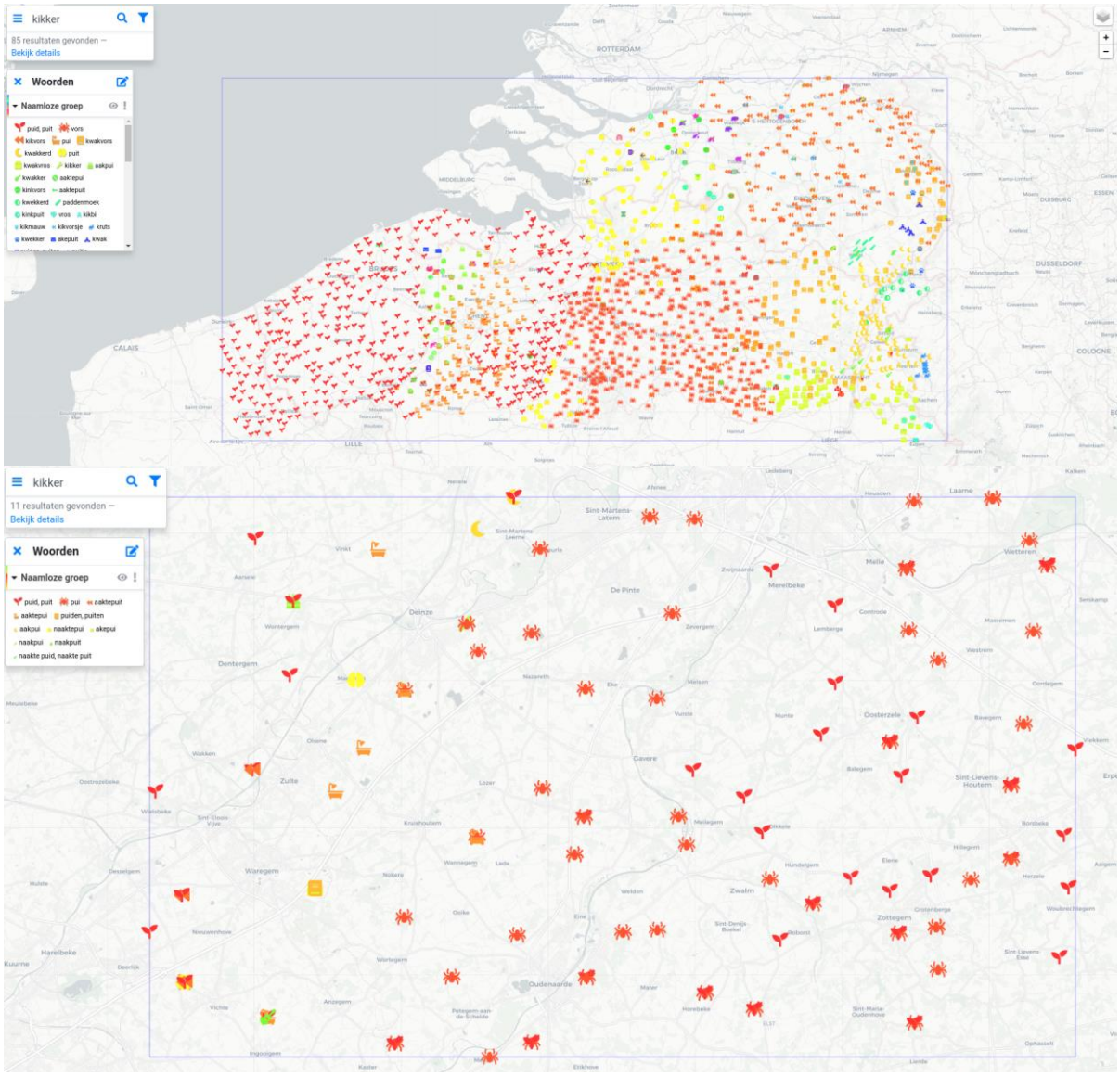
Before being able to start with the data alignment work, some effort had to be put into correcting the original data from the three dictionaries. As often is the case with dictionary data that was not originally conceived for digital use, a number of smaller and also more significant corrections had to be made; for instance, standardising the spelling of the keywords (*paddestoel* / *paddenstoel*) and combining diminutives with their simplex (*schommel* / *schommeltje*).

Concept linking was done by adding an additional DSDD concept layer on top of the concept layer of each individual dictionary. The DSDD concept was determined after careful analysis of the concepts in each dictionary. In many cases the mapping of the concepts was quite straightforward, but in other cases a choice had to be made. For instance *muil* (WVD), *bek* (WBD) and *bek* (WLD) were all put under the

concept *bek* 'animal's mouth'. Sometimes hyponyms and hypernyms had to be combined to achieve a better coverage, for example, *mier* and *zwarte mier* were combined under the concept *mier* 'ant'. The data work was done using the data processing tool Lex'it, developed by the Dutch Language Institute.



The results of the pilot are available in the Dictionary portal of the Southern Dutch Dialects (DSDD). The portal is intended for everyone with an interest in dialects. The portal aims to be accessible for both a wider audience and dialect researchers. Dialect researchers have a lot to gain from the opportunities the portal offers, because of its extensive search options and the availability of an application to create interactive maps with the very precise location information available in the dialect data.
The portal's search engine enables searching for individual dialect words (dialect word *puit,* belonging to the concept *kikker*), for dialect words of a certain village or region (all dialect words from the village Lozer; all dialects from the city of Ghent) and all dialect words belonging to a certain concept (*kikker*) or theme (fauna). Filters are available to make the search query as accurate as possible.
Based on the search results, interactive maps can be compiled, visualising the data in its geographical setting. The results can be grouped in various ways, using for instance frequency or etymologically related entries (for instance a group 'kik' containing *kikker* and *kikvors*, and/or a group 'vors', containing *fros, vors* and *kikvors*). The maps are flexible enabling the user, for example, to select the required level of detail to be visualised for the chosen region. A set of symbols and colours is available and can be arranged according the user's needs. The area map can be enlarged or reduced in size and printed or downloaded.



This portal is the result of a pilot aimed to demonstrate the opportunities of the combined dialect data of the Southern Dutch Dialects (Flemish, Brabantic and Limburgian). The project will be continued at the Dutch Language Institute. Apart from the addition of the remaining data, work will also be done to improve the search functionalities and the data manipulation. New dialect areas will also be added, beginning with the Zeelandic dictionary, which will complete the coverage of the Southern Dutch Dialects area. However, the addition of this final dictionary will be challenging since this is not an ononmasiological dictionary like the three dictionaries already available in the portal but a semasiological dictionary. Once we have developed a strategy to add this dictionary, we intend to add dictionaries from other Dutch dialect areas.

Keywords: Dialects, Dutch, Geo-mapping, Dictionary portal, Data visualisation.

1. Instituut voor de Nederlandse Taal (Dutch Language Institute), Leiden.
2. Universiteit Gent (Ghent University).

Hand in hand or separate ways: Representation of related BODY PART multiword expressions in the microstructure of learners' dictionaries online

Sylwia Wojciechowska

Faculty of English, Adam Mickiewicz University in Poznań, Poznań, Poland

Abstract

The surge of interest in phraseology since the advent of corpus linguistics has not translated into studies of the representation of multiword expressions (henceforth MWEs) in dictionaries, and their status in lexicography still remains unsettled. One reason is lack of consensus concerning the typology of MWEs (Moon 1998: 19-20), another is the privileged status of the (orthographic) word in lexicography (Lew 2012). This results in inconsistent treatment of MWEs not only in printed dictionaries, as observed by Oppentocht and Schutz (2003: 218), but also in present-day monolingual English learners' dictionaries (henceforth MELDs) online, as the current study shows.

The aim of the paper is to investigate the representation of MWEs containing nominal forms of body part names such as *hand*, *head*, *face*, *shoulder*, *eye* and *ear* in the "Big Five" MELDs online. The study focuses on the arrangement of the examined MWEs in the dictionary microstructure and includes not only MWEs defined on the page of the body part headword, but also the ones which are cross-referenced in the form of hyperlinks. MWEs with body parts names have been chosen for the present study because they constitute a fairly homogenous group, related by means of metonymic motivation, and are therefore expected to be treated in a fairly consistent way.

The evaluation of the arrangement of body part MWEs in the entry structure is carried out from the perspective of cognitive linguistics (henceforth CL), and attention is paid to the ways of demonstrating the metonymic motivation of the analysed expressions as well as reflecting the semantic relations which hold between them. Adopting the CL approach in this paper is supported by the fact that increasing numbers of (meta)lexicographers and cognitive linguists recognise the advantages of its application in lexicography, e.g. van der Meer (1999), Moon (2004), Geeraerts (2007), Adamska-Sałaciak (2008), Atkins and Rundell (2008), Kövecses and Csábi (2014), Ostermann (2015), the last of whom in fact coined the term *cognitive lexicography*. Moreover, studies show that idiomatic expressions, are learned and recalled more effectively when their metaphorical underpinnings are explicated (Boers and Lindstromberg 2006).

The list of MWEs for the selected body part headwords was derived from the five dictionaries, and then finalised by choosing only those shared by all the MELDs. The total number analysed and annotated approximates 150 items. The MWEs under scrutiny were tagged in Excel files according to: metonymy type, position on the page of the body part headword, presence or absence of a cross-reference (= hyperlink), nesting or separating related expressions, and presence of entry navigation devices (menus and guidewords).

The results of the study indicate that the arrangement of body part MWEs within the entry structure tends to be dictionary-specific. In four out of five dictionaries (CALD, COBUILD, LDOCE and MEDO), the MWEs are hyperlinked in the overwhelming majority of cases (around 90%) rather than defined under the entry for the body part name. OALD turns out to offer a completely opposite approach, and the prevailing treatment is defining MWEs in an alphabetically arranged *Idioms* box at the bottom of the entry. Each of the "Big Five" offers three types of access routes to body part MWEs with a varying degree of preference, depending on the dictionary: (1) under distinct senses, (2) in dedicated idioms/phrases boxes/sections and (3) via search panels that provide potential target items for the looked-up lexemes. Additionally, in three of the five dictionaries there are a few hyperlinks placed separately at the bottom of the entry without a clear policy behind this treatment.

As for nesting body part MWEs with common metonymic motivation, no consistent approach is found in the MELDs under analysis. Some related expressions are subsumed under one sense and tagged with

a navigation device in the form of the metonymic target, but the majority of them are treated separately, hence the semantic links between them get broken in the entry structure. The former representation is observed for instance in LDOCE, where MWEs with the same underlying metonymic motivation such as *do sth in your head*, *come into/pop into your head* and *get/put something out of your head* are nested together under the sense "your mind or mental ability" with the metonymic target MIND attached to it as the signpost. The latter treatment can be found e.g. in MEDO, where three variant forms of the same expression: *hands off*, *keep your hands off* and *take/get your hands off* receive separate hyperlinks in the alphabetically ordered *Phrases* section which take the user to three differently worded definitions conveying the same meaning. Another case of breaking the semantic link is separating the expressions *give somebody a hand* and *want a hand* from *lend somebody a hand* in COBUILD, with the first two featuring in example sentences under sense 4 ("help" meaning), and the third one detached from them and hyperlinked under sense 36. Other cases of nesting and separating related body part MWEs in the MELD microstructure will be discussed in the full paper.
It should be stressed that none of the "Big Five" emerges as the winner in this comparative analysis, as each of them subsumes some of the related body part MWEs under the same senses, but fails to show the semantic links between the other ones. Solutions are proposed in the framework of CL in order to offer a more homogeneous representation of the examined MWEs in the entry structure.

Keywords: multiword expressions, monolingual English learners' dictionaries online, dictionary microstructure, cognitive linguistics, metonymy, cross-references, hyperlinks, navigation devices

References:

Dictionaries

CALD: *Cambridge Advanced Learner's Dictionary*. Accessed at: http://dictionary.cambridge.org. (25 January 2020)
COBUILD: *COBUILD Advanced English Dictionary*. Accessed at: http://www.collinsdictionary.com. (25 January 2020)
LDOCE: *Longman Dictionary of Contemporary English*. Accessed at: http://www.ldoceonline.com. (25 January 2020)
MEDO: *Macmillan English Dictionary Online*. Accessed at: http://www.macmillandictionary.com. (25 January 2020)
OALD: *Oxford Advanced Learner's Dictionary*. Accessed at: http://www.oxfordlearnersdictionaries.com. (2 September 2019)

Other references

Adamska-Sałaciak, A. (2008). Prepositions in dictionaries for foreign learners: A cognitive linguisitc look. In E. Bernal, J. DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress*, Barcelona: Universitat Pompeu Fabra. CD-ROM, pp. 1477-1485.
Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
Boers, F. & Lindstromberg, S. (2006). Cognitive linguistic applications in second or foreign language instruction: Rationale, proposals, and evaluation. In G. Kristiansen, M. Achard, R. Dirven, F.J. Ruiz de Mendoza Ibáñez (eds.), *Cognitive Linguistics: Current Applications and Future Perspectives*, Berlin: Mouton de Gruyter, pp. 305-355.
Geeraerts, D. (2007). Lexicography. In D. Geeraerts, H. Cuyckens (eds.), *The Oxford Handbook of Cognitive Linguistics*, Oxford: Oxford University Press, pp. 1160-1174.
Kövecses, Z & Csábi, S. (2014). Lexicography and cognitive linguistics. *Revista Española de Lingüística Aplicada*/*Spanish Journal of Applied Linguistics* 27(1), pp. 118-139.
Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger, M. Paquot (eds.), *Electronic Lexicography*, Oxford: Oxford University Press, pp. 343-361.
Meer, G. van der. (1999). Metaphors and dictionaries: The morass of meaning, or how to get two ideas for one. *International Journal of Lexicography* 12(3), pp. 195-208.
Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Oxford University Press.

Moon, R. (2004). On specifying metaphor: An idea and its implementation. *International Journal of Lexicography* 17(2), pp. 195-222.
Oppentocht, L. & Schutz, R. (2003). Developments in electronic dictionary design. In P. van Sterkenburg (ed.), *A Practical Guide to Lexicography*. (Terminology and Lexicography Research and Practice 6.) Amsterdam: John Benjamins, pp. 215-227.
Ostermann, C. (2015). *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. (Lexicographica. Series Maior 149.) Berlin: de Gruyter.

Variation and Semantic Change Automatic Tracking: a Combined Linguistic, Cognitive and Sociolinguistical Approach

Emmanuel Cartier

LIPN - RCLN UMR 7030 CNRS, University Sorbonne Paris Nord, Paris, France

Abstract

Lexical Change and Variation are both essential characteristics of linguistic systems. In this paper, we will present and illustrate three main NLP approaches to track semantic change and variation, as well as to follow the life-cycle of emerging new words and/or meaning. They rely respectively on cognitive, linguistical and sociolinguistical properties of the evolution of form-meaning pairs (Schmid, 2020).

The first and most evident approach consists in tracking the frequency evolution of words through time and through varieties of language. Frequency has long been recognized as a huge signal of exposure and entrenchment of lexical usage (Ellis et al., 2012, for example). We will detail several simple and nevertheless powerful techniques enabling to detect emergence, obsolescence, trends of evolution, and temporal clusters (Koplening, 2018; Kulkarni et al., 2015; Hilpert and Gries, 2016). The model of successful innovation (Rogers, 2003 [1962]) will be critically presented through examples, showing the various paths of evolution (Nevalainen, 2015, 2018; Feltgen et al., 2017).

A more linguistically-grounded approach consists in studying the combinatorial profile (Gries, 2010) of lexemes and its evolution through time. Based on notions like collocation and collostruction (Stefanovistsch et Gries, 2003) , from quantitatively significant corpus, several measures have been proposed to approximate the behavior of lexemes and detect lexical, syntactic or lexico-syntactic change, signaling new meanings.

A second and complementary approach, grounded on the distributional hypothesis that words sharing the most contexts are most semantically similar (Harris, 1954; see Turney et Pantel, 2010 and Baroni et Lenci, 2010 for a computational presentation), enables to follow semantic change through the evolution of the cluster of similar words (i.e. notably synonyms, antonyms, hypernyms, hypernyms, co-hyponyms and meronyms). From the word2vec (Mikolov et al., 2013a and 2013b) initial popular model to the BERT family transformers (Vaswani et al., 2018; Devlin et al. 2018), which have become an essential initial step in most NLP systems, we will show through concrete examples how these models can help detect and follow the evolving linguistic properties of lexemes.

A more sociolinguistical-based approach, notably grounded on the social network structure (Milroy and Milroy, 1985) and the community practice concept (Eckert, 2012), enables to follow the life-cycle of lexical usage through the linguistic communities. Some preliminary experiments have been setup, mainly from online social networks (Eisenstein et al, 2014; Grieve et al., 2018), showing the paths of diffusion by using sociological properties of people and the structure of communication ties (see Nguyen et al., 2016; Clem, 2016 for a review).

Practical examples demonstrating the current state-of-the-art in semantic change automatic tracking will illustrate the methods, their strengths and limits, and avenues for future research and collaboration.

Keywords: Semantic Change, Variation, NLP, Lexical Diffusion, Emergence, Diffusion and Institutionalization

References

Baroni Marco et Lenci Alessandro, 2010, « Distributional memory: A general framework for corpus-based semantics ». Computational Linguistics, 36(4):673-721.

Clem, Emily, 2016, « Social network structure, accommodation, and language change ». UC Berkeley PhonLab Annual Report, 12(1).

Devlin Jacob, Chang Ming-Wei, Lee Kenton et Toutanova Katarina, 2018, « Bert: Pre-training of deep bidirectional transformers for language understanding ». In Proceedings of the 2019 Conference of NAACL-HLT, 2018.

Eckert Penelope, 2012, « Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation ». Annual review of Anthropology, 41:87-100.
Eisenstein, Jacob, O'Connor, Brendan, Smith, Noah A., et al., 2014, « Diffusion of lexical change in social media ». PloS one, 9(11).
Ellis Nick C., 2012, « What can we count in language, and what counts in language acquisition, cognition, and use?" In S. Th. Gries ; D. S. Divjak (Eds.) Frequency effects in language learning and processing (Vol. 1). (pp. 7-34). Berlin: De Gruyter Mouton.
Feltgen Quentin, Fagard Benjamin et Nadal Jean-Pierre, 2017, « Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change ». Royal Society Open Science, The Royal Society, 2017, 170830, 4 (11).
Goldberg, Yoav, et Omer Levy, 2014, « word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method », arXiv preprint arXiv:1402.3722.
Gries Stefan Th., 2012, « Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics ». In Gonia Jarema, Gary Libben, and Chris Westbury (eds.), Methodological and analytic frontiers in lexical research, 57-80. Amsterdam and Philadelphia: John Benjamins.
Grieve, Jack, Andrea Nini, and Diansheng Guo, 2018, « Mapping lexical innovation on American social media ». Journal of English Linguistics 46.4 (2018): 293-319.
Hamilton, William L., Jure Leskovec, et Dan Jurafsky, 2016, « Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change », Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1489–1501, Berlin, Germany, August 7-12, 2016
Hilpert Martin et Gries Stefan T., 2016, « Quantitative approaches to diachronic corpus linguistics ». The Cambridge handbook of English historical linguistics, 36-53.
Jawahar Ganesh et Seddah Djamé, 2019, « Contextualized Diachronic Word Representations ». 1st International Workshop on Computational Approaches to Historical Language Change 2019 (colocated with ACL2019), Aug 2019, Florence, Italy.
Koplenig Alexander, 2018, « Using the parameters of the Zipf-Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes - a large-scale corpus analysis ». Corpus Linguistics and Linguistic Theory, 14(1), 1-34.
Kulkarni Vivek, Al-Rfou Rami, Perozzi Bryan et Skiena Steven, 2015, « Statistically significant detection of linguistic change ». In Proceedings of the 24th International Conférence on World Wide Web, WWW '15, pages 625-635.
Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, et Erik Velldal. 2018. « Diachronic word embeddings and semantic shifts: a survey ».
Lim, Kyungtae, Niko Partanen, et Thierry Poibeau. 2018. « Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian ».
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, et Jeffrey Dean, 2013a, « Distributed Representations of Words and Phrases and their Compositionality », Advances in neural information processing systems, pp. 3111-3119.
Mikolov, Tomas, Kai Chen, Greg Corrado, et Jeffrey Dean, 2013, « Efficient Estimation of Word Representations in Vector Space », arXiv preprint arXiv:1301.3781.
Milroy James et Milroy Lesley, 1985, « Linguistic change, social network and speaker innovation ». Journal of Linguistics, 21(2):339-384.
Nevalainen Terttu, 2015, « Descriptive adequacy of the S-curve model in diachronic studies of language change ». Studies in Variation, Contacts and Change in English 16.
Nevalainen Terttu, Laitinen Mikko, Nevala Minna et al., 2018, « Changes in different stages: From nearing completion to completed ». In T. Nevalainen, M. Palander-Collin, and T. Saily (Eds.), Patterns of Change in 18th-century English: A Sociolinguistic Approach (pp. 251-254).
Nguyen, Dong, Doğruöz, A. Seza, Rosé, Carolyn P. et al., 2016, « Computational sociolinguistics: A survey ». Computational linguistics, 42(3), 537-593.
Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, et Luke Zettlemoyer. 2018. « Deep contextualized word representations », Proceedings of NAACL-HLT 2018, pages 2227–2237, New Orleans, Louisiana, June 1 - 6, 2018.

Rogers Everett M., 2003 [1962], Diffusion of innovations (5th ed.). New York, NY: Free Press.
Ruder, Sebastian, Ivan Vulić, et Anders Søgaard. 2017. « A Survey Of Cross-lingual Word Embedding Models », Journal of Artificial Intelligence Research, 65, 569-631.
Schmid, Hans-Jörg, 2020. The dynamics of the linguistic system. Usage, conventionalization, and entrenchment, Oxford University Press.
Stefanowitsch Anatol et Gries Stefan T., 2003, « Collostructions: Investigating the interaction of words and constructions ». International journal of corpus linguistics, 8(2):209-243.
Tahmasebi Nina, Borin Lars et Jatowt Adam, 2018, « Survey of Computational Approaches to Lexical Semantic Change ». arXiv preprint arXiv:1811.06278.
Turney Peter D. et Pantel Patrick, 2010, « From frequency to meaning: Vector space models of semantics ». Journal of artificial intelligence research, 37, 141-188.
Vaswani Ashish, Shazeer Noam, Parmar Niki et al., 2017, « Attention Is All You Need ». In Advances in Neural Information Processing Systems.
Yao Zijun, Sun Yifan, Ding Weicong et al., 2018, « Dynamic word embeddings for evolving semantic discovery ». WSDM '18, pages 673-681, ACM.

www.euralex2020.gr

Extended Abstracts

Lexicography for special needs

A Glossary to support Deaf access to Scientific Journal Systems

Ronnie Fagundes de Brito[1], Vera Lúcia de Souza e Lima[2], Noriko Lúcia Sabanai[3]

[1] *Coordenação de Articulação, Geração e Aplicação de Tecnologia, Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, Brasil.*
[2] *Departamento de Engenharia Civil, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, Brasil.*
[3] *Secretaria de Estado de Educação do Distrito Federal, Brasília, Brasil*

Abstract

This research presents a proposal for structuring a Bilingual Glossary (Brazilian Sign Language / Brazilian Portuguese) of terminology related to scientific communication. And also, demonstrates its application in the development of an user interface that enables deaf accessibility to Open Journal Systems, a software widely used in the publication of Scientific Journals.

Deaf representativeness in the digital environment is a constant challenge to be overcome by the information society in which we live. Sign language, one of the communication means used by deaf people, have benefited from the ability to communicate via video and also written notations which may be more easily represented in digital media.

In this scenario, the development of information systems that present sign languages as the main mean of communication in their interfaces is a determining factor for the affirmation of Deaf`s linguistic rights.

One of the problematic issues that Deaf people face is the scarcity of terminological lexicon in sign language, in the case of Brazil, Libras (Língua Brasileira de Sinais). According to Lima (2014), the lack of technical terms in different disciplines makes access to knowledge inefficient, discouraging them, for example, from reaching higher education.

In the proposed glossary, a term will be represented in three different ways: (1) Portuguese in written form; (2) Video of the sign in Libras, (3) Libras in written form. Signwriting (SUTTON 2009; BARRETO 2015) will be used for Libras writing. This form of writing is derived from a notation for recording dance movements, and has been adapted for recording movements and gestures of sign languages, becoming increasingly popular with deaf community and used from their literacy learning to scholarly communication.

In the Open Journal Systems interface, illustrated in figure 1, the signs in Libras will be presented using Signwriting, where an interaction model, to be developed, will allow access to the record in the glossary of each term or phrase.

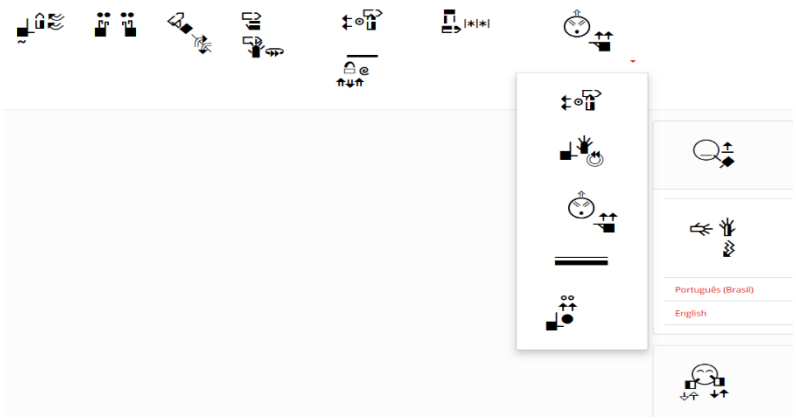


Figure 1: OJS User Interface
Source: http://swojs.ibict.br/ojs/

In addition to terminological development, there is also a challenge regarding the presentation of signs

on the software interface, given that Signwriting occupies more space than the characters of the Latin alphabet.
This study is expected to expand access to scientific and technological information for the Deaf, and also to offer a tool for learning Signwriting.
The research has a multidisciplinary character, proper to Terminology, aiming at the production of glossaries in the scientific, technological, linguistic, artistic and cultural areas. The theoretical assumptions used were related to Lexicology and Terminology: Lima (2014) among others, aiming at the production of Bilingual Terminological Glossaries (Brazilian Portuguese / Brazilian Sign Language). Research in the area of Signwriting: Barreto (2015) and typological functional linguistics, such as: Aikhenvald (2008), Brito (1995, 2003), Dixon (2010), Felipe (2001), Lillo-Martin, D.; Klima (1990), Meir, I. et al. (2006), Quadros (1999, 2003), Quadros; Karnopp, (2001), Schembri (2003), Stokoe (1960) etc.
Finally, it is understood that Sign Language is the main Assistive Technology that Deaf need.

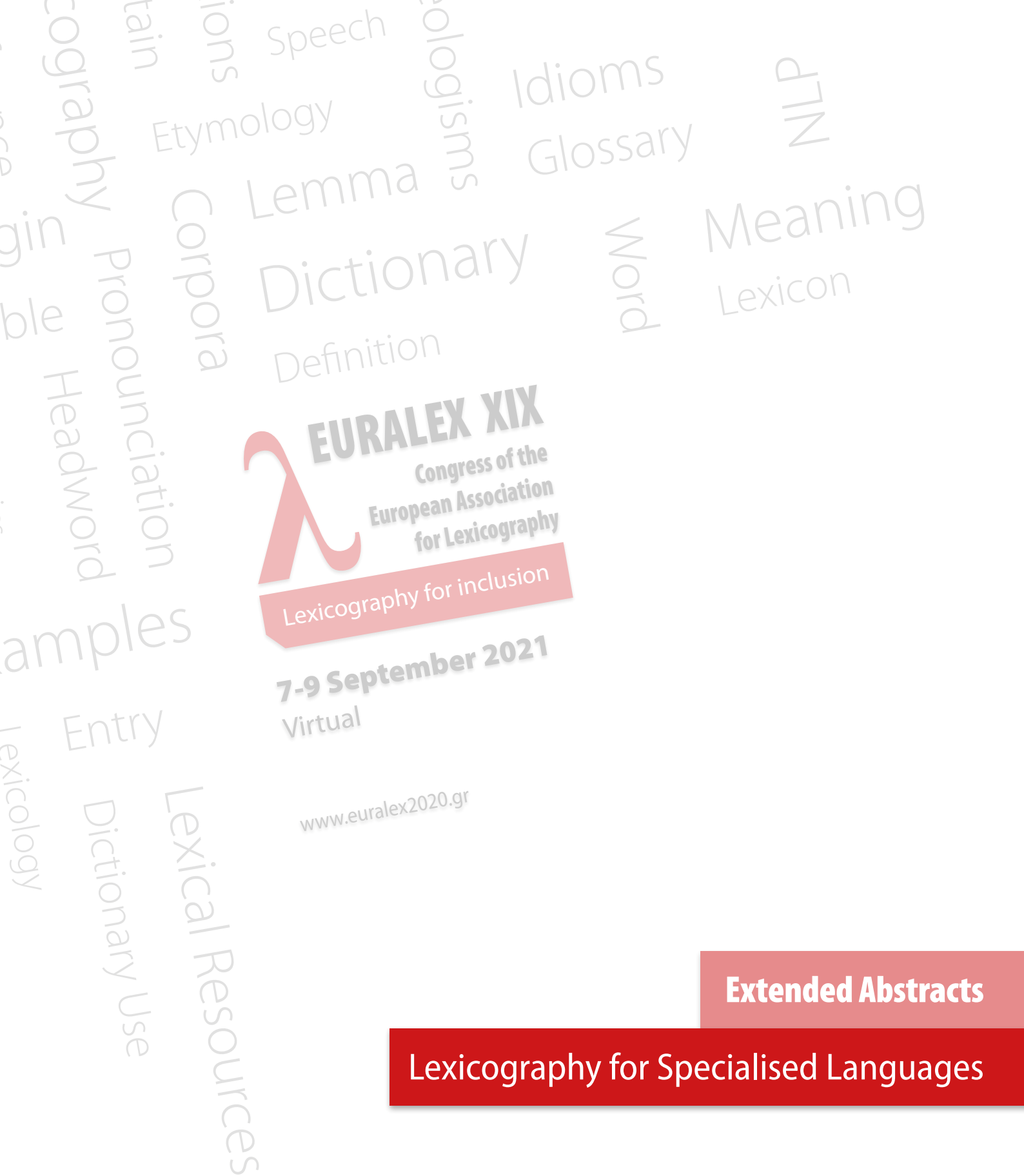
This work has the support of CNPq and FINEP (Edital Viver Sem Limites / 2015)

Keywords: Brazilian Sign Language, Deaf Accessibility, Scientific Communication, Open Journal Systems, Signwriting.

References

AIKHENVALD, A. Y. Classifiers: a typology of noun categorization devices. New York: Oxford University Press, 2008.
BARRETO, M. Escrita de sinais sem mistérios. Salvador, v.1: Libras Escrita, 2015.
BRITO, L. F. Por uma gramática de línguas de sinais. Rio de Janeiro: Tempo Brasileiro: UFRJ, 1995.
_____. Língua brasileira de sinais – LIBRAS. In: BRASIL, Secretaria de Educação Especial. Língua Brasileira de Sinais/ organizado por Lucinda F. Brito Et. AL. Brasília: SEESP, 1997. VIII. (série Atualidades Pedagógicas, n.4). p. 19-61. LIDDELL, S. K. Grammar, gesture, and meaning in American Sign Language. New York: Cambridge University Press, 2003.
DIXON, R. M. W. Basic linguistic theory: grammatical topics. New York: Oxford University press, v. 1, 2010.
FELIPE, T. Libras em contexto: curso básico, livro do estudante cursista. Brasília: Programa Nacional de Apoio à Educação dos Surdos, MEC; SEESP, 2001.
LIMA, V. L. de S. Língua de sinais [manuscrito]: proposta terminológica para a área de desenho arquitetônico. Tese (doutorado) - Universidade Federal de Minas Gerais, Faculdade de Letras. Belo Horizonte, Minas Gerais. 2014
LILLO-MARTIN, D.; KLIMA, E. E. Pointing out differences: ASL pronouns in syntactic theory. In Theoretical issues in sign language research, V. 1: Linguistics, eds. Susan D. Fischer and Patricia Siple, 191-210, Chicago: University of Chicago Press, 1990.
MEIR, I. et al. Re-Thinking sign language verb classes: the body as subject. In: QUADROS, R. M. de; VASCONCELLOS, M. L. B. de (Org). Questões teóricas das pesquisas em línguas de sinais. TISLR 9º THEORETCAL ISSUES IN SIGN LANGUAGE RESEARCH CONFERENCE. Florianópolis, SC: Editora Arara Azul, 2006. p. 82-101.
QUADROS, R. M. de. Phrase structure of Brazilian Sign Language. 1999. Tese de Doutorado. Pontifícia Universidade Católica do Rio Grande do Sul. Porto Alegre. 1999.
_____. Phrase structure of Brazilian Sign Language. In: Cross-linguistic perspectives in sign language research. Selected papers from TISLR 2000. Signum Press: Hamburg. 2003. p.141-162.
QUADROS, R. M. de; KARNOPP, L. B. Língua de sinais brasileira: estudos lingüísticos. Porto Alegre: Artes Médicas, 2004. SANDLER, W.; LILLO-MARTIN, D. Natural sign languages. In: ARONOFF, M.; REES-MILLER, J. (Eds.). Handbook of linguistics. 2001. pp. 533-562.
SCHEMBRI, A. The syntax and morphology of classifiers in Sign Language: rething 'classifiers' in signed languages. In: EMMOREY, K. Perspectives on classifier construction in sign languages. Mahwah, New Jersey: Lawrence Erbaum Associates, Publishers, 2003. P. 03-34.

STOKOE, W. Sign Language structure: an outline of the visual communication systems of the american deaf. Studies in linguistics: occasional papers (Nº 8). Buffalo: Dept. of Anthropology and Linguistics, University of Buffalo, 1960.
SUTTON V. SignWriting: as línguas gestuais são línguas escritas! Manual 1: Noções básicas sobre SignWriting. Tradução de Rafaela Cota Silva. 2009.



Extended Abstracts

Lexicography for Specialised Languages

The contribution of specialized glossaries in heritage cultural conservation: The glossaries of tobacco, silk and Thracian cuisine

Penelope Kambakis-Vougiouklis[1], Asimakis Fliatouras[2], Zoe Gavriilidou[3]

[1] *Department of Greek Philology, Democritus University of Thrace, Komotini, Greece*

Abstract

The aim of present paper is to introduce the three open-access e-glossaries of the Linguistics Laboratory SYNMOPRHOSE (Department of Greek Philology, DUTH, synmorphose.gr/index.php/el) and elaborate on the theoretical principles employed for their compilation. These specialized e-glossaries concern dialectal/idiomatic professional jargon connected with regional/ traditional activities of the Eastern Macedonia & Thrace Prefecture, which is proposed to be included in the list of good safeguarding practices of cultural heritage. More specifically, there are presented (a) The glossary of tobacco in Eastern Macedonia & Thrace, (b) The glossary of silk in Soufli (town of Thrace) and (c) The glossary of Thracian cuisine (mainly of Evros).

The presentation will shed light on:

1. Both the methodology of material collection, based on fieldwork, student work, bibliography etc. and the lexicographical presentation. With the ambitious aim of filling editorial gap in mind, the whole procedure focuses on the implementation stage of specialized glossaries in the Greek language. By merging computer-assisted term extraction with data collected from experts' knowledge, fieldwork and the existing specialized literature, three lists of candidate headwords were drafted. The replicability of the methodology applied makes this pilot lexicographic work procedure generalizable, thus fostering the compilation of specialized glossaries connected to other fields or disciplines.

2. The linguistic canvas of language material, such as semantic (sub-)fields, etymology, quantitative data etc.

3. The contribution of the glossaries to: (a) Linguistics, since they provide untreasured or endangered material and they give evidence for language contact and loan integration between the Greek and other languages (Turkish, Slavic etc) of the area (lexical overlapping/supplementation, code-switching etc.), (b) History/folklore, since they reflect the political, social, economic, cultural and commercial heritage of the area, mainly of previous era, as these activities have been heavily restricted or fading away nowadays. Thus, these glossaries help document and safeguard the local cultural heritage, (c) Academic strategies, since they contribute to the connection of academic work with local community.

Keywords: specialized glossaries, heritage cultural conservation, tobacco, silk, cuisine.

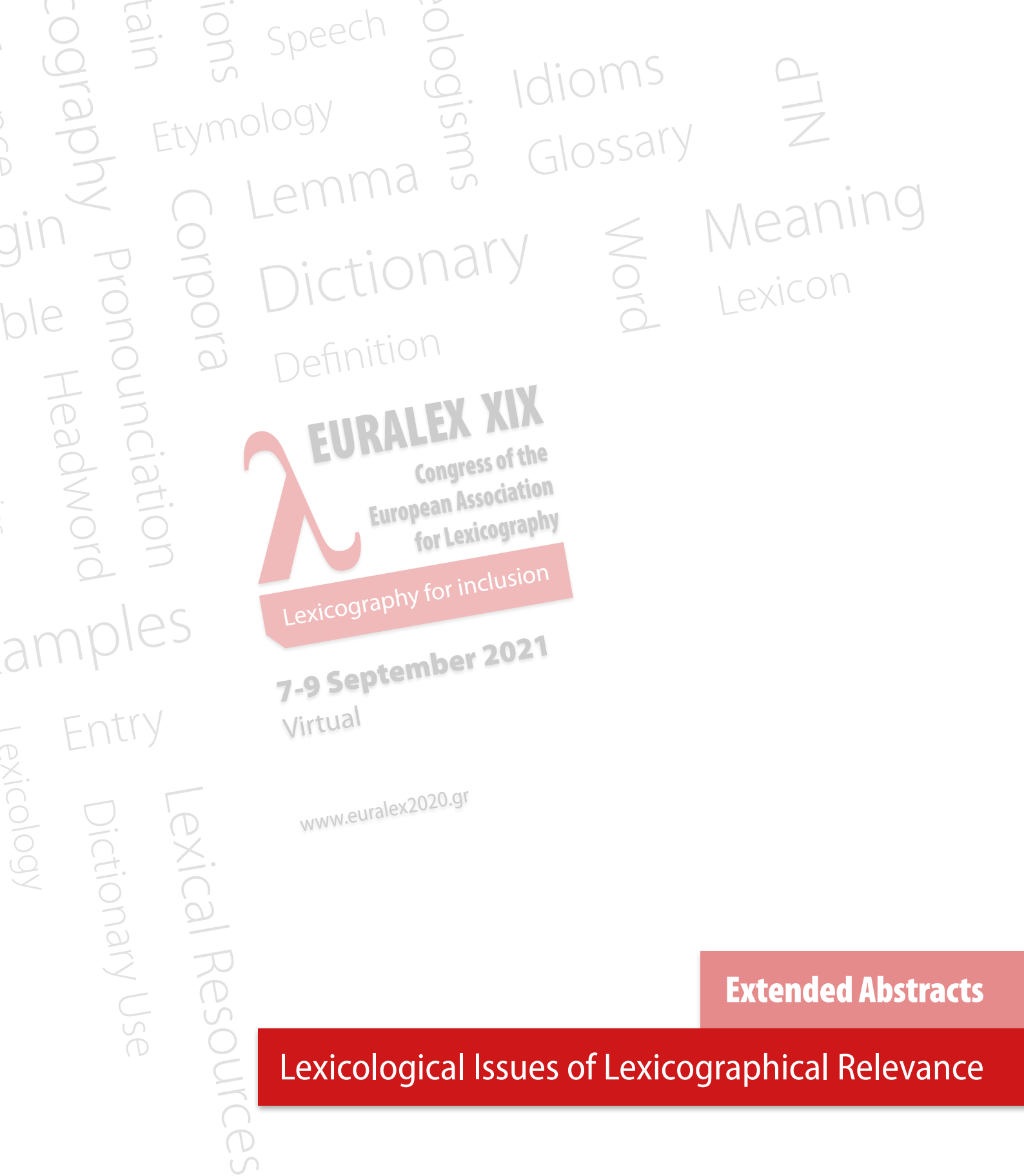
References

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

Bergenholtz, H. (1995). Basic Issues in Specialized Lexicography. In H. Bergenholtz, S. Tarp (eds.) Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries. Amsterdam & Philadelphia: John Benjamins, pp. 14-47.

Bowker, L. (2003). Specialized Lexicography and Specialized Dictionaries. In P. Van Sterkenburg (ed.) A Practical Guide to Lexicography. Amsterdam & Philadelphia: John Benjamins, pp. 154-164.

Bowker, L., Pearson, J. (2002). Working with Specialized Language. A Practical Guide to Using Corpora. London & New York: Routledge.



Extended Abstracts

Lexicological Issues of Lexicographical Relevance

Some Lexical and Lexicographic Issues in Translating Japanese Named Entities

Jack Halpern

The CJK Dictionary Institute, Inc., Niiza, Saitama, Japan

Abstract

1. Introduction

A major issue in both human and machine translation from Japanese is the translation of named entities. The major factors that contribute to this are:

1. The highly irregular orthography of Japanese makes it difficult to identify the large number of orthographic variants (Halpern, 2008).
2. The morphological complexity of Japanese requires the use of a robust morphological analyzer for accurate segmentation and lemmatization (Brill et al., 2001; Yu et al., 2000; Goto et al., 2001).
3. The difficulty of accurately translating proper nouns and points of interest (POIs).
4. The lack of comprehensive lexical resources for named entities, especially for POIs.

This paper discusses some of the key linguistic and lexicographic issues related to Japanese named entities, including (1) orthographic variation, (2) the challenges of Japanese POIs, and (3) the integration of lexicons into NMT systems. It also introduces several large-scale lexical resources, consisting of millions of named entities, and argues that such resources significantly enhance translation accuracy.

2. Japanese Orthographic Variants

2.1 Irregular Orthography

Japanese has a highly irregular orthography resulting from, among other things, the unpredictable interaction between four scripts: kanji, hiragana, katakana and Latin. For example, in 金の卵を産む鶏 /Kin no tamago wo umu niwatori/ 'A hen that lays golden eggs,' /tamago/ has four variants (卵, 玉子, たまご, タマゴ), /niwatori/ three (鶏, にわとり, ニワトリ) and /umu/ two (産む, 生む) (24 permutations). Since such variants occur frequently, MT systems have little hope of identifying them as instances of the same lexeme without support for orthographic disambiguation, (normalization).

2.2 Variant Typology

There are eight types of orthographic variation in Japanese (Halpern, 2008). The three most important ones are:

1. **Okurigana variants** (kana endings attached to kanji) such as in *atarihazure*, 'hit and miss' which can be written in six ways: 当たり外れ, 当り外れ, 当外れ, 当外, 当たりはずれ, あたり外れ Identifying and normalizing the numerous okurigana variants is a major nuisance. An effective solution is to use a lexicon of orthographic variants.
2. **Cross-script variants** refers to variation across multiple Japanese scripts, e.g. /ninjin/ 'carrot' written in kanji (人参), hiragana (にんじん) or katakana (ニンジン). Such variants are common and unpredictable, negatively impact recall, and pose a major challenge to accurate translation.
3. **Katakana variants**, such as コンピュータ vs. コンピューター and チーム vs. ティーム are a major nuisance in translating from Japanese.

3. Lexicons in MT

Large-scale lexicons have dramatically improved translation quality in traditional MT systems (Mediani et al., 2014). Many attempts to replace lexicons with algorithmic solutions, such as for orthographic disambiguation, have been made (Brill et al. (2001); Goto et al.), but such efforts have met with only limited success.

A major problem with these methods is that they often fail to achieve high accuracy unless supported by large-scale lexicons. For example, Emerson (2000), Nakagawa (2004) and others have shown that MT systems and robust morphological analyzers capable of processing *lexemes,* rather than n-grams, must be supported by large-scale lexicons.

NMT systems handle named entities poorly. Although it is technically difficult to integrate lexicons

into such systems, Arthur et al. (2016) showed that integrating "discrete translation lexicons" achieved substantial increases in BLEU (2.0-2.3) and NIST (0.13-0.44) scores.

4. Experiments and Results

4.1 Experiments on POI

NMT systems trained on large corpora often translate high-to-mid-frequency named entities accurately, but for less-frequent (low tail) POIs, the failure rate is high. Below are the results of tests of major NMT engines on less-frequent POIs and orthographic variants, compared to the results of our (CJKI) large-scale lexicons.

Japanese	Google	Bing	CJKI
海の中道線	Midair line of the sea	The middle line of the sea	Umi-no-Nakamichi Line
三角線	Triangle	Triangular line	Misumi Line
鬼の城公園	Demon Castle Park	Demon Castle Park	Oninojo Park

Table 1. POIs by Google and Bing

Japanese	Baidu	NICT	CJKI
海の中道線	The sea line	海の中道線	Umi-no-Nakamichi Line
三角線	Misumi	Misumi Line	Misumi Line
鬼の城公園	Demon Castle Park	Oni Castle Park	Oninojo Park

Table 2. POIs by Baidu and NICT

4.2 Evaluation of results

For less-frequent POIs, the success rate of the above MT engines is less than 50%. “Success” means outputting the "official translation" of that entity, if it exists, which is identical to CJKI’s POI databases.

Google	47%
Microsoft	40%
Baidu	39%
NICT	47%

Table 3. Success rate

The MT engines gave surprisingly poor results, often due to literal translation, e.g.鬼の城公園, correctly 'Oninojo Park', was translated as 'Demon Castle Park' (鬼の 'demon's + 城 'castle' + 園 'park).

4.3 Orthographic variation

MT engines do not perform well in the orthographic normalization of Japanese. Consider the following:

English	Reading	Var. 1	Var 2	Var. 3
sun	hi	日	陽	
mansion	yashiki	屋敷	邸	
shine	sasu	差す	さす	射す

Table 4. Typical variants in Japanese

A sentence including those words can have over ten permutations, such as 日の差さない屋敷, 陽の射さない屋敷 and 陽のささない屋敷, which translates to "A mansion that gets no sunshine". Google and Bing give:

No.	Japanese	Google	Bing
1	日の差さない屋敷	A dwindling residence	A house with no sun
2	日の射さない屋敷	A mansion that does not shine.	She mansion of the day.

3	陽のささない屋	A ya man who does not sunlight.	A house with no sunshine

Table 5. Japanese variants by Google and Bing

Except for Bing (3), all the translations are incorrect. For example, Google translated 陽 to the mysterious 'ya man' and is not aware that it's a variant of 日. In the case of Bing, "She mansion of the day" makes no sense. Baidu and NICT (not shown here) give similarly poor results. Clearly, none of the MT engines surveyed are performing orthographic normalization correctly.

5. Lexical Resources

The CJK Dictionary Institute (CJKI), which specializes in CJK and Arabic computational lexicography, has for decades been engaged in the compilation of learner's dictionaries (Halpern, 2017) and comprehensive lexical resources. Our lexical resources (referred to as Very Large-Scale Lexical Resources or VLSLR), have a special emphasis on named entities. Below are the principal resources designed to enhance the accuracy of Japanese named entity translations.

1. The multilingual Japanese Personal Names Database covers over five million entries, including millions of romanized variants.
2. The Japanese Lexical/Orthographic Database covers about 400,000 entries, including okurigana, kanji, and kana variants for orthographic disambiguation.
3. The Comprehensive Database of Japanese POIs and Place Names, which covers about 3.1 million entries in 14 languages, along with romanized variants.
4. The Database of Katakana Loanwords covers about 50,000 entries.
5. The Database of Japanese Companies covers about 600,000 entries.

6. Conclusions

With computer memory being inexpensive and virtually unlimited, it is not necessary for MT systems to over-rely on corpora and algorithmic solutions. The time has come to leverage the full power of large-scale lexicons to significantly enhance the accuracy of MT. As for NMT, although "lexicon integration" does pose some technical challenges, it is a worthwhile goal to pursue. Ideally, a new kind of “hybrid NMT” that leverages the power of traditional MT systems combined with neural networks and large-scale lexicons will achieve higher quality than ever before.

Keywords: named entities, POI, Japanese, lexicons, MT

References

Brill, E., Kacmarcik, G. and Brockett, C. (2001). Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pages 393-399, Tokyo, Japan.

Emerson, T. (2000). Segmenting Chinese in Unicode. In Proceedings of the 16th International Unicode Conference, Amsterdam

Goto, I., Uratani, N. and Ehara, T. (2001). Cross-Language Information Retrieval of Proper Nouns using Context Information. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, pages 571-578, Tokyo, Japan

Halpern, J. (2008). Exploiting Lexical Resources for Disambiguating Orthographic CJK and Arabic Orthographic Variants. In Proceedings of LREC 2008. Marrakesh, Morocco.

Halpern, J. (2017). Very Large-Scale Lexical Resources to Enhance Chinese and Japanese Machine Translation, Niiza, Japan. http://www.cjk.org/cjk/reference/VLSLR_MT_final.pdf. Retrieved January 30, 2020.

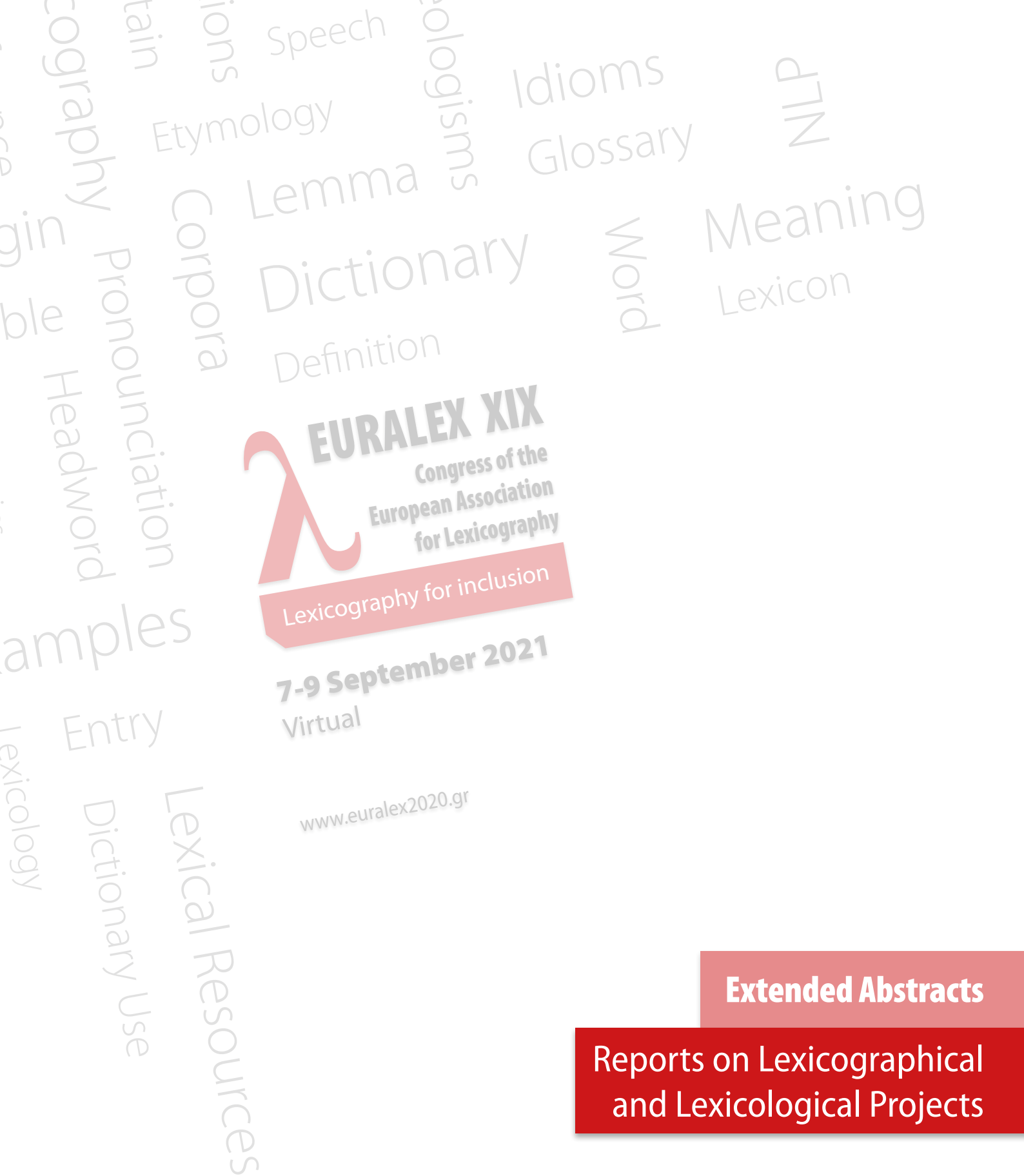
Mediani, M., Winebarger, J. and Waibel, A. (2014). Improving In-Domain Data Selection For Small In-Domain Sets. In Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, USA.

Nakagawa, T. (2004). Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information. In Proceedings of the 20th international conference on Computational Linguistics, p.466-es, Geneva, Switzerland.

Sumita, E. (2013). Multi-Lingual Translation Technology: Special-Purpose System for Multi-Lingual High-Quality Translation. Journal of the National Institute of Information and Communications Technology, pages 35-39, Tokyo, Japan.

Yu, S., Zhu, X. and Wang, H. (2000). New Progress of the Grammatical Knowledge-base of Contemporary Chinese. Journal of Chinese Information Processing, Institute of Computational Linguistics, Peking University,

Vol.15 No.1.



Extended Abstracts

Reports on Lexicographical and Lexicological Projects

Lexico-Semantic Database of Czech

Ondřej Tichý[1], Aleš Klégr[1], Zora Obstová[2], et al.

[1] *Department of English Language, Faculty of Arts, Charles University, Prague, Czech Republic*
[2] *Department of Romance Languages, Faculty of Arts, Charles University, Prague, Czech Republic*

Abstract
This paper is a report on an ongoing project whose aim is to create a digital lexico-semantic database of Czech. The initial structure and data will be based on the only two Czech onomasiological printed dictionaries: Haller's (1969-77) *Český slovník věcný a synonymický* and Klégr's (2007) *Tezaurus jazyka českého*. Haller's dictionary (over 1700 pages, excluding the index) is modelled on Hallig and von Wartburg's (1963) conceptual scheme. Klégr's thesaurus is an adaptation of Roget's Thesaurus (2002) to Czech, following Roget's structure and format (500 pages of entries, 1189 pages including the index). The IPR have been duly secured for both source dictionaries and the resulting dataset will therefore be made publicly available under one of the usual open licenses.
By a conservative estimate, Klégr's thesaurus with 885 heads-concepts (subdivided into noun, verb, adjective and adverb sections) each of which is composed of 300 to 400 lexical units per head totals about 300 000 lexical units (the number of lexical items will naturally be somewhat lower). Haller's dictionary which features 3193 heads quite variable in length and structure (though essentially comparable to those of the thesaurus) can be realistically expected to contain twice to three times as many lexical units (i.e. at least 600 000 units). This vast vocabulary is arranged into multiple fine-grained lexico-semantic fields (hierarchies, linear and cluster-like structures).
Given the differences in the design, conception and focus of the two dictionaries and their entries, their merging poses a formidable technical and labour-intensive task. The outcome should be an integrated and multi-functional semantic network resulting from the alignment of the two dictionaries. The final dataset is expected to be larger than either of the two dictionaries, but its total size is impossible to predict with any accuracy, since some overlap is to be expected, but its extent will only be apparent when both dictionaries have been digitized and their entries for the same lexical fields and items compared.
The database will be useful both to the NLP specialist as well as anyone interested in or working with Czech (such as writers, translators, journalists, etc.). The latter group will benefit from a more complete, highly searchable, freely and easily available thesaurus-like application online, while the former group will be able to use the database to conduct semantic and lexicological research: using the online app, the programmatic API or working directly with the exportable dataset.
Moreover, access to semantic relations captured in our database is expected to help future lexicographers with their work or corpus linguists with tasks like semantic tagging or disambiguation. The API will also allow third party applications that may benefit from semantic fields, e.g. to increase search yields.
The publication dates indicate that the dictionaries cover the lexis from the mid-20th century up to the new millennium and although they do not contain later additions found in contemporary corpora, they represent core vocabulary of the language, and certainly all its essential lexico-semantic fields and relations relevant for NLP.
The crucial merit of our dataset, i.e., the mapping of networks of semantic relations in the language carefully hand-crafted by the authors of the dictionaries, is notoriously difficult to generate satisfactorily by means of corpus linguistic tools without a huge amount of time and manual input. Our dataset will be relatively simple to update (in terms of lexical neologisms and losses in respective fields) using contemporary corpus data and it can be easily enriched with frequency and collocability data on individual lexical items.
Currently we are in the digitization stage with the data already scanned and OCRed. The data from Klégr's dictionary are cleaned and structurally tagged, while the data from Haller's dictionary still need to be processed. Since the structure of this dictionary is more complex and less consistent, we have decided to use Grobid-dictionaries, a machine learning technology geared towards lexicographical digitization. While it requires manual training, it should save time and increase precision in the long

run (although we plan to revise the OCR and tagging manually as well). The tool is currently being adapted, since this will be the first time it is used for an onomasiological dictionary. The result of this stage will be a TEI-XML dataset, but it has already become obvious that transforming the data into a fully compliant format will pose a number of theoretical, technical and methodological difficulties (not least because TEI has not to our knowledge so far focused on onomasiological lexicography, though a lot of work is currently being done to accommodate lexicographers).

Concurrently to the data processing stage, the programming stage has started this year with the development of data storage, API and search facilities, which will be followed by the front-end of the online application.

At this key milestone in the project we are reaching out to lexicologists, lexicographers and other prospective users for feedback, best practices and feature requests. It should also be noted that the project is not necessarily limited to speakers or scholars of Czech since we plan to connect the database with the English WordNet and thereby potentially with other semantic databases in other languages.

Keywords: lexicography, lexicology, semantics, onomasiology, Czech, thesaurus, dictionary, digitization, lexical database.

References

Haller, Jiří (1969-77) Český slovník věcný a synonymický: 1-3 [The Czech conceptual and Synonymous Dictionary]. Praha: Státní pedagogické nakladatelství.

Hallig, Rudolf, von Wartburg, Walther (1963) Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas. Berlin: Akademie Verlag.

Khemakhem, Mohamed / Herold, Axel / Romary, Laurent (2018) Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018. Miyazaki: Japan. hal-01708137v2.

Klégr, Aleš (2000) Rogetův Thesaurus a onomaziologická lexikografie [Roget's Thesaurus and onomasiological lexicography]. Časopis pro moderní filologii 82/2, pp. 65-84.

Klégr, Aleš (2007) Tezaurus jazyka českého. Slovník českých slov a frází souznačných, blízkých a příbuzných [Thesaurus of the Czech Language. A Dictionary of Synonymous, Similar and Related Words and Phrases]. Prague: Nakladatelství Lidové noviny.

Klégr, Aleš (2008) Turning Roget's Thesaurus into a Czech Thesaurus. In E. Bernal, J. De Cesaris eds, Proceedings of the XIII Euralex International Congress. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pp. 697-702.

Nimb, S., Trap-Jensen, L., & Lorentzen, H. (2014). The Danish Thesaurus: Problems and Perspectives. Proceedings of the XVI EURALEX International Congress: The User in Focus, 15–19.

Obstová, Zora (in press) Zwei onomasiologische Wörterbücher als Basis für eine lexikalisch-semantische Datenbank des Tschechischen. In M. Šemelík / V. Kloudová (eds), Spielräume der modernen linguistischen Forschung. Praha: Karolinum.

Roget's Thesaurus of English Words and Phrases (red. R.A. Dutch) (1962) Longman [1982 (red. S.M. Lloyd), 1987, 1998 (red. B. Kirkpatrick), 2002 (red. G. Davidson, Penguin)].

Consolidated dictionary of Russian dialects: current status and prospects

Olga Krylova[1], Elena Syanova[1]

[1] *Department of dialect lexicography and linguo-geography of the Russian language, Institute for Linguistic Studies Russian Academy of Sciences, Russia, St.-Petersburg*

Abstract

«Dictionary of Russian folk dialects» – the only modern project to create a consolidated dictionary of Russian folk dialects. The lexicographic project includes materials from dialects that exist throughout the functioning of the Russian language. This project is a continuation of work on the study of the dialect macrostructure of all Russian dialects on the materials of the XIX–XXI centuries. The Dictionary is based on direct observations of lively folk speech, works of oral folklore. Such a dictionary is an urgent task of Russian dialect lexicography. Its creation was prepared by the whole course of theoretical and practical development of the problems of Russian dialectologists for over 70 years and is based on the achievements of related sciences (linguistic geography, dialect lexicology).

The first issue of the Dictionary of Russian Folk dialects was published in 1965. By 2020, 51 issues with a total volume of more than 18 thousand pages were published (letters «A» – «Kh»). The Dictionary is based on the data of the Card index of the dictionary of Russian folk dialects – the largest repository of Russian dialect vocabulary: more than 2 million cards illustrating more than 300 thousand dialect words and meanings. The file contains vocabulary denoting the flora and fauna, relief, reservoirs, meteorological phenomena, a person with his characteristics, tools, crafts and crafts, customs, ceremonies, games, housing, clothes, shoes, etc. This is a kind of encyclopedia of the life of the Russian people, their material and spiritual culture, reflected in the language. Of particular value are the manuscript materials of the Archive ILI RAS. They recorded many words or meanings of words not noted in other sources. More than 600 domestic and foreign researchers worked in the file cabinet of the Dictionary.

The dictionary records the lexical and phraseological richness of Russian folk speech with all the meanings of words and phraseological combinations, shows the features of their use in Russian dialects of the XIX–XXI centuries. The Dictionary provides a detailed description of the semantic content of the vocabulary of all Russian dialects, traces the development of secondary meanings, gives grammatical, stylistic, compatibility, areal and chronological characteristics of dialect vocabulary.

In the process of lexicographic development of dialect vocabulary, many thousands of words, entire categories and layers of vocabulary were put into scientific circulation, the features of their semantics were shown, and the conditions for their functioning were revealed. The materials contribute to the establishment and refinement of the distribution boundaries of dialect words, their isogloss, the development of the problem of the ratio of dialect and popular meanings in a word, and the tracking of the penetration of dialect words into a literary language.

A theoretically important feature of the Dictionary is the historical approach to dialect facts. The historical aspect in the study of dialect vocabulary is realized in the development of such phenomena as polysemy, polysemy of words. A number of other serious lexicographical issues are associated with it: order, sequence of meanings in semantic structures, etc.

At the Institute for Linguistic Studies Russian Academy of Sciences on the basis of 1) the Big Vocabulary of the ILI RAS, 2) the vocabulary catalog of the Russian folk dialects, 3) unique archival data, 4) handwritten materials, 5) original dialect materials collected annually by employees of the institute and colleagues from universities in the country during the field expeditions (dialect units operating in modern Russian dialects of the Russian Federation), work began on creating a database

of Russian folk dialects as a storage system with a wide range of functional capabilities of information processing. The linguistic resource in the project are the actual language resources – dialect texts, lexical units (individual words, stable combinations, phraseological structures, proper phraseological units), grammatical forms. The following information will be included: graphic and phonetic appearance of the dialect word (units), lexical meaning, illustration (-s), distribution area, fixation date, information on the informant, information on the researcher (collector) who made the recording (if relevant data is available). The methods and techniques of linguistic support of the automated system are developed, the tasks of preserving the authenticity and at the same time correctness and completeness of dialect materials representing the live speech of a modern informant are being solved. One of the priorities is the task of promptly retrieving and processing the original dialect data.
A new generation of electronic dictionary is supposed to be based on the data. The dictionary should contain comprehensive information, allowing to represent the word as a unit of language and at the same time as a unit operating in a broad cultural and historical context.

Keywords: Lexicography, consolidated dictionary, dictionary of Russian dialects, dialectology, history of the Russian language, lexicology, semantics, phraseology, ethnography, history of Russian folk culture.

Where Does Invisible Culture Belong In Dictionaries?

Lauren Sadow

School of Literature, Languages, and Linguistics, The Australian National University, Canberra, Australia

Abstract

Language learners and their teachers need to understand culture in order to use and teach the language. Dictionaries (in all their forms) are already key language resources for both learning and teaching foreign languages, yet provide scant opportunities for learning about that culture. Invisible culture – the values, attitudes, and beliefs which underpin a culture and a language – is key in both interpreting and producing language. It can be the difference to understanding different connotations of synonyms or why a word may be derogatory.

The extent to which culture is a part of meaning rather than scientific knowledge has shaped debates to the place of culture in dictionaries over many years (Silverstein, 2006). Silverstein argued for an "ethnographic lexicography" which included elements of culture as integral to language. Using this as a starting point, I also follow Goddard and Wierzbicka's (2013) argument that cultural perception, usage, and meaning are all intertwined, and therefore should be included in dictionaries and other lexicographical works.

From time to time, dictionaries which place a bigger focus on culture emerge in the English-language market, but are typically limited to culture-specific definitions of particular terms, or including encyclopaedic information in addition to the definition. However, there have also been some rare examples of attempting to include the elements of invisible culture in dictionaries (see for example Cummings & Wolf, 2011). Despite this, the term 'cultural dictionary' is a relatively unused descriptor.

The challenge with including elements of invisible culture in dictionaries is that in many cases, invisible culture is just that – invisible – and therefore does not have consistent lexical equivalents (such as the Anglo attitude of "not losing face when you ask for help", which could be expressed as "it's okay to ask for help" or "no one thinks badly of you if you need help" or any other number of phrases, but no single word), which makes it difficult to reference in a dictionary. This is further complicated by the fact that a single element of invisible culture is important for so many aspects of language. Not only in terms of lexical items, but also in terms of the pragmatics attached to those items (such as the above attitude being linked to Anglo expectations of learning, to the phrase "can I help you with anything?", to attitudes in relationships, and so on).

The question then becomes how can invisible culture be included in dictionaries, and how much of it should be included?

This paper uses the Australian Dictionary of Invisible Culture for Teachers (AusDICT) project – a dictionary of invisible culture in Australian English, targeted at English language teachers of migrant students – as a starting point for answering these questions. The AusDICT Project engaged with target users in several rounds of consultation and testing to develop a product which was maximally useful and useable. The final AusDICT contained both entries which described culturally significant words and those which described values and ways of interacting. In both cases, the focus was on the culturally conditioned components in meaning, including social evaluation.

A challenge in this dictionary project was organising the entries for users. Because of most entries lacking an established word or phrase by which to refer to them, an alphabetical arrangement was not practical for users. Here, onomasiological lexicography (e.g. Ostermann, 2015) provided a way to approach a non-alphabetical structure, by instead focusing on the cognitive connections between entries, and grouping them by themes or topics, such as "Education", "Work", "Expressing opinions" "Humour" and so on. Because of the interrelated nature of many entries in the AusDICT, these themes were a mix of domain-specific and interaction-specific themes, with entries included based on user feedback about where they would search for such topics. This structure also worked best for the target user because they tend to work in language teaching classrooms teaching by themes and topics similar to these ones. This meant that they would then be able to use the dictionary more consistently in their classroom

practice. The AusDICT then used extensive cross-referencing between entries to enhance these cognitive connections, which created a network of ideas then listed in indices.
To make the themes and indices as intuitive for users as possible, electronic lexicography was an essential component of this project, with the first version being in an eBook format. In particular, natively built cross-referencing within that format meant that users were able to move easily between connected ideas to explore ideas in greater depth, or compare entries. Such in-built features also allowed for several indices of entries and tables of contents using different search pathways (such as by part of speech, alphabetical order, or listed entries in each theme), enhancing the users' ability to find entries no matter their perspective on the value or attitude.
This project built on Silverstein's idea of "ethnographic lexicography" to show that the place for invisible culture may in fact be front and centre, and not relegated to an appendix or featured section of a dictionary. While including invisible culture – both in lexical entries and in higher level values – required that the dictionary be fundamentally reorganised, the new networks of values, attitudes, words, and phrases highlight a broader picture of the English language. Such a picture then fosters a deeper understanding of the language to teachers, and through them, to their students.

Keywords: Language learning, invisible culture, ethnographic lexicography, cognitive lexicography.

References

Cummings, P., & Wolf, H. (2011). A Dictionary of Hong Kong English: Words From the Fragrant Harbor (Hong Kong English). Hong Kong University Press.
Ostermann, C. (2015). Cognitive lexicography: a new approach to lexicography making use of cognitive semantics. De Gruyter.
Silverstein, M. (2006). Old Wine, New Ethnographic Lexicography. Annual Review of Anthropology, 35(1), 481 496. https://doi.org/10.1146/annurev.anthro.35.081705.123327
Wierzbicka, A., & Goddard, C. (2013). Words and Meanings. Oxford University Press.

Extended Abstracts

Research on Dictionary Use

A hunt for synonyms: results of user research

Kristina Koppel[1], Maria Tuulik[1]

[1] *Institute of the Estonian Language, Tallinn, Estonia*

Abstract

Starting from 2017, a dictionary writing system Ekilex (Tavast et al., 2018) has been in development in the Institute of the Estonian Language. One of the main goals of Ekilex is to join information from the already existing monolingual and bilingual dictionary databases and termbases into one aggregated EKI Combined Dictionary (EKI ühendsõnastik, CombiDic). All dictionary projects starting from 2019 are adding new types of information layers into CombiDic. To the general public, the information from Ekilex will be displayed in Sõnaveeb (Koppel et al., 2019)–a language portal being developed in parallel to Ekilex. We initiated a new project in 2019 to supplement the synonym layer in Ekilex.

Until the beginning of 2020, only synonyms originating from the Dictionary of Estonian (DicEst, Langemets et al., 2018)–which forms the semantic base of Ekilex–were shown to the users in Sõnaveeb. The authors of DicEst had a general principle not to add more than three synonyms per word sense. Synonyms were given primarily to support definitions, since the main emphasis in the dictionary was to explain word meanings, hence only synonyms that share the exact same definition (i.e. words that have identical meanings) were presented in DicEst. The aim of the new project is to supplement the synonym layer in Ekilex/Sõnaveeb with full as well as partial synonyms[1].

In 2019, a database of automatically generated synonym candidates was extracted into Ekilex from the existing dictionaries, benefitting from the tradition of using synonyms in definitions and other fields in the dictionary entry, as well as using semantic mirroring (Dyvik 1998, 2004). Distributional similarity (Turney & Pantel, 2010) was calculated from the multilingual FastText model (Grave et al., 2018). For post-editing the automatically generated lists of synonym candidates, a specialized tool was developed within Ekilex.

Our initial goal was to compile a synonym layer, i.e. post-edit the automatically generated candidates list, for the 10,000 most frequent Estonian words (in Ekilex, there are approximately 160,000 Estonian headwords). Before starting to compile the entries, we wanted to identify users' needs and monitor their behavior in looking up synonyms in Sõnaveeb. For that purpose, multiple designs of the UX prototype was prepared that were shown to the user one at a time, followed by a request to replace a word with a synonym in a specific sentence. To map users' preferences a semi-structured interview was conducted.

In the light of automated lexicography, the question arises whether lexicographic work will become obsolete. Our research tries to answer the questions: would users be satisfied with a synonym cloud generated by a combination of automated solutions (corpus summary + semantic mirroring of existing dictionaries) that requires them to do additional analysis if the cloud also contains more semantically distant words; would they prefer a cloud edited by lexicographers or would users find relevant information more easily from lists of synonyms divided according to word senses–a presentation which requires thorough semantic analysis from lexicographers.

In the user research three types of people were inquired: professionals with a linguistic background (e.g. editors, translators, linguists), general users (e.g. journalists, developers, social workers), and learners of Estonian as a second or a foreign language. We were looking for answers to the following questions:

- Where do users first look when they are searching for synonyms?
- Do users prefer synonyms being listed under the corresponding senses or do they favor a cloud of synonyms without sense division?
- Are synonym clouds edited by lexicographers preferred or would automatic clouds be acceptable as well?

[1] In Ekilex, full synonyms share a meaning, partial synonyms have meaning relations (read more about the data model for synonyms in Tavast et al., 2020, to be published in the XIX EURALEX International Congress proceedings).

- Does the noise in the automatically generated cloud of synonyms distract the users, e.g. does it puzzle them that the word *kaelkirjak* 'giraffe' is presented to them as a synonym for the headword *jänes* 'rabbit'?
- Do users understand the difference between full and partial synonyms and do they want the difference between the two types of synonyms to be displayed in an explicit way, e.g. using background colors similarly to Thesaurus.com?
- How many synonyms would users like to see and when does it become too many?
- Do users find style and register labels useful?

In the poster, we will provide an overview of users reasoning behind their preferences of the presentation of synonyms.

Keywords: synonyms, user research, Estonian language

References

Eesti keele sõnaraamat 2019. [The Dictionary of Estonian 2019, DicEst.] Tallinn: Eesti Keele Instituut, Sõnaveeb. Retrieved from https://sonaveeb.ee/.
EKI ühendsõnastik 2020. [EKI Combined Dictionary 2020, CombiDic.] Tallinn: Eesti Keele Instituut, Sõnaveeb. Retrieved from https://sonaveeb.ee/.
Ekilex. https://ekilex.eki.ee/ (21.07.2020).
Dyvik, H. 1998. A translational basis for semantics. Language and Computers, 24, 51–86.
Dyvik, H. 2004. Translations as semantic mirrors: from parallel corpus to wordnet. Language and Computers, 49(1), 311–326.
Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. 2018. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893.
Koppel, K.; Tavast, A.; Langemets, M. & Kallas, J. 2019. Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. – Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, Jose Pedro Ferreria, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek & Carole Tiberius (Eds), Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of eLex 2019 conference, 1–3 October 2019, Sintra, Portugal. Brno: Lexical Computing CZ, s.r.o., 434−452.
Langemets, M. Tiits, M., Udo, U., Valdre, T. & Voll, P. 2018. Eesti keel uues kuues: Eesti keele sõnaraamat 2018. Keel ja Kirjandus, 12, pp. 942–958.
Sõnaveeb. https://sonaveeb.ee/ (21.07.2020).
Tavast, A., Langemets, M., Kallas, J. & Koppel, K. 2018. Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In J. Čibej, V. Gorjanc, I. Kosem, S. Krek (eds.) Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts. Ljubljana, Slovenia, pp. 749–761.
Tavast, A., Koppel, K., Langemets, M., Kallas, J. 2020. Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations. [To be published in XIX EURALEX International Congress proceedings].
Thesaurus.com. https://www.thesaurus.com/ (21.07.2020).
Turney, P. D., & Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37(1), 141–188.

The usefulness of illustrations in dictionaries

Anna Dziemianko

Adam Mickiewicz University, Poland

Abstract

1.Introduction

Research shows that users expect illustrations in dictionaries (Klosa et al. 2014). A variety of illustration categories have been distinguished (e.g., Hupka 1989, 2003, Ilson 1987, Kemmer 2014a, Lew 2010, Stein 1991, Svensén 2009). Among them are single pictures and line drawings, dealt with below. Showing precisely one object, they are most common in dictionaries (Klosa 2015: 519), where their function is to support comprehension (Lew et al. 2018: 54). Empirical studies demonstrate that illustrations in dictionaries improve meaning reception (Nesi 1998), retention (Gumkowska 2008) and do not detract users' attention from definitions (Kemmer 2014b, Lew et al. 2018). It is thus fortunate that some issues concerning the inclusion of illustrations in paper dictionaries, like expensive typesetting and printing (especially in color), do not apply to e-dictionaries. While cost is less of a criterion for choosing between greyscale and color images for e-dictionaries than it is for paper ones, it is not known whether color makes illustrations any more beneficial (Gangla 2001, Biesaga 2016). This is an interesting question relevant not only to dictionaries in book form; even though e-dictionaries are often consulted on devices with color displays, those accessed on e-book readers and handheld portables are usually in greyscale. Besides, "there is no evidence that full-colour illustrations are in fact more effective in dictionaries than the more traditional simple iconic line drawings" (Lew 2010: 299). It is thus worthwhile to see whether pictures in color, greyscale or iconic line drawings are more useful for reception and retention. It is also advisable to compare entries with and without illustrations, as it is not known how the inclusion of illustrations affects consultation time (Lew et al. 2018: 75).

2.Aim

The aim of the paper is to investigate the effect of illustrations (color pictures, greyscale pictures and line drawings) in English dictionaries on language reception and learning. The following questions are posed:

1. Do color pictures, greyscale pictures and line drawings in English dictionaries affect meaning comprehension?
2. Is the time of meaning comprehension on the basis of dictionary consultation dependent on the presence and type of illustrations in entries?
3. Do color pictures, greyscale pictures and line drawings influence dictionary-based meaning retention (both immediate and delayed)?

3.Methods

An online experiment of four parts was conducted (a pre-test, a main test, immediate and delayed post-tests). The main test included 15 infrequent concrete English nouns whose meaning had to be explained. To perform the task, purpose-built monolingual entries were supplied, which differed only in illustrations. Four experimental conditions (dictionaries) were created: with color pictures, greyscale pictures, iconic line drawings and without any illustrations.

In the other tests, the nouns had to be explained without access to any dictionary. The pre-test identified the cases where the words were known. Meaning retention was checked in immediate and delayed post-tests.

238 learners of English (B2 in CEFR) participated in the study. 61 of them accessed dictionaries with color pictures, 60 – with greyscale pictures, 58 – with line drawings, 59 – with no illustrations.

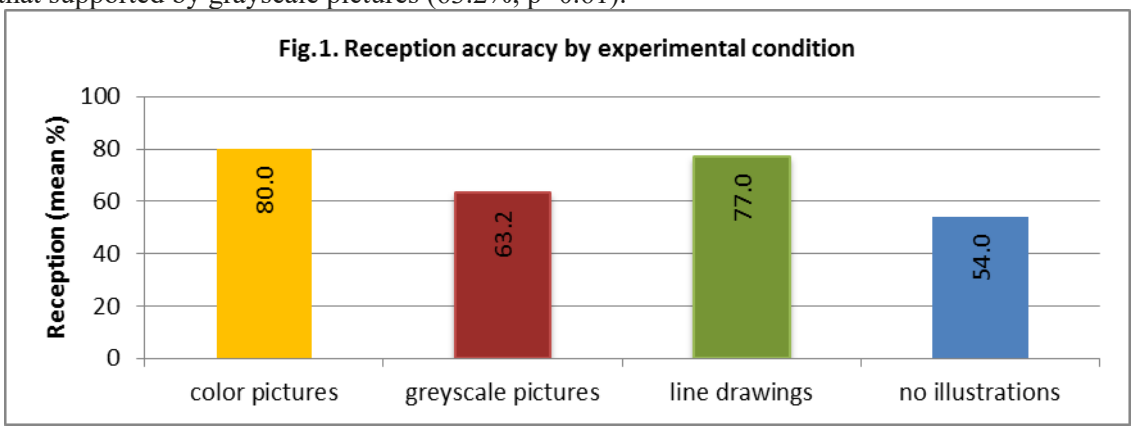
4.Results

One-way GLM ANOVAs were calculated for each dependent variable (reception accuracy, reception time, immediate retention, delayed retention). The Tukey HSD test was used to conduct post hoc comparisons.

4.1. Main test

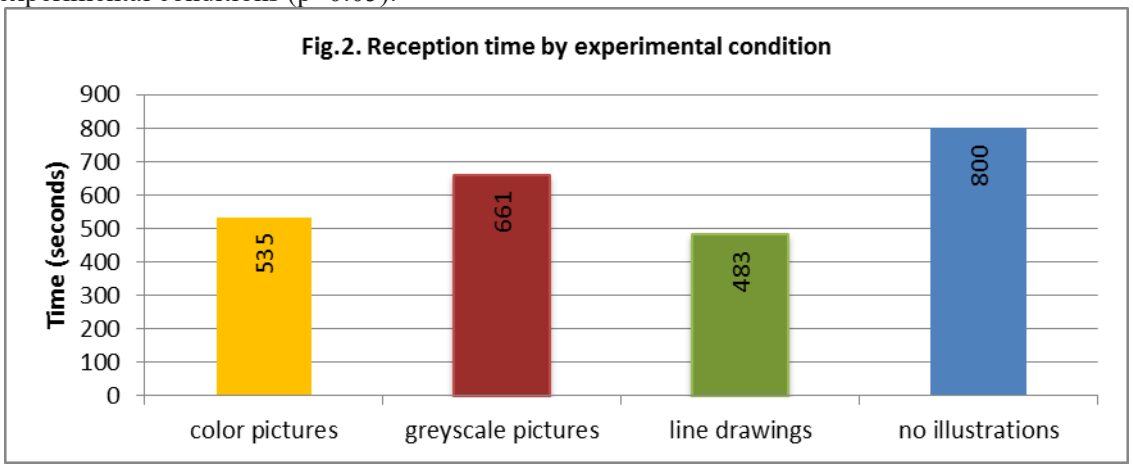
4.1.1. Reception

Illustrations had a statistically significant effect on reception (F=5.21, p=0.00, partial η^2=0.218, Fig.1). Meaning explanation was significantly less successful when no illustrations were given in entries (54%) than when color pictures (80%, p=0.01) or line drawings (77%, p=0.02) were present. There were no statistically significant differences in reception between the experimental conditions with illustrations (p>0.05). Also, reception not assisted by illustrations (54%) was comparable with that supported by grayscale pictures (63.2%, p=0.61).



4.1.2. Time

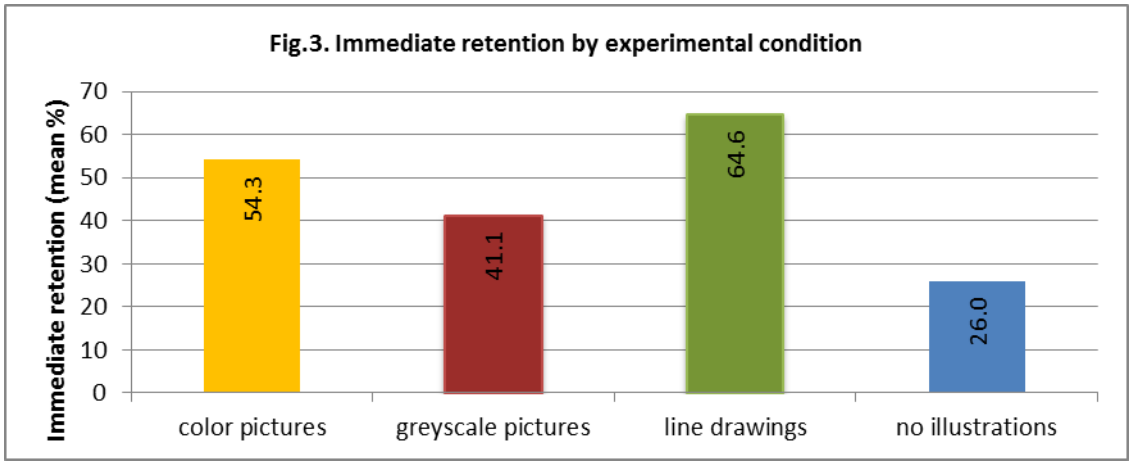
Illustrations significantly affected the time of meaning comprehension (F=18.06, p=0.00, partial η^2=0.492, Fig.2). Reception was the fastest when line drawings (483s) and color pictures (535s) were present, with no difference between these two conditions (p=0.70). Meaning comprehension took significantly longer when entries featured greyscale pictures (661s, p<0.05). Decoding not assisted by any illustrations (800s) was the most time-consuming, significantly more than in the other three experimental conditions (p<0.05).



4.2. Retention

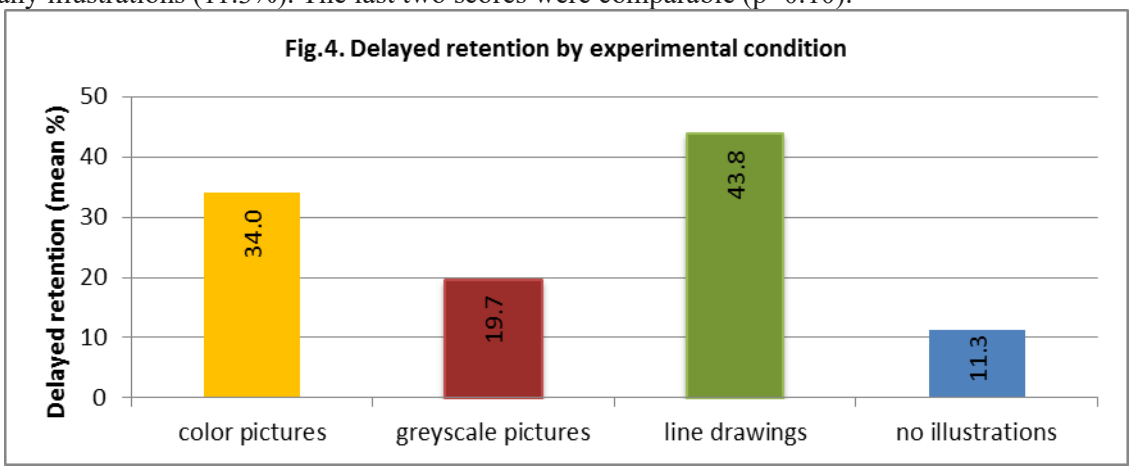
4.2.1. Immediate retention

Illustrations had a strong impact on immediate retention (F=20.14, p=0.00, partial η^2=0.519, Fig.3). Significantly more meanings were remembered in any condition with illustrations than in the one without them (p<0.05). Most meanings were retained when entries offered line drawings (64.6%), significantly more than when greyscale pictures were given (41.1%, p=0.00). The retention rate for entries with color pictures (54.3%) was not significantly different from the scores in the other two conditions with illustrations (p>0.05).



4.2.2. Delayed retention

Illustrations also had a powerful influence on delayed retention ($F=32.97$, $p=0.00$, partial $\eta^2=0.638$, Fig.4). In the long run, line drawings were the most beneficial for meaning retention. Over two fifths of meanings (43.8%) were remembered two weeks after exposure to line drawings, which is significantly more than in any other condition ($p<0.05$). Color pictures came off second best, since about one third of meanings (34%) were remembered with their help. This result significantly exceeded those in the remaining two conditions ($p<0.05$). Only one fifth of meanings (19.7%) were remembered after greyscale pictures had been seen in the main test, and one tenth – in the absence of any illustrations (11.3%). The last two scores were comparable ($p=0.10$).



5.Conclusions

Overall, graphic illustrations in dictionaries proved to be beneficial for language reception and learning.

Color pictures and line drawings help most to comprehend meaning, whereas greyscale pictures do not make any significant contribution in this respect in comparison with illustration-free dictionaries. Also, line drawings and color pictures ensure the most time-efficient reception. Meaning comprehension takes much longer when greyscale pictures are given, and the lack of illustrations extends reception most.

To stimulate immediate and delayed retention, line drawings are most recommendable, followed by color and greyscale pictures. Yet, in the long run, greyscale images do not help retention significantly more than illustration-free dictionaries.

Limitations of the study and implications for printed and e-dictionaries are discussed in the full paper.

Keywords: dictionary, illustrations, reception, retention

Selected references

Biesaga, Monika. 2016. Pictorial illustration in dictionaries: The state of theoretical art. In Margalitadze, T. and M. George (eds.), Proceedings of the XVII EURALEX International Congress. Tbilisi: Lexicographic Centre, Ivane Javakhishvili Tbilisi State University, 99– 108.

Biesaga, Monika. 2017a. Dictionary tradition vs. pictorial corpora: Which vocabulary thematic fields should be illustrated? Lexikos 27: 132–151.
Biesaga, Monika. 2017b. Pictorial illustrations in encyclopaedias and in dictionaries – A comparison. Kosem I., C. Tiberius, M. Jakubícek, J. Kallas, S. Krek and V. Baisa (eds), Electronic lexicography in the 21st century. Proceedings of eLex 2017 Conference, Leiden, the Netherlands, 19-21 September 2017. Brno: Lexical Computing: 221–236.
Faber, Pamela, Pilar León Araúz, Juan Antonio Prieto Velasco and Arianne Reimerink. 2007. Linking images and words: The description of specialized concepts. International Journal of Lexicography 20.1: 39–65.
Gangla, Lilian Atieno. 2001. Pictorial illustrations in dictionaries. MA Thesis, Pretoria: University of Pretoria.
Gumkowska, Anna. 2008. The role of dictionary illustrations in the acquisition of concrete nouns by primary school learners and college students of English. BA Thesis, Collegium Balticum.
Hartmann, Reinhard R. K. 2013. Mixed dictionary genres. In Gouws, R., U. Heid, W. Schweickard et al. (eds), Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent Developments with focus on electronic and computational lexicography. Berlin, Boston: De Gruyter Mouton, 381-393.
Höger, Rainer. 2005. Die Aufmerksamkeitsverteilung bei der Betrachtung von Bildern. In Sachs-Hombach, K. (ed.), Bildwissenschaft zwischen Reflexion und Anwendung. Cologne: Halem, 331–41.
Hupka, Werner. 1989. Wort und Bild. Die Illustrationen in Wörterbüchern und Enzyklopädien. (Lexicographica. Series Maior 22). Tübingen: Niemeyer.
Hupka, Werner. 2003. How pictorial illustrations interact with verbal information in the dictionary entry: A case study. In Hartmann, R. R. K. (ed.), Lexicography. Critical concepts. London: Routledge, 363–390.
Ilson, Robert F. 1987. Illustrations in dictionaries. In Cowie, A. P. (ed.), The dictionary and the language learner. (Lexicographica Series Maior 17). Tübingen: Niemeyer, 193–212.
Kemmer, Katharina. 2014a. Illustrationen im Onlinewörterbuch. Text-Bild-Relationen im Wörterbuch und ihre Empirische Untersuchung. Mannheim: Amades.
Kemmer, Katharina. 2014b. Rezeption der Illustration, jedoch Vernachlässigung der Paraphrase? Ergebnisse einer Benutzerbefragung und Blickbewegungsstudie. In Müller-Spitzer, C. (ed.), Using online dictionaries. (Lexicographica Series Maior 145.). Berlin: Walter de Gruyter, 251–278.
Klosa, Annette. 2015. Illustrations in dictionaries: Encyclopaedic and cultural information in dictionaries. In Durkin, P. (ed.), The Oxford handbook of lexicography. Oxford: Oxford University Press, 515–531.
Klosa, Annette, Alexander Koplenig and Antje Töpel. 2014. Benutzerwünsche und Benutzermeinungen zu dem monolingualen deutschen Onlinewörterbuch elexiko. In Müller-Spitzer, C. (ed.), Using online dictionaries. Berlin: de Gruyter. 281–384.
Langridge, Sabine. 1998. The genesis and development of illustration in the English dictionary. In De Beaugrande, R., M. Grosman and B. Seidlhofer (eds), Language policy and language education in emerging nations: Focus on Slovenia and Croatia and with contributions from Britain, Austria, Spain and Italy: Stamford, CT/London: Ablex, 69–76.
Lew, Robert. 2010. Multimodal lexicography: The representation of meaning in electronic dictionaries. Lexikos 20: 290-306.
Lew, Robert and Joanna Doroszewska. 2009. Electronic dictionary entries with animated pictures: Lookup preferences and word retention. International Journal of Lexicography 22.3: 239–257.
Lew, Robert, Rafał Kaźmierczak, Ewa Tomczak and Mateusz Leszkowicz. 2018. Competition of definition and pictorial illustration for dictionary users' attention: An eye-tracking study. International Journal of Lexicography 31.1: 53–77.
Liu, Xiqin. 2015. Multimodal definition: The multiplication of meaning in electronic dictionaries. Lexikos 25: 210–232.
Luna, Paul. 2013. Picture this: How illustrations define dictionaries. In Luna, P. and E. Kindel (eds), Typography Papers 9. London: Hyphen Press, 153–172.
Müller-Spitzer, Carolin and Alexander Koplenig. 2014. Online dictionaries: Expectations and demands. In Müller-Spitzer, C. (ed.), Using online dictionaries, Berlin-Boston: de Gruyter, 143–188.
Nesi, Hilary. 1989. How many words is a picture worth? A review of illustrations in dictionaries. In Tickoo, M. L. (ed.), Learners' dictionaries: State of the art. Singapore: SEAMEO, pp. 124-134.
Nesi, Hilary. 1998. Defining a shoehorn: The success of learners' dictionary entries for concrete nouns. In Atkins, B. T. S. (ed.), Using dictionaries. Studies of dictionary use by language learners and translators. (Lexicographica Series Maior 88). Tübingen: Niemeyer, 159–178.
Paivio, Allan. 1990. Mental representations: A dual coding approach. Oxford: Oxford University Press.
Rothenhöfer, Andreas. 2013. New developments in learner's dictionaries II: German. In Gouws, R., U. Heid, W. Schweickard et al. (eds), Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent Developments with focus on electronic and computational lexicography. Berlin, Boston: De Gruyter Mouton, 414–425.
Stein, Gabriele. 1991. Illustrations in dictionaries. International Journal of Lexicography 4.2: 99–127.

Svensén, Bo. 2009. A handbook of lexicography: The theory and practice of dictionary-making. Cambridge: Cambridge University Press.
Szczepaniak, Renata and Robert Lew. 2011. The role of imagery in dictionaries of idioms. Applied Linguistics 32.3: 323–347.

Information Needs and Contextualization in the Dictionary Consultation Process

Theo Bothma[1], Rufus Gouws[2]

[1] *Department of Information Science, University of Pretoria, Pretoria, South Africa*
[2] *Department of Afrikaans And Dutch, Stellenbosch University, Stellenbosch, South Africa*

Abstract

Dictionaries offer curated data to users, and, based on the scope of the dictionary, present items giving the paraphrases of meaning or translation equivalents for all the senses of the lexical item represented by the lemma sign being consulted This is in addition to other data, such as items giving example sentences, morphology, pronunciation, etc. Neither the items giving the different paraphrases of meaning nor those giving the translation equivalents are contextualized in terms of the context or situation of the information need of the user and in some dictionary articles no cotextual items like example sentences and collocations are included. The user therefore has to rely on their intuition or apply their mind to select the correct paraphrase of meaning or translation equivalent. In many dictionary articles the information retrieval is impeded due to the extent and density of the article and this requires careful analysis and a considerable amount of time to read and comprehend. In such cases, the user is often subjected to an unnecessary information overload (cf. Gouws and Tarp 2017).

The main issue to be addressed in this paper is to reflect on how dictionaries can be improved by providing more items that can enhance the contextualization and cotextualization of information presented to users.

Tarp and Gouws (2019) provide an example of a Write Assistant for L1/L2 text production, which provides contextualized information to the user: "Write Assistant 'observes' its users and, on this basis, chooses the set of lexicographical data most likely to meet their needs in each concrete situation" (p. 265). It is unclear what the underlying technologies for this "observation" is, and from the example it seems as if the user still needs to navigate through possibly a large amount of data (albeit not a full dictionary article) before the correct translation equivalent in the specific context is identified.

For text reception, Tarp and Gouws (2019:262ff) discuss the linking of individual words in an e-text to user-specified e-dictionaries. The user is presented with a "traditional dictionary article" (p. 264), and they acknowledge that contextualization in these situations implies only that "the user gets access to the lexicographical service directly in the context where the information need occurs" (p. 264); in addition, the user is presented with information that is irrelevant in a text reception situation. Even though such linking is very useful, the linking technology is not always correct, as is described in some detail in Bothma and Prinsloo 2013.

Contextualization described in the preceding paragraphs may simplify the dictionary consultation process considerably. However, in none of the cases is the system fully aware of the pragmatic environment, the context, or the syntactic environment, the cotext (cf. Lettner 2020) in which the word occurs, and the user is offered a number of options from which to choose. The user therefore still has to apply their mind to identify the correct sense or translation equivalent in context.

In this paper, we intend to expand on the preceding brief description, and discuss a number of examples that show the complexity of contextual analysis (and to a lesser degree the cotextual situation) that may be required for e-dictionaries to provide the user with more relevant information in a specific information need situation. The focus will be on text reception, i.e., when the user reads a text in L1 or L2 and needs to understand a word or phrase in the text. Each text has a specific context, but the dictionary provides all possible senses for a word in all possible contexts and cotexts. To automate or partially automate the interaction between the text and the dictionary, the system has to select a specific context from all the senses in the comment on semantics of a given dictionary article. The text has specific attributes, for example, it is a historical novel set in a specific period and country, a contemporary newspaper article reporting on recent political events, or an academic article dealing with a contentious topic in a specific discipline. A human reading the text understands the context; in many cases the dictionary could provide pragmatic details, for example through labels, indicating that a

specific sense is archaic/regional/scientific etc. To what extent can such and more complex mappings be automated and what technologies are required? Does a text need to have formal indications about context? Should the cotext be negotiated? Therefore, should software be able to analyse the text easily to establish context according to the cotext? What role could granular metadata mark-up of textual data play in establishing context? Would the structure of the dictionary database need to be adapted or redesigned to accommodate such potential mappings? How does the interaction between the text and the dictionary database occur? Could it make use of traditional search technologies to establish links, or should there be more complex software running in the background? Such software, if needed, should not be something the user should have to manipulate, and the software should run in a "black box" that interprets the text, interacts with the dictionary database and provides the user with a contextualized result. What are the characteristics and modular components of such software? To what extent would part-of-speech parsers, natural language processing technologies, artificial intelligence etc. be required to enable this interaction?

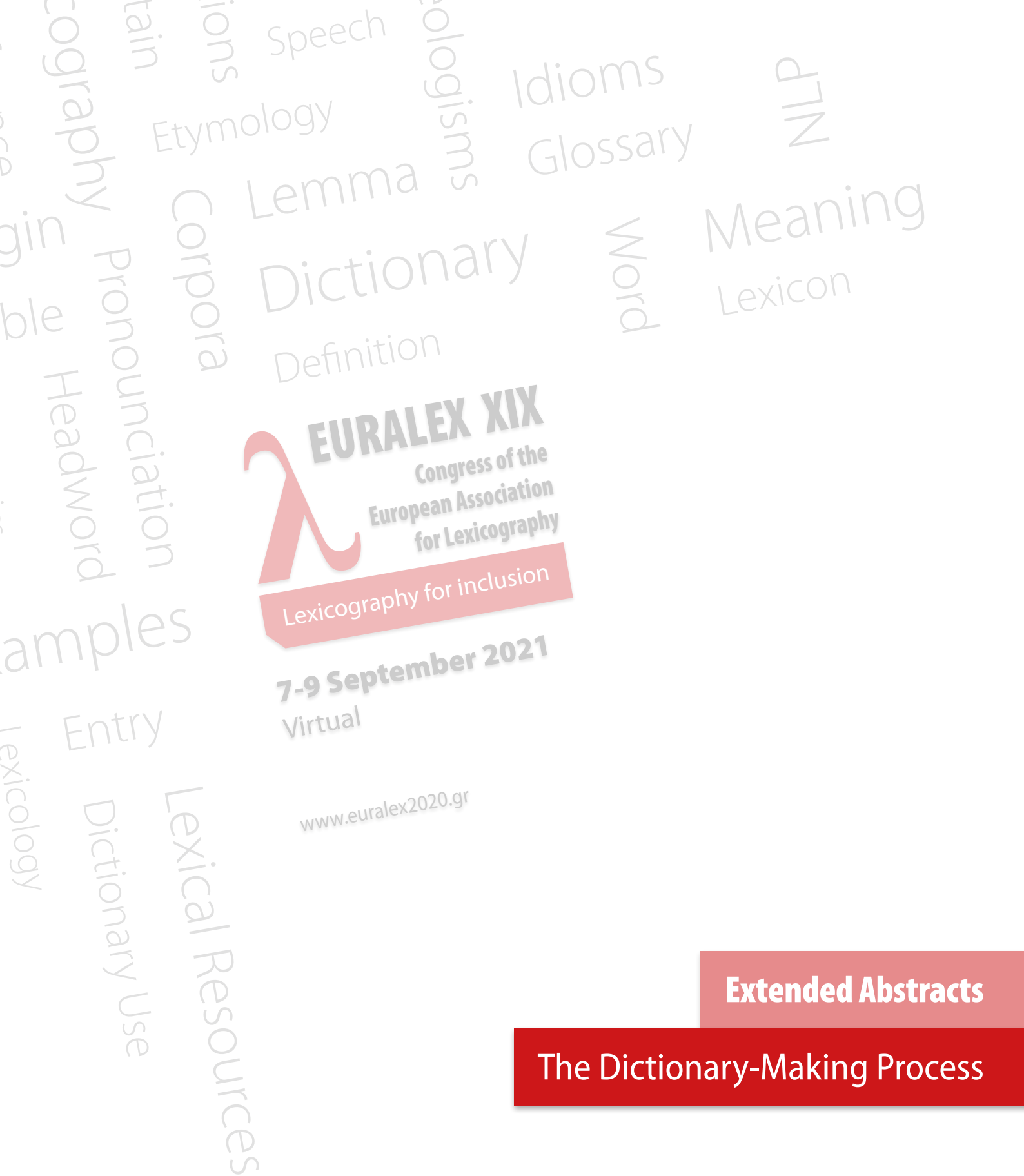
Even though users would not be required to understand the detail of how such a system works, they should nevertheless understand the basic principles. For this, dictionary literacy, as a subset of digital and information literacy, is essential.

Contextualization is an essential goal that needs to be addressed in future research and we concur with the final comment of the article by Tarp and Gouws (2019): "We have work to do" (p. 266). To attain this goal would need extensive research and collaboration between lexicographers, information scientists, computer scientists and others, and an easy solution is not envisaged. We do not pretend to propose definitive solutions but intend to contribute to this important discussion for the development of smart e-dictionaries.

Keywords: Information Needs, Contextualization, Dictionary Consultation Process.

References

Bothma, TJD and Prinsloo, DJ. 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries. Lexicographica 29(1): 165-198.
Gouws, RH and Tarp, S. 2017. Information Overload and Data Overload in Lexicography. International Journal of Lexicography 30(4): 389-415.
Lettner, K. 2020. Zur Theorie des lexikographischen Beispiels. Berlin: De Gruyter.
Tarp, S and Gouws, RH. 2019. Lexicographical Contextualization and Personalization: A New Perspective. Lexikos 29: 250-268. DOI: https://doi.org/10.5788/29-1-1520.



Extended Abstracts

The Dictionary-Making Process

'Help, my XML is too complex!' – the problem of excessive structural markup in dictionaries

Michal Měchura

Natural Language Processing Centre, Masaryk University, Brno, Czech Republic

Abstract

Dictionary encoding is the activity of taking an inventory of lexicographic object types such as headword, part-of-speech label, sense and translation, and expressing them formally in a data serialization language such as XML. But the use of XML for dictionary encoding often leads to excessively complex markup, with multi-layered embedding of elements inside other elements inside yet more elements. The following code shows how a pair of translations would typically be encoded in a bilingual dictionary (adapted from the New English–Irish Dictionary [Ó Mianáin and Convery 2014]).

Code sample 1

```
<translations>
  <translationContainer>
    <translation>leasú</translation>
    <pos>n-masc</pos>
  </translationContainer>
  <translationContainer>
    <translation>athchóiriú</translation>
    <pos>n-masc</pos>
    <usage>formal</usage>
  </translationContainer>
</translations>
```

The only XML elements here that contain actual human-readable information are <translation> (the translation's wording), <pos> (its part of speech) and <usage> (its usage label). The remaining XML elements are purely structural, used for grouping other elements together. Arguably, their presence here distracts a human XML reader (and even more so, a human XML writer) from lexicographic information which is otherwise simple and could be expressed more economically in some other (not yet existent) serialization language such as the pseudo-code in the following code sample.

Code sample 2

```
translation: leasú
  pos: n-masc
translation: athchóiriú
  pos: n-masc
  usage: formal
```

The distracting presence of purely structural elements in lexicographic XML is often acknowledged as an inconvenience in e-lexicographic circles informally but, to the author's knowledge, no serious attempts have been made yet to analyze or solve it.

We can define *purely structural markup* as such XML elements which contain no text nodes as their direct children: all their child nodes are other XML elements. Purely structural elements tend to be called *groups, containers* or *blocks* in the entry schemas of various dictionaries. For example, the entry schema for the DANTE project (Atkins, Kilgarriff and Rundell 2010) consists of elements such as <CollocGp> (collocate group) as a wrapper for a sequence of one or more collocates, <CollocCont> (collocate container) as a wrapper for a single collocate along with additional information about it (usage labels, example sentences, translations etc.) and finally <COLLOC> as a wrapper for the actual collocate (a text node). The first two of these three element types are purely structural. Broadly speaking, we tend to find two patterns of purely structural markup in lexicographic XML.

List pattern. The first kind is a pattern in which a parent element wraps a sequence of child elements which are all of the same type, such as <CollocGp> for a series of collocates in Dante, or <translations> for a series of translations in Code sample 1. They are almost always unnecessary in the sense that they

convey no useful information. They are there because the designer of the entry schema probably thought it 'logical' to group elements of the same type under a common parent element. But the usefulness of this grouping is debatable: the group thus created does not seem to represent any lexicographic fact which a lexicographer might want to communicate to the dictionary's end-users. Unnecessary grouping of this kind can be found in XML outside lexicography too and tends to be advised against in XML styleguides (eg. Ogbuji 2004).

Headed pattern. The second kind is a pattern in which a parent element wraps child elements of different types, one of which can be considered the head and the others can be seen as providing additional information about the head. An example is <translationContainer> in Code sample 1 which can be said to be headed by <translation>, while the other children <pos> and <usage> provide additional information about the head. In DANTE, a similar example is <CollocCont> which is headed by <COLLOC> (the actual collocate) while other child elements of <CollocCont> provide additional information about the head (usage labels, example sentences, translations etc.). Unlike the list pattern, the headed pattern cannot be explained away as a bad practice. Its purpose is to encode a lexicographic fact which the lexicographer wants to communicate to the end-user: for example, which <pos> element modifies which <translation> element. The purely structural <translationContainer> element is a tool for encoding that fact.

In this paper I will focus on the headed pattern of purely structural markup and dicuss it in depth. I will identify several subtypes of this pattern and show examples from real-life dictionaries. I will discuss whether it is possible to encode the headed pattern in XML *without* recourse to purely structural markup (for example by using XML attributes), but I will reach the conclusion that some amount of purely structural markup is unavoidable and that the problem is unsolvable, as long as one insists on using XML.

Secondly, I will evaluate other popular data serialization languages such as JSON a YAML and show that they, too, lead to excessive structural markup. What XML, JSON and YAML have in common is that they take no account of the inherent *headedness* of many lexicographic information objects such as collocations, example sentences and translations. In XML and other languages, the only way to encode the relation between a head (such as <translation>) and its modifiers (such as <pos> or <usage>) is to wrap them inside a common parent (such as <translationContainer>), which unavoidably leads to the proliferation of excessive structural markup we see in lexicographic XML everywhere.

An ideal lexicographic serialization language would be one which respects the inherent headedness of lexicographic data. In conclusion I will propose the creation of such a *lexicographic lightweight markup language* (à la Benko 2018). The language would read similarly to the pseudo-code in Code sample 2 and could be used in dictionary writing systems either as a replacement for XML or as a non-persisten surface representation on top of XML in the fashion of *Invisible XML* (Pemberton 2013).

Keywords: XML, dictionary encoding, dictionary editing, lexicographic lightweight markup language

References

Atkins, B. T. S.; Kilgarriff, A.; Rundell, M (2010). Database of ANalysed Texts of English (DANTE): the NEID database project. In: Proceedings of the Fourteenth EURALEX International Congress, EURALEX 2010.

Benko, V. (2018). In Praise of Simplicity: Lexicographic Lightweight Markup Language. In: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.

Ogbuji, U. (2004). Considering container elements: When to use elements to wrap structures of other elements. In: Principles of XML design, IBM, https://www.ibm.com/developerworks/library/x-contain/index.html (accessed 6 September 2019).

Ó Mianáin, P.; Convery, C. (2014). From DANTE to Dictionary: The New English-Irish Dictionary. In: Proceedings of the Sixteenth EURALEX International Congress, EURALEX 2014.

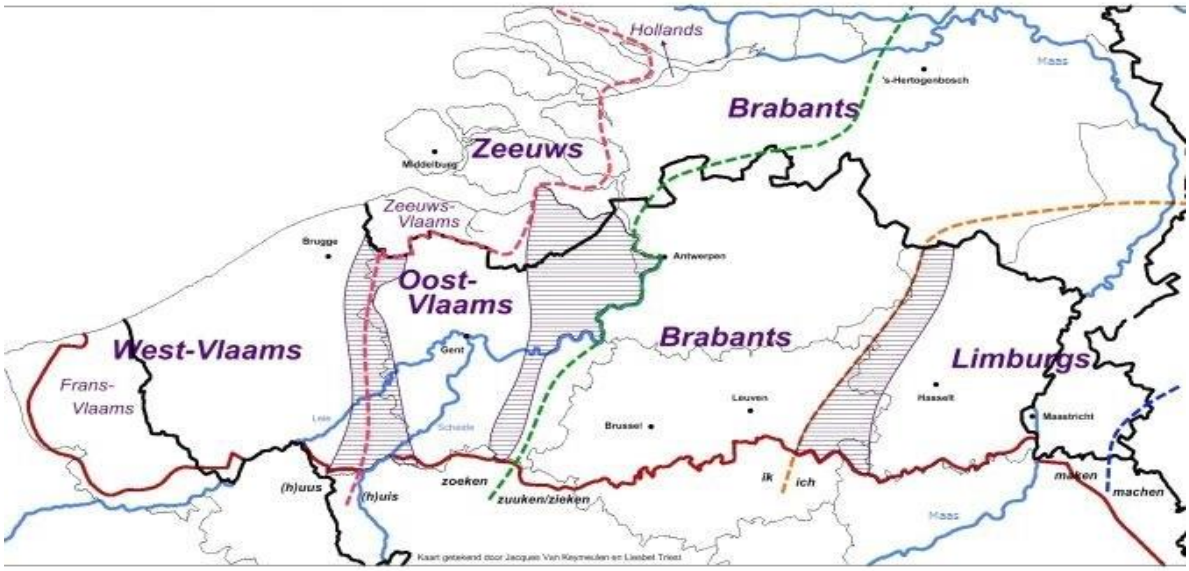
Pemberton, S. (2013). Invisible XML. In: Proceedings of Balisage: The Markup Conference 2013. Balisage Series on Markup Technologies, vol. 10.

The Dictionary portal of the Southern Dutch Dialects

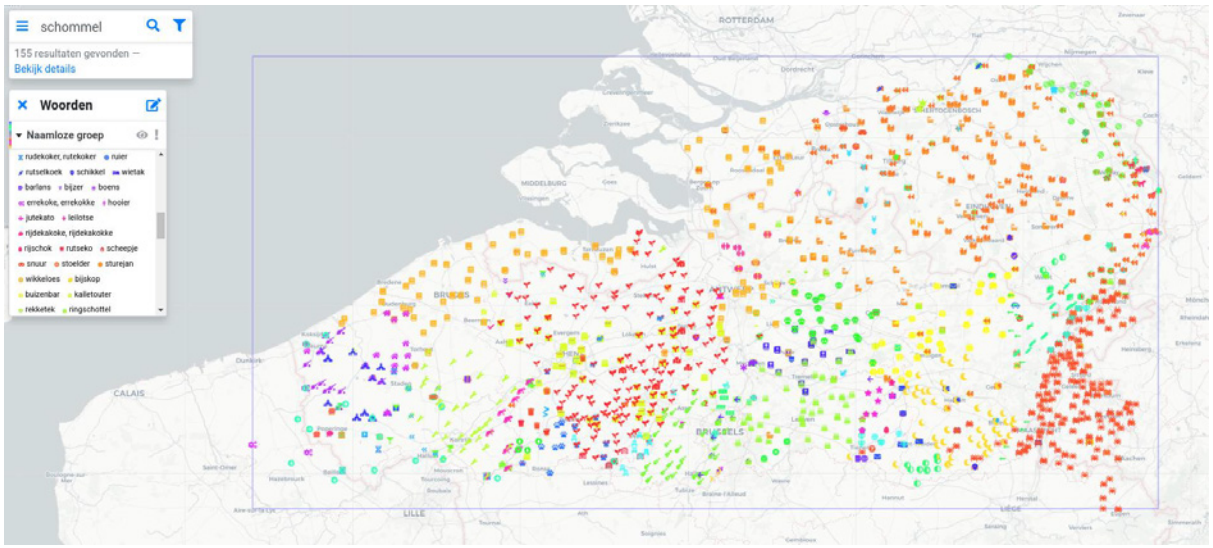
Veronique De Tier[1], Katrien Depuydt[1], Jesse de Does[1], Tanneke Schoonheim[1], Jacques Van Keymeulen[2], Sally Chambers[2]

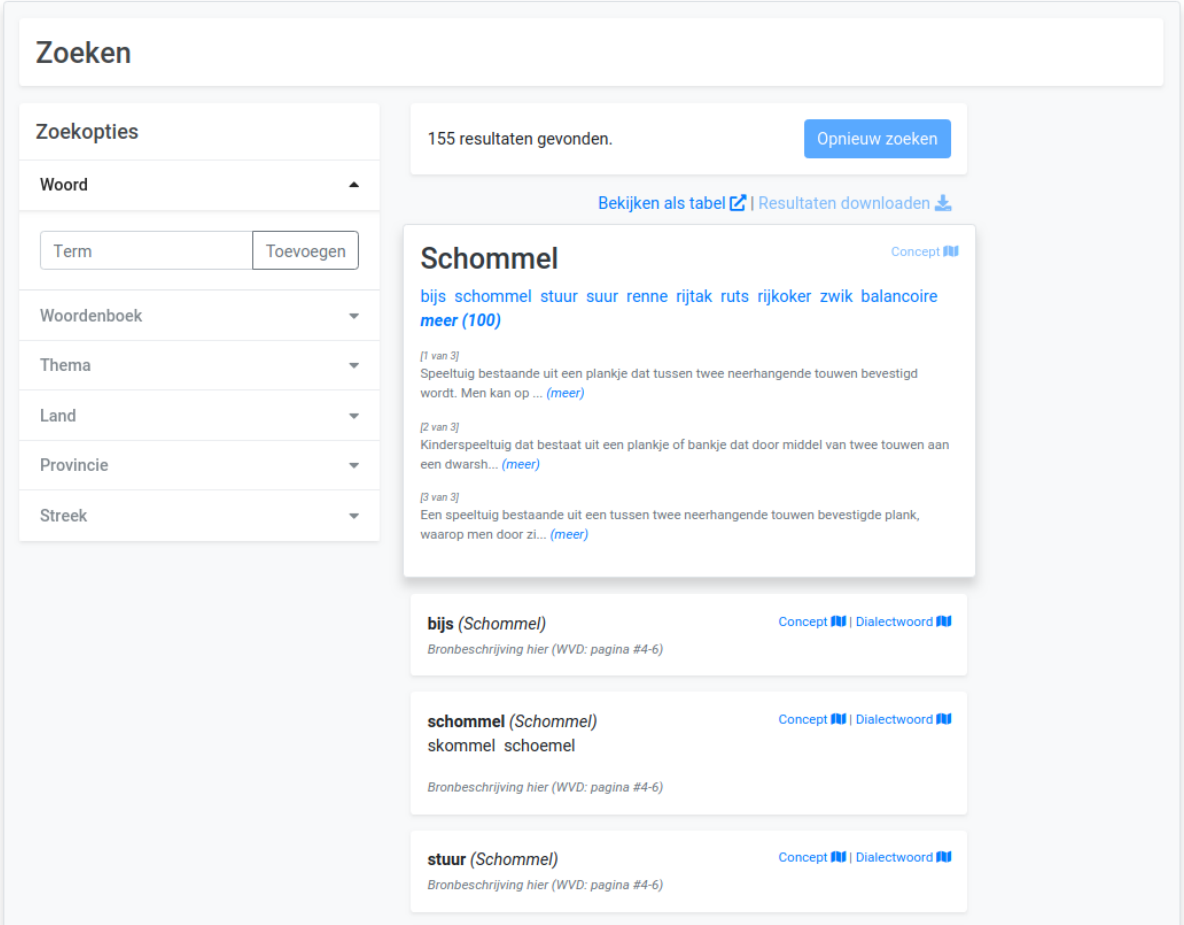
[1] *Ghent University, Belgium*
[2] *INT, Netherlands*

Abstract
The southern Dutch dialect area consists of four dialect groups, partly found in the Netherlands, in Belgium and in a small part of France: the Flemish, Brabantic, Limburgian and the Zeeland dialects. Each of these dialects has been described in a separate dictionary. The Flemish, Brabantic and Limburgian dialect dictionaries are more or less similarly constructed onomasiological dialect dictionaries (thematically arranged), whereas for the Zeeland dialect, it is a semasiological dictionary (alphabetically arranged). In 2016 an infrastructure project started to combine these dictionaries in one dictionary portal: the database of the Southern Dutch Dialects (DSDD). This project was initiated by Ghent University and undertaken in close collaboration with the Dutch Language Institute (Leiden).



The aim of the project was to work with a pilot dataset of 1500 concepts. Principles, strategies and tools for data alignment were developed and a prototype for online application, the dictionary portal. As far as the design of the portal was concerned, inspiration was found in the online versions of the individual dialect dictionaries (e-WVD, e-WBD, e-WLD), in similar projects like Verba Alpina (Switzerland, Italy, Germany) and Regionalsprache.de (REDE - Germany), and in dedicated workshops with partners in the field of geography/cartography. The portal application will be demonstrated here.

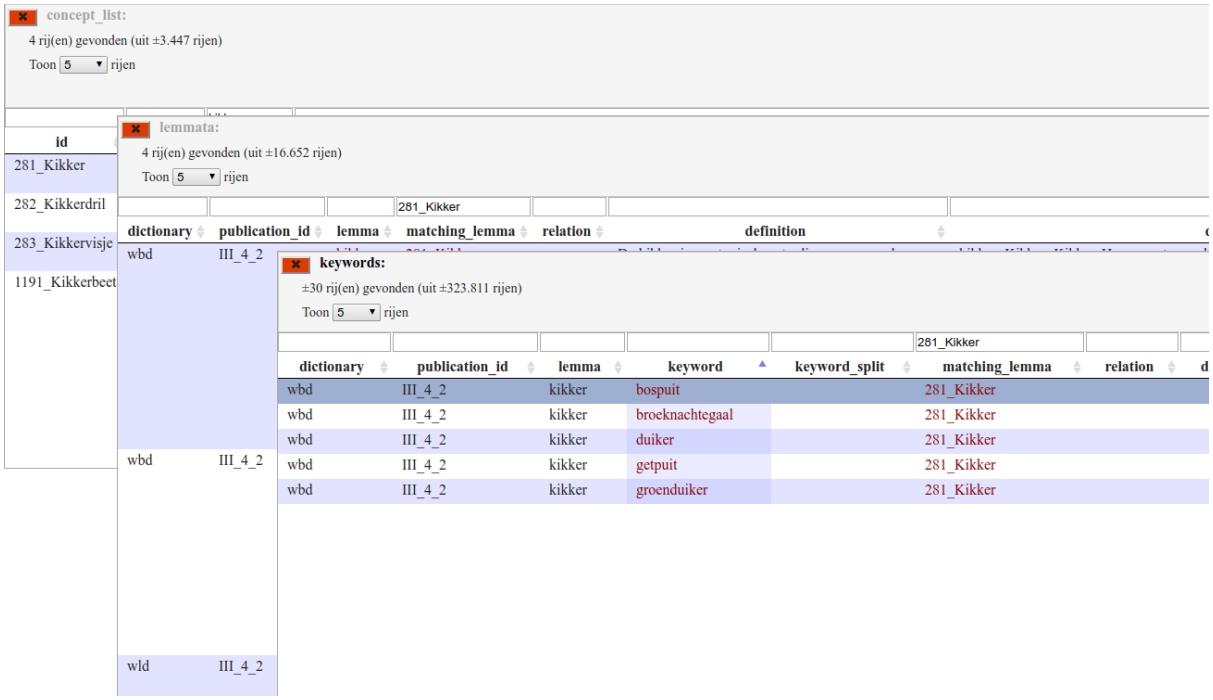




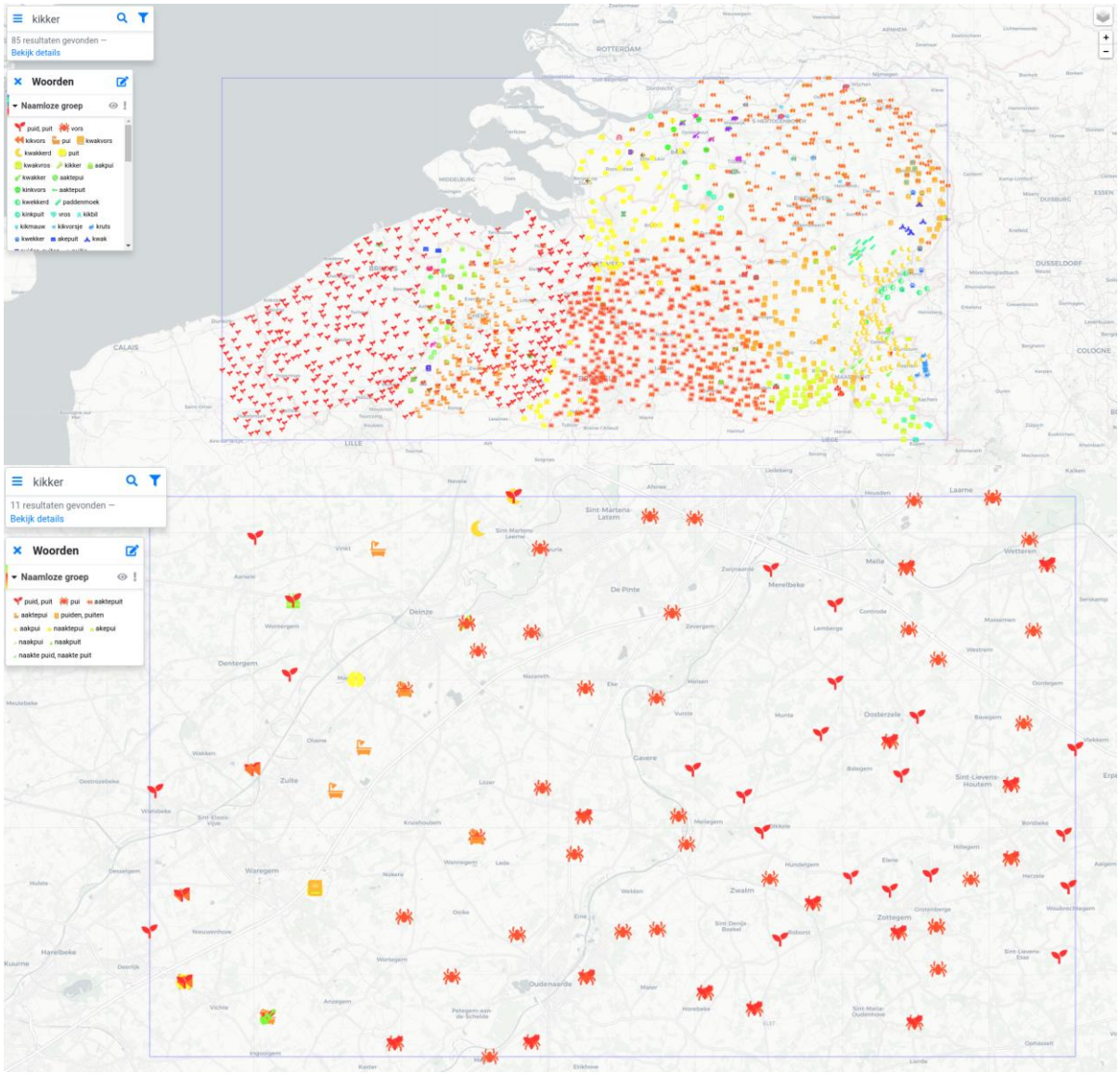
Before being able to start with the data alignment work, some effort had to be put into correcting the original data from the three dictionaries. As often is the case with dictionary data that was not originally conceived for digital use, a number of smaller and also more significant corrections had to be made; for instance, standardising the spelling of the keywords (*paddestoel* / *paddenstoel*) and combining diminutives with their simplex (*schommel* / *schommeltje*).

Concept linking was done by adding an additional DSDD concept layer on top of the concept layer of each individual dictionary. The DSDD concept was determined after careful analysis of the concepts in each dictionary. In many cases the mapping of the concepts was quite straightforward, but in other cases a choice had to be made. For instance *muil* (WVD), *bek* (WBD) and *bek* (WLD) were all put under the

concept *bek* ‘animal’s mouth’. Sometimes hyponyms and hypernyms had to be combined to achieve a better coverage, for example, *mier* and *zwarte mier* were combined under the concept *mier* ‘ant’. The data work was done using the data processing tool Lex’it, developed by the Dutch Language Institute.



The results of the pilot are available in the Dictionary portal of the Southern Dutch Dialects (DSDD). The portal is intended for everyone with an interest in dialects. The portal aims to be accessible for both a wider audience and dialect researchers. Dialect researchers have a lot to gain from the opportunities the portal offers, because of its extensive search options and the availability of an application to create interactive maps with the very precise location information available in the dialect data.
The portal’s search engine enables searching for individual dialect words (dialect word *puit,* belonging to the concept *kikker*), for dialect words of a certain village or region (all dialect words from the village Lozer; all dialects from the city of Ghent) and all dialect words belonging to a certain concept (*kikker*) or theme (fauna). Filters are available to make the search query as accurate as possible.
Based on the search results, interactive maps can be compiled, visualising the data in its geographical setting. The results can be grouped in various ways, using for instance frequency or etymologically related entries (for instance a group ‘kik’ containing *kikker* and *kikvors*, and/or a group ‘vors’, containing *fros, vors* and *kikvors*). The maps are flexible enabling the user, for example, to select the required level of detail to be visualised for the chosen region. A set of symbols and colours is available and can be arranged according the user’s needs. The area map can be enlarged or reduced in size and printed or downloaded.



This portal is the result of a pilot aimed to demonstrate the opportunities of the combined dialect data of the Southern Dutch Dialects (Flemish, Brabantic and Limburgian). The project will be continued at the Dutch Language Institute. Apart from the addition of the remaining data, work will also be done to improve the search functionalities and the data manipulation. New dialect areas will also be added, beginning with the Zeelandic dictionary, which will complete the coverage of the Southern Dutch Dialects area. However, the addition of this final dictionary will be challenging since this is not an ononmasiological dictionary like the three dictionaries already available in the portal but a semasiological dictionary. Once we have developed a strategy to add this dictionary, we intend to add dictionaries from other Dutch dialect areas.

Keywords: Dialects, Dutch, Geo-mapping, Dictionary portal, Data visualisation.

1. *Instituut voor de Nederlandse Taal (Dutch Language Institute), Leiden.*
2. *Universiteit Gent (Ghent University).*

Mapping domain labels of dictionaries

Ana Salgado[1,2], Rute Costa[1], Toma Tasovac[3]

[1] *NOVA CLUNL Universidade NOVA de Lisboa, Lisbon, Portugal*
[2] *Academia das Ciências de Lisboa, Lisbon, Portugal*
[3] *Belgrade Center for Digital Humanities, Belgrade, Serbia*

Abstract

The purpose of this paper is to compare and analyse the use of domain labels in three large scholarly dictionaries – *Dicionário da Língua Portuguesa Contemporânea* (DLPC), published in 2001 by the Academia das Ciências de Lisboa (ACL); the 23th online edition of *Diccionario de la lengua española* (DLE), published by Real Academia Española (RAE); and the 9th online edition of *Dictionnaire de l'Académie Française* (DAF), a work in progress – in order to a) highlight the commonalities and differences in their editorial practices and approaches to knowledge organisation; b) report on a mapping exercise for a particular domain (GEOLOGY) which can serve as a test case for establishing procedural rules for the alignment of domain labels in general language dictionaries. We show how "meta-labels" can be used to optimise the alignment of specialised senses in lexicographic works.

General dictionaries register, describe and define the specialised senses of lexical items, or terms, specific to different areas of knowledge. As a result of technological changes, the evolution of society, and globalisation, the number of terms found in dictionaries entry has increased (Wiegand, 1984, Boulanger and L'Homme, 1991 and Ahumada, 2002). The labels assigned to these specialised senses are called "domain labels". As markers which identify the specialised field of knowledge in which a lexical unit is mainly used (Salgado et al., 2019), domain labels can serve multiple functions: aiding lexicographers by providing specific information and by identifying specialised lexica in general language dictionaries that can serve as terminology control mechanisms; facilitating user searches by grouping lexical items according to a field so that the user can determine beforehand if the complete lexicographic article is relevant for them; facilitating end-user word sense disambiguation tasks; facilitating terminology extraction in diverse languages; enhancing machine translation and NLP projects.

A domain can be the name of a field in which a specific knowledge area is developed (GEOLOGY) or the specific object of the knowledge area (SHOEMAKING). Lexicographers often make subjective assignments according to a certain tradition they subscribe to (Ptaszynski, 2010, p. 413). For example, the dictionaries we analysed contain labels for domains such as CYNEGETICS (DLPC, DLE) and HUNTING (DAF) but not for MANAGEMENT or TOURISM.

All three Academy dictionaries lack explicit explanatory information regarding their labelling practices (Salgado et al., 2019). Our previous work on DLPC (Salgado and Costa, 2019) has already detected the problematic use of: i) domains with multiple labels, for example, football terms were found to be classified under the SPORT and FOOTBALL labels in DLPC (e.g. *líbero* [sweeper] in SPORT and *lateral* [back] in FOOTBALL); ii) unlabeled equivalent headwords, for example, *paleozóico* [palaeozoic] *adj.* is unlabeled and *primário* [primary] *adj.*, a synonym, appears with a GEOLOGY label; iii) combinations of labels referring to closely related domains, such as *antracite* [anthracite] being associated with both MINERALOGY) and GEOLOGY or *glaciar* [glacier] being associated with both the GEOLOGY and GEOGRAPHY domains. Such inconsistencies can lead to numerous issues that complicate the sharing, aligning, and linking data.

Atkins and Rundell (2008) argue that instead of conceiving "a totally 'flat' (non-hierarchical list of domains)", "it is more practicable to try to build a domain list with a certain hierarchical structure" (p. 184). Applying previously organised hierarchical structure is advantageous both when composing and

when editing a lexicographic resource because it helps the lexicographer control the terminology. The geology domain was reorganised to illustrate examples of existing frameworks (WordNet Domains Hierarchy[1]; Dewey Decimal Classification[2]).

In this paper, we will present the theoretical framework, a threefold methodology and the analysis of the chosen domain:

1) Theoretical framework: The theoretical framework upon which this research is based is summarised to provide related background information (e. g. assumptions about domain labelling by Atkins and Rundell (2008); labelling classifications by Hausmann (1989), followed by Svensén (2009); works on WordNet domains by Magnini and Cavaglià (2000), Bentivogli et al. (2004), Gella et al. (2014)), and to argue for a conceptual modelling based on ISO standards (704:2009; 1087:2019) for terminology.
2) The methodology applied in this research:
   i) Monolingual dictionaries were chosen due to their highly discursive properties. Academy dictionaries were selected for study due to their authoritativeness.
   ii) Datasets were compiled manually from dictionary abbreviation lists. Three hundred eighty-seven multilingual domain labels were collected. There were 184, 74, and 237 domain labels in DLPC, DLE, and DAF, respectively. Generic domains and subdomains coexisted. We noted the case of MATHEMATICS and its sub-domains ALGEBRA (DLPC, DAF), ARITHMETIC (DLPC, DAF), GEOMETRY (DLPC, DLE, DAF) and TRIGONOMETRY (DLPC) or STATISTICS (DLE, DAF). In our comparison, a flagrant imbalance in the number of domains was found: the DLE contains generic domains alone, whereas the DLPC and DAF register multiple subdomains and even multiple labels for the same or very similar domains (e.g. COURSES DE CHEVAUX and COURSES HIPPIQUES [horse races] in DAF).
   iii) In order to systematise the labels and to detect overlapping, the compiled domain label lists were compared. The DLPC list was set as baseline, against which the DLE and DAF counterparts were compared. DLE and DAF were also separately compared. Domain labels were manually mapped using semantic properties such as "exact" and "related" (to a generic domain) and "none". The equivalent English term was assigned as the "meta-label" of the corresponding domain (Table 1 and Appendices).

DLPC	RELATION	DLE	RELATION	DAF	METALABEL
Acústica	EXACT	acústica	EXACT	Acoustique	acoustics
Aeronáutica	EXACT	aeronáutica	EXACT	Aéronautique	aeronautics
Agricultura	EXACT	agricultura	EXACT	Agriculture	agriculture
Anatomia	EXACT	anatomía	EXACT	Anatomie	anatomy
Antropologia	EXACT	antropología	EXACT	Anthropologie	anthropology
Arqueologia	EXACT	arqueologia	EXACT	Archéologie	archeology
Arquitectura	EXACT	arquitectura	EXACT	Architecture	architecture
Astrologia	EXACT	astrologia	EXACT	Astrologie	astrology
Astronomia	EXACT	astronomía	EXACT	Astronomie	astronomy

TABLE 1 – A fragment of domain labels with an "exact" correspondence – 61 domains were mapped to an equivalent domain.

3) Domain analysis: Example entries are presented from the domain GEOLOGY. Using the DLPC as the baseline, this domain was found to have branches that were considered subdomains of a generic domain. GEOLOGY include CRYSTALLOGRAPHY, MINERALOGY, and PALAEONTOLOGY. The corresponding dictionary definitions for each of these terms were compared to clarify, if possible, the underlying reasoning for these subdivisions.

The multilingual domain map constructed in this study will support future standardisation efforts. Standardisation of the domain labelling process and associated encoding tasks are required in order to achieve structured, organised, accessible, and interoperable lexical resources.

[1] http://wndomains.fbk.eu/hierarchy.html
[2] http://www.gutenberg.org/files/12513/12513-h/12513-h.htm

APPENDICES

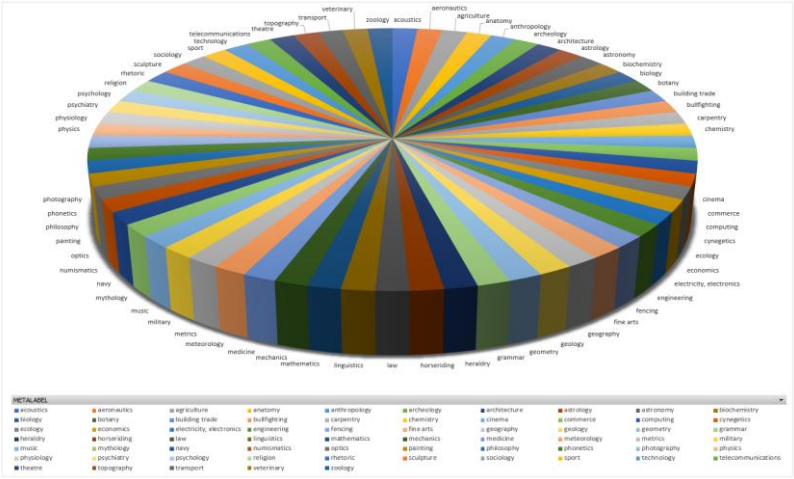


Fig. 1 DLPC vs. DLE – Correspondence between domain labels (65)

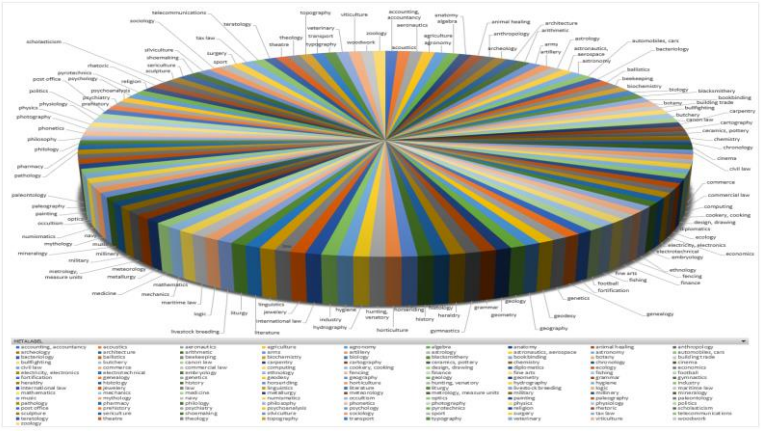


Fig. 2 DLPC vs. DAF – Correspondence between domain labels (136)

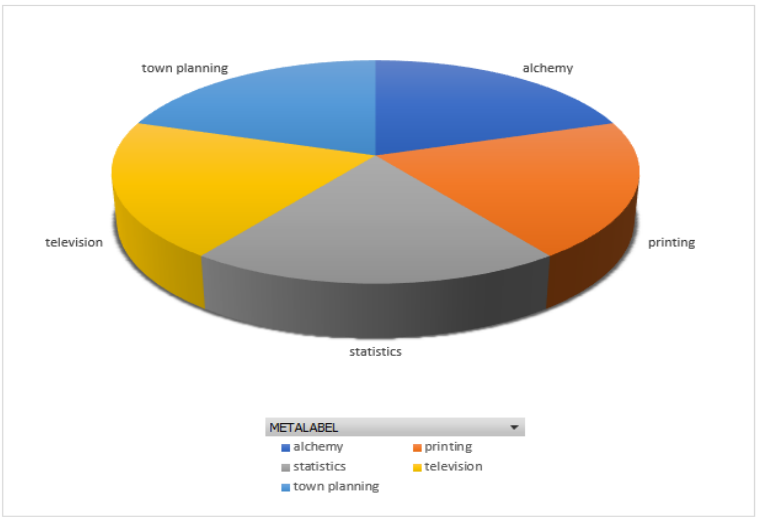


Fig. 3 DLE vs. DAF – Correspondence between domain labels (5)

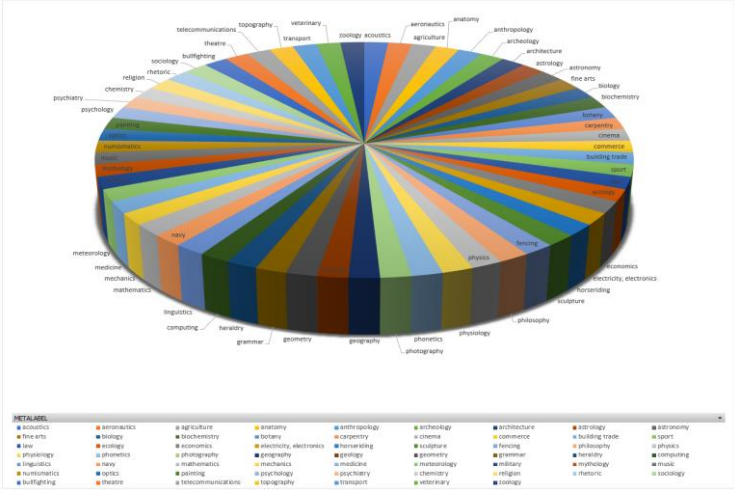


Fig. 4 DLPC vs. DLE vs. DAF – Correspondence between domain labels (61)

Acknowledgements

Research was financed by Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020, and by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 731015 (ELEXIS). The authors would also like to thank ILex (Institute of Lexicography of the RAE) for granting us access to their data and for hosting Ms. Salgado's 3-week research visit within the scope of an ELEXIS grant. We would also like to thank the Académie Française for sharing domain lists from the current version of their digital dictionary (February 2020, 9th edition) – work completed up to letter S (savoir).

Keywords: Academy of Sciences dictionaries, domain label, Lexicography, Terminology

References

Dictionaries

Dicionário da Língua Portuguesa Contemporânea. 2001. João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa and Editorial Verbo. New digital edition under revision, Ana Salgado (coord.).

Diccionario de la Lengua Española (24.ª ed.). Real Academia Española, 2001–2020, www.rae.es/rae.

Dictionnaire de l'Académie Française (9.ª ed.). Académie Française, 2020, http://www.dictionnaire-academie.fr/.

Other literature

Ahumada, I. (ed.) (2002). Diccionarios y lenguas de especialidad. Jaén: Universidad de Jaén.

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

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In COLING 2004, Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, 101–108.

Boulanger, J. C. and L'Homme (1991). Les techonolectes dans la pratique dictionnaire générale: quelques fragments d'une culture. In Meta, vol. 36(1), 23–40.

Gella, S., Strapparava, C., and Nastase, V. (2014). Mapping WordNet Domains, WordNet Topics and Wikipedia Categories to Generate Multilingual Domain Specific Resources. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Iceland: Reykjavik, 1117–1121.

Hausmann, F. J. (1989). Die Markierung in eineim allgemeinen einsprachigen Wörterbuch: eine Übersicht. In F. J. Hausmann, O. Reichmann, H. E. Wiegand and L. Zgusta (eds.). Wörterbücher. Ein internationales Handbuch zur Lexikographie, 649–657. Berlin: Walter de Gruyter.

International Organization for Standardization. 2009. ISO 704: Terminology Work – Principles and Methods. Geneva: ISO.
International Organization for Standardization. 2019. ISO 1087–1: Terminology Work – Vocabulary – Part 1: Theory and Application. Geneva: ISO.
Magnini, B., and Cavaglià, G. (2000). Integrating Subject Field Codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (eds.). Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May–2 June 2000, 1413–1418.
Ptaszynski, M. O. (2010). Theoretical Considerations for the Improvement of Usage Labelling in Dictionaries: A Combined Formal-Functional Approach. In International Journal of Lexicography, Volume 23, Issue 4, December 2010, 411–442. DOI: https://doi.org/10.1093/ijl/ecq029.
Salgado, A., and Costa, R. (2019). Marcas temáticas en los diccionarios académicos ibéricos: estudio comparativo. RILEX. Revista Sobre Investigaciones léxicas, 2(2), 37–63. DOI: https://doi.org/10.17561/rilex.v2.n2.2.
Salgado, A., Costa, R., and Tasovac, T. (2019). Improving the consistency of usage labelling in dictionaries with TEI Lex-0. Lexicography ASIALEX 6, 133–156. DOI: https://doi.org/10.1007/s40607-019-00061-x.
Svensén, B. (2009). A Handbook of Lexicography: The Theory and Practice of Dictionary Making. Cambridge: Cambridge University Press.
Wiegand, H. E. (1984). On the Structure and Contents of a General Theory of Lexicography. In R.R.K. Hartmann (ed.), LEXeter'83. Tübingen: Max Niemeyer, 13–30.

Using log files to improve the Brazilian Portuguese Olympic Dictionary

Bruna da Silva[1,2], Rove Chishman[1], Gilles-Maurice de Schryver[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

Abstract

1.Online dictionaries may be seamlessly 'tracked' through a study of server log files. These do not even need to be customised as one can get a long way by simply crunching the data contained in default Web analytic reports. A recent example of the power of doing so may be found in de Schryver et al. (2019). Both AWStats (2000-2021) and Google Analytics (2005-2021) are available for the *Dicionário Olímpico* (Chishman and colleagues 2016), an online dictionary based on Frame Semantics (Fillmore 1982), but with adaptations (da Silva and Chishman 2018; Chishman et al. 2019). At present, mainly the data from AWStats has been analysed, using a sample of the 'top 1000 pages/URLs per month' for the most popular months overall – being August, September and November in the bracket 2016-2019. Even though the study is presently limited in scope, some surprising results are already apparent.

2.We take a bird's-eye view of the logged data. Firstly, with regard to the 40 SPORTS, known as 'superframes', one notices that not all of the sports appear in the top 1000 of the most-frequently visited pages per month. Strangely, over the years, there seems to be a concentration on ever fewer sports. This is so despite the fact that the dictionary is increasingly popular, meaning that over the years, more and more pages are viewed for a smaller number of sports. Not only are several sports missing, not all sports are as frequently searched for, with one sport in particular standing out: 'rhythmic gymnastics'. Page views for this sport are 2.5 times more frequent than those for the next-most-frequent sport, 'volleyball'. From then on, the distribution of the sports appears to be Zipfian.

When it comes to the SCENARIOS ('frames' in Fillmore-speak), of which there are currently 780 in all, only about half are ever seen in the top 1000. The number of scenarios visited is stable over the years, i.o.w. it is not the case that, e.g., more scenarios are being studied at the expense of, say, the description of a sport (on the level of the 'superframe') or at the expense of the definition of a specific term (on the level of the 'word', or 'lexical unit' in Fillmore-speak).

The most interesting type of analyses concern case studies of the various SCENARIOS per SPORT. Doing so allows one to track which frames remain popular through time. For the sport *ginastica de trampolim* (trampolining), for instance, the most popular scenarios from 2017 to 2019 remain *equipamentos 3* (equipment #3) and *saltos* (jumps). Note that the year 2016 is always an outlier, as that is the period when the bulk of the dictionary was being compiled. For many pages of the dictionary, the visits are actually most frequent then. A case in point is the sport *canoagem slalom* (canoe slalom), for which a total of 15 scenarios were visited in August, September and November 2016, but only one to two scenarios thereafter (i.e., in 2017, 2018 and 2019). The dictionary compiler thus 'viewed' far more scenarios (and far more frequently at that) than all the dictionary users over the next three years combined!

Lastly, on the level of the (currently 3930) WORDS — at which point one has reached the typical dictionary-entry level — the logs reveal which words the dictionary users are most interested in, and how that interest changes with time.

3.A first major finding concerns the fact that most of the dictionary is actually hardly seen/used. This begs the question: What is the most user-friendly attitude which the dictionary compilers could adopt? There are two possible ways to look at this. On the one hand, they could engage in improving the sport entry (or do this for the top few sport entries) that is (are) already most popular by far, i.e. 'rhythmic gymnastics', and thus really strive to make that sport entry (or those few sport entries) as perfect as possible. This is in line with what modern social media do, whereby one is only fed with what one will like and agree with (which in turn leads to ever more navel-gazing and an increasingly polarised world). This wonderful/sad* [*select the appropriate word depending on one's standpoint] feature has actually already entered the field of lexicography: When, in 2019, typing *Dicionário Olímpico* into Google (1997-2021), the search engine immediately 'predicts' that one will actually want to go to *Ginástica*

Rítmica (Rhythmic Gymnastics). A detailed analysis of the various referrals will thus be in order to be able to judge whether dictionary users are still in control when they seek specialised encyclopaedic and lexical information, or whether our lives and the knowledge we are fed with is already 'controlled' by proprietary black boxes. On the other hand, the compilers could work on ways of encouraging users to consult the other 39 sports more frequently, through for instance a better development of the contents for these sports or via strategies such as "sport/scenario/word of the day".

Secondly, according to the Google Analytics data, users tend to start their searches at the dictionary homepage, while the number of page views decreases as one goes down the hierarchical structure of the dictionary. This finding in itself is not conclusive. Users could be satisfied with the encyclopaedic information found on the level of the sports and scenarios and not really be interested in the lexical analysis of meanings at the words. That would explain why they do not go deep(er) into the dictionary, and thus why they hardly explore the word level. Although from the point of view of the lexicographer this could be frustrating, since all the work on the word level would be 'lost', it would remain a good result. However, the massive drop-off rate at each level could also mean that users are not satisfied with the general information found in the sports and scenario pages.

How to know which one it is here? What is needed is for users to be able to express their (dis)satisfaction with their search experience. Feedback forms, e-mail, user interaction buttons and even emoticon-based Likert scales have already been employed in digital lexicography — see, respectively, de Schryver and Joffe (2004), Klosa and Gouws (2015), Liu (2017), and Efthimiou et al. (2019). We will show how these can be adapted to the *Dicionário Olímpico* to improve the user experience.

Keywords: digital lexicography, online dictionary, sports terminology, log files, user-friendliness

References

AWStats. 2000-2021. *Awstats – Open Source Log File Analyzer for Advanced Statistics. Available online at: https://awstats.sourceforge.io/.*

Chishman, R. and colleagues. 2016. *Dicionário Olímpico [Olympic Dictionary]. Available online at: http://www.dicionarioolimpico.com.br/.*

Chishman, R., A. N. dos Santos, B. da Silva and L. Brangel. 2019. *'Challenges and Difficulties in the Development of Dicionário Olímpico (2016)' In Kosem, I., T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek and C. Tiberius (eds), Electronic Lexicography in the 21st Century (Elex 2019): Smart Lexicography. Conference Proceedings. Sintra, Portugal, 1-3 October 2019. Brno: Lexical Computing, 622–641.*

da Silva, B. and R. Chishman. 2018. *'O Papel dos Frames na Organização de Dicionários Online [The Role of Frames in the Organization of Online Dictionaries].' Calidoscópio 16.3: 450–459.*

de Schryver, G.-M. and D. Joffe. 2004. *'On How Electronic Dictionaries Are Really Used' In Williams, G. and S. Vessier (eds), Proceedings of the Eleventh Euralex International Congress, Euralex 2004, Lorient, France, July 6-10, 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, 187–196.*

de Schryver, G.-M., S. Wolfer and R. Lew. 2019. *'The Relationship between Dictionary Look-up Frequency and Corpus Frequency Revisited: A Log-File Analysis of a Decade of User Interaction with a Swahili-English Dictionary.' GEMA Online® Journal of Language Studies 19.4: 1–27 + supplementary material online.*

Efthimiou, E., S.-E. Fotinea, T. Goulas, A. Vacalopoulou, K. Vasilaki and A.-L. Dimou. 2019. *'Sign Language Technologies and the Critical Role of Sl Resources in View of Future Internet Accessibility Services.' Technologies 7.1: 1–21.*

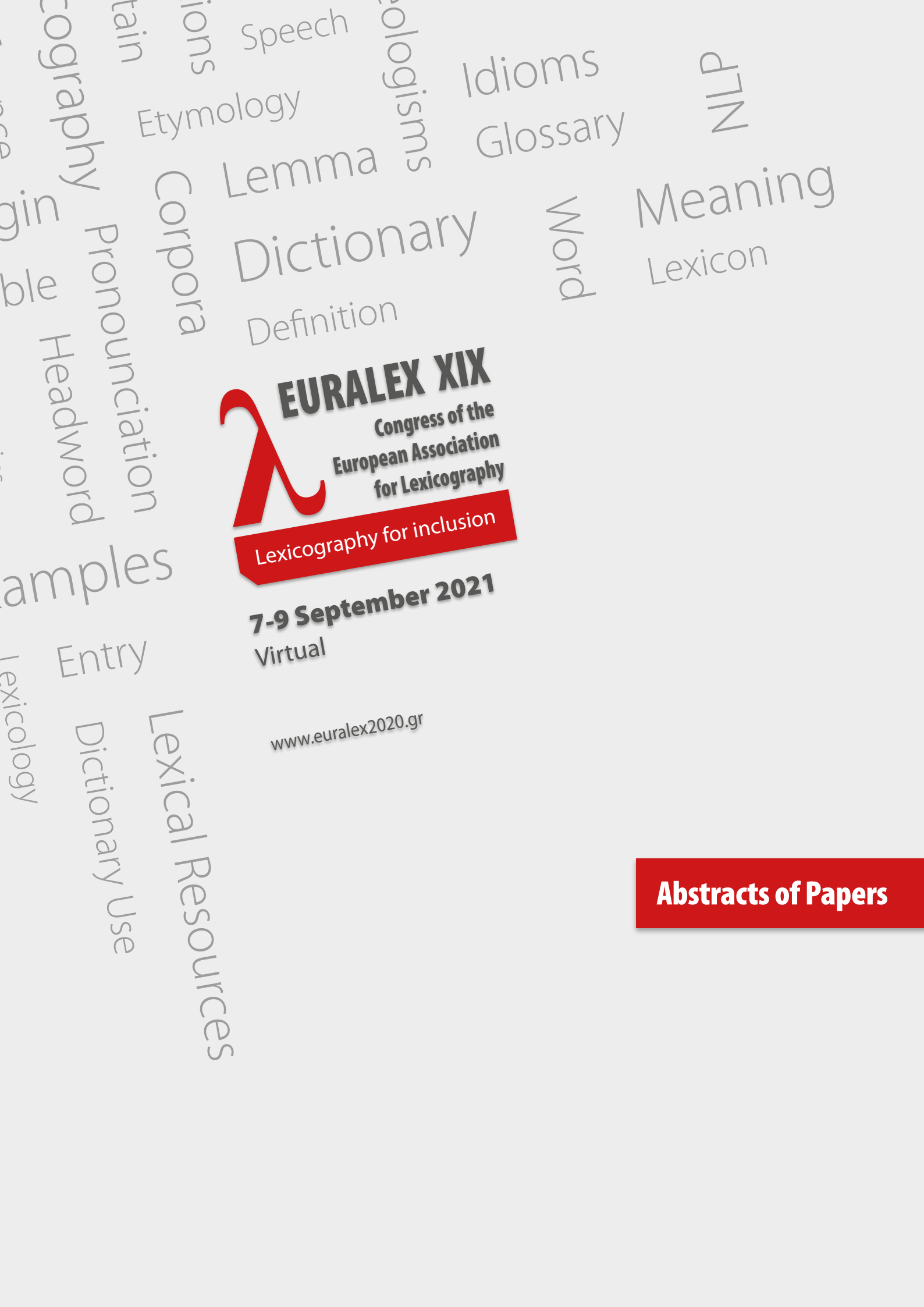
Fillmore, C. J. 1982. *'Frame Semantics' In The Linguistic Society of Korea (ed.), Linguistics in the Morning Calm. Seoul: Hanshin Publishing Co., 111–137.*

Google. 1997-2021. *Google – Web Search Engine. Available online at: https://google.com/.*

Google Analytics. 2005-2021. *Google Analytics – a Web Analytics Service Offered by Google. Available online at: https://marketingplatform.google.com/about/analytics/.*

Klosa, A. and R. H. Gouws. 2015. *'Outer Features in E-Dictionaries.' Lexicographica: International Annual for Lexicography 31: 142–172.*

Liu, X. 2017. *'Multimodal Exemplification: The Expansion of Meaning in Electronic Dictionaries.' Lexikos 27: 287–309.*



Abstracts of Papers

www.euralex2020.gr

Abstracts of Papers

Neology

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

Vassiliki Afentoulidou, Anastasia Christofidou

Research Centre for Scientific Terms and Neologisms, Academy of Athens, Greece

In this paper we discuss three main different views on the dictionarization 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 *stage of consolidation* (according to certain criteria) before their entering into the *establishment stage* (Kerremans 2015). The 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. facilitates the answer to the desideratum of the inclusion (or not) decision for 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 words, 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 ΝΕΟΔΗΜΙΑ research project at the Academy of Athens, comprising newspaper discourse spanning 5.4 years, we track the frequency development of selected new words which emerged during the Greek debt crisis and discuss their evolution in time.

Keywords: neology, dictionarization, corpora, consolidation, peakedness, dispersion

When neologisms don't reach the dictionary: occasionalisms in Spanish

Pedro Javier Bueno Ruiz

Universitat Pompeu Fabra, Barcelona, Spain

In this research we have analysed a type of neologism that does not become part of the usual speech of speakers and so has no chance of being incorporated into a general language dictionary: the *occasionalisms*. They have been defined by several authors as volatile language events difficult to detect and typified by their context-dependency and language creativity. With this research we want, on the one hand, to contribute with the description of the theoretical concept of *neologism* by defining one of the groups that are obtained following the lexicographical criterion and, on the other hand, to play a part in the definition of occasionalisms by finding regularities and their tendencies in formation processes and in formation rules. The data we have used belongs to *Neómetro* database and applying the methodology described below we have obtained an occasionalism corpus. After analysing the data, we can confirm that occasionalisms are not linguistic random acts and they are lexical units that fulfil an expressive function in a specific context.

Keywords: occasionalism, nonce-word, neology, lexicographical criterion

www.euralex2020.gr

Abstracts of Papers

Bi- and Multilingual Lexicography

The Online Dutch-Frisian dictionary in *Poarte ta it Frysk*

Eduard Drenth[1], Hindrik Sijens[1], Hans Van de Velde[1,2]

[1]*Fryske Akademy, Leeuwarden, Netherlands*
[2]*Utrecht University*

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, dictionary, translation, Dutch

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

Peter Meyer[1], Gerd Hentschel[2]

[1] *Department of Lexical Studies, Leibniz Institute for the German Language, Mannheim, Germany*
[2] *Institute of Slavic Studies, Carl von Ossietzky University of Oldenburg, Oldenburg, Germany*

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

Thematic Dictionary for Doctor-Patient Communication: The Principles and Process of Compilation

Igor Kudashe[1], Olga Semenova

Faculty of Information Technology and Communication Sciences (ITC), Tampere University, Tampere, Finland

Increasing internationalisation has resulted in a constantly growing need for community interpreting worldwide. Healthcare is one of the most challenging domains for community interpreters, as misunderstandings, especially those caused by the use of incorrect terminology, may cost lives. In this paper, we describe the process of planning and compiling the *Finnish-Russian Thematic Dictionary for Doctor-Patient Communication* aimed at professional community interpreters and university-level students of community interpreting. We start by describing the theoretical background of dictionary planning and analysing the information needs of the target groups. We then describe and justify the selection of dictionary sources as well as the mega-, macro-, and microstructure of the dictionary. The dictionary has been compiled using a tailored version of the in-house dictionary writing system MyTerMS. We briefly report the details of the technical implementation of the project. Finally, we reflect on some challenges encountered in this project as well as its future prospects. The dictionary can be further developed by increasing the volume of its disease-specific part, adding verbs and usage examples, and customising the electronic version for various target groups and purposes.

Keywords: medical dictionary, medical glossary, medical terminology, doctor-patient communication, community interpreting, healthcare interpreting

Arabic Loanwords in English: a Lexicographical Approach

Pierre Fournier, Rim Latrache

Université Sorbonne Paris Nord, Pleiade Research Unit 7338, Villetaneuse, France

This article deals with Arabic loanwords in English with a lexicographical perspective. To create a representative corpus of Arabic loanwords in English, items are extracted from the *Oxford English Dictionary*database (henceforth *OED*) with an etymological advanced search. Among the criteria affecting the etymological tagging, the concept of two languages of origin is probably the most difficult one for lexicographers to deal with. This study presents some of theissues lexicographers are faced with in the dictionary-making process. Following which, Arabic loanwords are classified according to semantics,along with the date oftheir first attestationin the *OED*database. This quotation dating work that the *OED*systematically performsisnot onlyanimmense task,but also an essential one,asit enables researchers to determine the semantic spheres these corresponding loanwords are integrated into,as well as the cultural relationship between Arabic-speaking countries and English-speaking countries.

Keywords: Arabic loanwords, contemporary English, dictionary-based study

A Morpho-Semantic Digital Didactic Dictionary for Learners of Latin at Early Stages

Manuel Márquez Cruz, Ana Mª Fernández-Pampillon Cesteros

Complutense University of Madrid, Spain

Diccionario Didáctico Digital de Latín (Digital Didactic Dictionary for Latin) is an open-access lexicographic work, created and hosted by the Universidad Complutense de Madrid. It is a bilingual dictionary (Latin-Spanish) that faces the challenge of providing Spanish-speaking students of Latin with an innovative lexicographic tool that facilitates the learning of basic Latin. Based on the theoretical principles of valences described in Tesnière's Dependency Grammar, Lyons' ontologies, and Fillmore's theory of semantic frameworks, the dictionary has been conceived as a linguistic tool to understand how Latin works at the semantic, morphological, and syntactic levels. It is a qualitative dictionary created ad hoc, used as an auxiliary tool to answer linguistic questions raised by an inductive didactic methodology that makes the Latin learning-teaching process available even to those students who lack basic syntactic knowledge. A structure of Hierarchical Faceted Categories that constitutes the lexicographic model provides various ways to access the lemmatised lexicon, facilitating intuitive navigation through the dictionary.

Keywords: Computational Lexicography, Learner's Dictionaries, Latin Lexicography, Bilingual Lexicography, Digital Dictionaries, Dictionary for special needs

Lemma selection and microstructure of a domain-specific e-dictionary of the mathematical field of graph theory

Theresa Kruse, Ulrich Heid

Institute of Information Science and Natural Language Processing, University of Hildesheim, Hildesheim, Germany

We design an electronic dictionary for the domain of graph theory. In this paper we will present the criteria for the corpus-based lemma selection for the dictionary as well as its planned microstructure.
The intended user group are mathematics students attending lectures on graph theory. According to Wiegand et al. (2010) they are semi-experts in this domain as they already have basic mathematical knowledge. The dictionary should help them in cognitive and communicative situations (cf. Fuertes-Olivera & Tarp, 2014; Tarp, 2008): They have to read and understand papers in English which is generally their L2; and later they have to give presentations or write theses in German which is generally their L1. The target group as well as the functions of the dictionary were already described in Kruse & Giacomini (2019).
For creating the dictionary, we built two comparable corpora consisting of texts the students use during their studies in the field. Therefore, the selection was based on the bibliography used in the course as well as on a survey we carried out with students. The English corpus contains eight books and 26 scientific papers (about one million tokens) and the German corpus consists of the lecture notes as well as of (parts of) nine books (about 700.000 tokens). Each corpus comprises about 30.000 types.
One obstacle in the creation process of the corpus was to deal with mathematic formulae. Due to different source file formats it was not possible to use a single workflow. Therefore, one has to keep in mind that as a result of these differences the number of tokens for the same formula may vary in different texts. But this is of no concern as the focus of this project is on the language and not on the formulae.
The process of corpus-based lemma selection is based on different steps. We first extract definition patterns which are typical for the mathematical language (Pagel & Schubotz, 2014). Each of these patterns expresses a certain semantic relation which can be used in the further development of the dictionary (cf. Kruse & Giacomini, 2019). This pattern-based approach will be combined with data produced by other term extraction tools (e.g. Rösiger et al., 2016). The merged results are assessed by three expert raters (inter-rater reliability to be computed). This will lead to the final lemma list.
Further, we designed a microstructure for the dictionary. The bilingual dictionary will include English and German equivalents for each item. Additionally, as an ontology is the backbone of the dictionary, it will provide links to lemmas which are semantically related to a given term. When used on desktop computers, it might be possible to show a shortened definition while hovering over the terms. We present the relations of synonymy, hypernymy / hyponymy, meronymy / holonymy, antonymy, eponymy, pertonymy, mapping, alternatives, attribution, medium and analogy (see below). Of course, not every lemma will have indications for each of the relations.
Another question is the order in which these relations should be presented in the dictionary. It might be useful to give the German/English equivalent first or even visually marked-off as the user often either wants an explanation or a translation. Next, it seems useful to give synonyms as the user might recognize terms which he or she already is familiar with and therefore he or she does not need any further explanations. The synonyms can be followed by the hypernyms in order for the user to learn that the concept looked up is a subtype or an example of that of some other term. The mathematical language is structured in a strongly hierarchical way. Therefore, the given hypernym will always be the direct hypernym on the next higher level.
Further information can be arranged in blocks which the user can open manually (information on demand). One block may contain holonyms / meronyms, pertonyms and antonyms as they are somehow linguistically related with the term. The other block provides information on domain-specific relations as it contains terms related as mediums, analogies, alternatives, attributes, mapping and eponyms (see below).
According to scholars such as e.g. Coady (1997), definitions should be part of a learner dictionary. What these definitions should look like was an issue massively discussed in the pedagogical lexicography of the mid 1990s, when the Cobuild dictionary was published (e.g. Allen (1996), Bogaards (1996), Herbst (1996)). One of its biggest innovations was to use whole-sentence definitions instead of traditional ones which define words

by giving a similar word with the same part of speech. In addition to using this device in definitions, we also propose to use it when we explain the above mentioned relations, since the intended public of the dictionary is not familiar with linguistic terminology. We paraphrase the relations using expressions of general language: synonyms (is also called), hypernyms (is always a) / hyponyms (examples), meronyms (is part of) / holonymy (is composed of), eponymy (is named after), pertonymy (linguistically related), mapping (is usually mapped to / is canonically mapped to), alternatives, attribution (possible properties), analogy. For example, one can describe an *Eulertour* with the following entry: "An Eulertour in a graph is a closed walk which contains each edge exactly once. - An Eulertour is always a cycle. It is part of an Eulerian graph. - It can be computed with Hierholzer's algorithm or Fleury's algorithm. It is named after Leonard Euler."
In the paper we plan to discuss the above-mentioned issues, and we show further examples of articles.

Keywords: LSP-dictionary, microstructure, lemma selection

Term variation in terminographic resources: a review and a proposal

M. Cabezas-García, P. León-Araúz

University of Granada, Spain

Term variation or the coexistence of different terms to name the same concept (e.g. contamination and pollution) is frequent in specialized language (Fernández-Silva 2018). Since variants are not always interchangeable, language users such as translators or terminologists need to know when and why a variant should be used in preference to another. Terminographic resources should facilitate this task by including different variants as well as the criteria guiding their selection. However, variants are not usually fully covered and, when they are included, indicators regarding semantics, pragmatics, or usage are not often provided. This paper investigates the representation of term variation in terminographic resources. Our goals were (i) to confirm whether term variants are underrepresented and usage indications are not usually provided, (ii) to collect the data categories and fields employed in the description of term variants, and (iii) to propose a model of representation of term variation in the terminological knowledge base EcoLexicon. Our results showed that despite the prevalence of term variation, terminographic resources do not usually describe the different possibilities and/or the criteria guiding their selection. In contrast, those which attempt to add pragmatic information do not show this kind of data in a parameterized way.

Keywords: term variation; terminology; terminographic resources

LBC-Dictionary: A Multilingual Cultural Heritage Dictionary. Data Collection and Data Preparation

A. Farina[1], C. Flinz[2]

[1]University of Florence, Italy
[2]University of Milan, Italy

An increasing number and a wide variety of texts on Italian cultural heritage are available today, both online and on paper. However, there are no specific tools (dictionaries, reference materials on technical translations) that can train and support specialists involved in cultural tourism. Mainly focusing on Florence and its cultural heritage, the LBC project (Farina 2016) will try to fill this gap by providing tools for those who have to write/translate for dissemination in various languages: in a first step by building monolingual corpora (English, French, German, Italian, Russian, Spanish) that the user can freely search; in a second step by realizing a plurilingual LSP internet dictionary on cultural heritage which uses the above-mentioned corpora as a primary source. The aim of this paper is to give an insight in the lexicographical process of the LBC-Dictionary, concentrating in particular on data collection and data preparation, which, as is usual for dynamic dictionaries, are open-ended and ever ongoing (Klosa 2013). In particular, we will illustrate the main characteristics of the French and German LBC Corpora and reflect on the provisional French and German entry list, also illustrating the procedure adopted, an alternation of corpus-driven and corpus-based steps (Tognini-Bonelli 2001), for their extraction.

Keywords: Corpora; cultural heritage; internet dictionary

Determining Differences of Granularity between Cross-Dictionary Linked Senses

Eirini Kouvara, Meritxell Gonzàlez, Julian Gross, Roser Saurí

Oxford University Press, United Kingdom

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

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

Ενδογλωσσική και διαγλωσσική προσέγγιση της συνωνυμίας. Συγκριτική μελέτη λογοτεχνικών μεταφράσεων με δίγλωσσα και μονόγλωσσα λεξικά

Ανθούλα Ροντογιάννη

Πανεπιστήμιο Θεσσαλίας, Ελλάδα

Η παρούσα μελέτη προσεγγίζει το επίθετο σύμφωνα με τη ψυχομηχανική / ψυχοσυστηματική θεωρία του G.Guillaume και εστιάζει στις ιδιαιτερότητες τόσο της φύσης του όσο και της λειτουργίας του. Με βάση το ίδιο θεωρητικό πλαίσιο της ψυχομηχανικής, όπως διατυπώθηκε στη σημασιολογική και λεξικολογική της διάσταση, από τους Pottier, Martin και Picoche, προσεγγίζονται και τα ζητήματα που παρατηρούνται κατά τη μελέτη της συνωνυμίας του. Στόχος μας είναι να ερευνήσουμε τη συμβολή της μετάφρασης στη διαβάθμιση και οριοθέτηση της συνωνυμίας του επιθέτου. Συγκρίνοντας και αντιπαραβάλλοντας τις λογοτεχνικές με τις λεξικογραφικές μεταφράσεις, αλλά και τους ορισμούς των μονόγλωσσων λεξικών, ερευνάται η δυνατότητα άντλησης παραδειγμάτων από έργα έγκριτων μεταφραστών ως μέσο εμπλουτισμού των λημμάτων των δίγλωσσων λεξικών.

Keywords: Ψυχομηχανική, Λεξικολογία, Σημασιολογία, Λεξικογραφία, Μετάφραση

Towards the Superdictionary: Layers, Tools and Unidirectional Meaning Relations

Arvi Tavast, Kristina Koppel, Margit Langemets, Jelena Kallas

Institute of the Estonian Language, Tallinn, Estonia

We report on the ongoing project of developing the Ekilex dictionary writing system and joining existing dictionaries into the EKI Combined Dictionary. To facilitate the joining, several tools have been developed to solve data quality issues and turn textual data into structured entities. The resulting superdictionary thus contains various sets of information, which we call layers, either transformed from existing dictionaries or authored already in Ekilex. Our current focus is on the layers for synonyms and equivalents, which we describe in terms of their data model, lexicographic processes and lexicographer feedback from the first six months of Ekilex in production. As it turns out, the layer system may need expanding to accommodate an ever-growing list of requirements. The unidirectional data model for synonyms fully conforms to its design specification and received favourable first impressions, but extended use has started to cast doubt on the optimality of the model. We describe the pros and cons of this model and possible alternatives.

Keywords: synonyms, equivalents, data modelling, unified dictionary

www.euralex2020.gr

Abstracts of Papers

Historical and Scholarly Lexicography and Etymology

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

I. Manolessou, G. Katsouda

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

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

New words in old sources: Additions to the lemma list of a historical scholarly dictionary

Ellert Thor Johannsson, Simonetta Battista

Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark
ellert@hum.ku.dk, sb@hum.ku.dk

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

Stereotypes and Taboo Words in the Dictionaries from a Diachronic and Synchronic Perspective – The Case Study of Croatian and Croatian Church Slavonic

Daria Lazić[1], Ana Mihaljević[2]

[1] *Institute of Croatian Language and Linguistics, Zagreb, Croatia*
[2] *Old Church Slavonic Institute, Zagreb, Croatia*

The paper deals with the lexicographic treatment of sensitive and derogatory 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

Revised entries in the multi-volume edition and TEI encoding: a case of the historical dictionary of Russian

Olga Lyashevskaya[1,2], Jana Penkova[2]

[1] School of Linguistics, National Research University Higher School of Economics, Moscow, Russia
[2] V. V. Vinogradov Russian Language Institute RAS, Moscow, Russia

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. Starting from 1975, the dictionary was originally published in hardcopy (the 32st volume is now in print). 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.

Keywords: historical lexicography, TEI encoding; retro-digitizing, dictionary content revision, Old Russian, Dictionary of Russian Language of the 11th — 17th centuries

John Pickering's *Vocabulary* (1816) Reconsidered: America's Earliest Philological Exploration of Lexicography

Kusujiro Miyoshi

Soka Women's College, Japan

John Pickering is the author of the first dictionary of Americanisms, the *Vocabulary, or Collection of Words and Phrases Which have been Supposed to Be Peculiar to the United States of America* (1816). Allen Read, a masterly scholar of Americanisms, regards the dictionary as "an important landmark in the study of the English language in America", acclaiming Pickering as "one of the most perceptive linguists America has produced". However, there seems to be the situation that research on the *Vocabulary* has scarcely been done since the 1950's. Then, has research on the *Vocabulary* been exhausted? My answer to the question is "Never in the least". When browsing through the *Vocabulary*, we can notice Pickering having finely used quite a few reference materials, thus the body of the *Vocabulary* becoming highly scholarly. As far as I can judge, this fact has not been pointed out to date. My intention in this paper is to clarify Pickering's use of English dictionaries out of such materials. To summarize my analysis, Pickering was versed in wide range of English dictionaries, making the fullest use of them for his investigation on the historical background of Americanisms.

Keywords: John Pickering; Americanisms; use of historical dictionaries

Indexing Paper Quotation Slips with the *Electronic Dictionary of the 17th and 18th Century Polish*

Joanna Bilińska[1], Ewa Rodek[2]

[1] *Institute of Western and Southern Slavic Studies, University of Warsaw, Warsaw, Poland*
[2] *Institute of Polish Language, Polish Academy of Sciences, Warsaw, Poland*

The paper presents the results of experimental paper quotation slips' tagging that was conducted to investigate the possibility of electronic indexing of scanned paper quotation slips constituting a citation archive (a card-index) for the *Dictionary of the 17th- and 1st half of the 18th-Century Polish* (e-SXVII https://sxvii.pl).
The paper citation archive consists of more than 3 million paper quotation slips posing an exemplification of ca. 116,000 of words, which means 86,000 dictionary entries – all of them placed in 836 boxes. There is the need for integration of the archive and the lexicographic panel in order to accelerate the lexicographic work and eliminate human-related mistakes. The test allowed the authors to determine the project priorities, main methodological problems and to decide on future project proceedings. The presented case study may be interesting for other lexicographic teams facing the same problems and looking for an efficient, cheap and quick solution to the problem of using such an abundance of available data.

Keywords: quotation slips, historical dictionary, indexing, card-index, citations archive

The Electronic Dictionary of the 17th- and 18th-century Polish - Towards the Open Formula Asset of the Historical Vocabulary

Renata Bronikowska[1], Magdalena Majdak[1], Aleksandra Wieczorek[1], Mateusz Żółtak[2]

[1] Institute of Polish Language, Polish Academy of Sciences, Poland
[2] Austrian Centre for Digital Humanities and Cultural Heritage, Austria
renata.bronikowska@ijp.pan.pl, magdalena.majdak@ijp.pan.pl, aleksandra.wieczorek@ijp.pan.pl, mateusz.zoltak@oeaw.ac.at

The paper discusses the *Electronic Dictionary of the 17th- and 18th-century Polish* (abbreviated e-SXVII), an important resource for the study of language, history and culture of the period. After several dozen years of gathering material, conceptual work, and after the publication of five fascicles, the print dictionary project was discontinued in favour of a digital version. The work has since accelerated significantly, although development is still ongoing. The paper focuses on new aspects of the methodology stemming from the open form of the dictionary. The innovation offers significant benefits both for the editors and the users. This formula also allows for e-SXVII to be integrated with other electronic language resources, like corpora and digital libraries – a feature currently under intense development.

Keywords: electronic dictionaries, integration of linguistic resources, Middle Polish, historical vocabulary

Announcing the dictionary: Front Matter in the Three Editions of Furetière's Dictionnaire Universel

Geoffrey Williams[1,3], Ioana Galleron[2], Clarissa Stincone[2], Andres Echevarria[3]

[1] *UMR Litt & Arts, Université Grenoble Alpes*
[2] *UMR LATTICE, Université Sorbonne Nouvelle*
[3] *Master Métiers du Livre et de l'Édition, Université Bretagne Sud*

The front matter of a dictionary provides important information as to the background to a work and what is to be expected inside. Although they can be read as standalone texts, it is only when linked to the actual dictionary content that their full potential is realised. This is very much the case for the prefaces to Furetière's *Dictionnaire Universel*, first published posthumously in 1690 and then to go through two major revisions in 1701 and 1725/27. As Furetière left no preface, we start with his *factums*, texts that details his fight with the *Académie Française* who wanted to impede publication. We then have the preface by Bayle of 1690 and then the front matter produced by the two revisers, Basnage de Beauval and Brutel de la Rivière. This was a highly innovative dictionary as both an encyclopaedic work and one with a pedagogical intention. We explore the declarations in the prefaces and the encyclopaedic and linguistic content concentrating on the 1701 edition that is currently being fully digitised in XML-TEI. Citations from largely contemporary texts were used to illustrate entries leading to a very wide knowledge network of late seventeenth century science. Basnage also experimented in illustrating usage through examples, grammatical and pronunciation information.

Keywords: Front matter, historical dictionaries, digitisation, history of ideas

Studying language change through indexed and interlinked dictionaries

C.E. Ore[1], O. Grønvik[2]

[1]University of Oslo, Norway
[2]University of Bergen, Norway

In this paper we study how to use the Meta Dictionary of the Norwegian Language Collections to measure lexical stability in standard dictionaries across a timespan. The Meta Dictionary uses the lexical item as its core unit, expressing each lexical unit in a separate Meta Dictionary entry. The success of this model rests on having access to electronic versions of major and generally accepted dictionaries from the different stages of the orthography of a language. With this documentation it is possible to see - for instance - how much and which parts of the 1873 lexicon (Norwegian vernacular) is present in modern Nynorsk and Bokmål respectively, and whether this lexicon is present in its original orthography, or not. This method for studies of the lexical development is comparable to remote sensing in archaeology and distant reading in literary studies. As an extended example of the application of the method we study a few issues related to the position of the pioneering lexicographers Ivar Aasen (1813-1896) and Hans Ross (1831-1912) in the description of Nynorsk, as shown in more recent lexicographical works, and in particular in two school dictionaries from 1954 and 1970 which border on being spellers.

Keywords: lexical item, lexicon, language change, dictionary, Meta Dictionary model, standard language, orthography, Norwegian

Creating a DTD template for Greek dialectal lexicography: the case of the Historical Dictionary of the Cappadocian dialect

A. Karasimos[1], I. Manolessou[1], D. Melissaropoulou[2]

[1] *Academy of Athens, Greece*
[2] *Aristotle University of Thessaloniki, Greece*

This article reports on the compilation of a full dictionary, both print and digital, of Cappadocian Greek, one of the major Modern Greek dialects. This bilingual (Cappadocian Greek-Standard Modern Greek) dictionary is one of the products of the 'DiCaDLand' dialectological project, funded by the Hellenic Foundation of Research and Innovation (http://cappadocian.upatras.gr/en). Its compilation is based on the powerful professional dictionary editing software TLex Suite, after extensive parameterization in order to meet the needs and the particularities of both the project and the dialectal variety in study. More specifically, we present a sophisticated and state-of-the art e-lexicographic annotation template capable of handling and describing the complex data of an obsolescent and "aberrant" dialect, without written tradition, heavily influenced by language contact (with Turkish), and presenting considerable variation and serve as a model for future approaches to Greek digital lexicography.

Keywords: e-lexicography, dialectology, historical e-dictionary, Cappadocian Greek

7-9 September 2021
Virtual

www.euralex2020.gr

Abstracts of Papers

Lexicography and Corpus Linguistics

A lexicographic platform for migration terminology: problems and methods

Isabella Chiari

Dipartimento di Lettere e Culture Moderne, Sapienza Università di Roma, Roma, Italy

Language on the Fly is 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 has an international level where it is defined for example by institutions like the EU regulations (both legal and administrative); there is a national level, where general international procedures are modified and adapted to the specific country administrative and general migration policies and a last ordinary level that is interlinked to aspects that migrants have to face in their interactions with institutions (for 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

Verb patterns and metaphors: What semantic types can explain about meaning differentiation

Irene Renau

Pontificia Universidad Católica de Valparaíso

In this study, we wonder if we can find regularities in combinations of verb patterns, and if these regularities can be used to find recurrent metaphors in discourse. As 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 of the database (78%) had 2 or more transitive structures and were included in the study. Results show how a few 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

Les termes des arts dans les dictionnaires de la tradition française et dans les corpus de dernière génération: une relation d'inclusion réciproque?

Valeria Zotti

Dipartimento LILEC, Università Alma Mater Studiorum, Bologna, Italia

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: dictionnaire, corpus, termes, art, peinture, sculpture, architecture

Building a Controlled Lexicon for Authoring Automotive Technical Documents

Rei Miyata, Hodai Sugino

Nagoya University, Japan

We describe the framework and the process of building a controlled lexicon, specifically intended for authoring Japanese automobile repair manuals. Focusing on verbs, we seek to control two types of linguistic variations: (1) synonymous words and (2) case (argument) order variations. For synonymous words, we comprehensively extracted verb tokens from a large text data set and classified each verb type as approved or unapproved. For case order variations, we descriptively analysed case structures of Japanese sentences in the data set and defined the canonical order. We also examined the status of the constructed lexicon in terms of coverage, which enables us to establish a tangible goal of future lexicon building. The resultant controlled lexicon with 910 verbs and 954 case patterns can help authors choose appropriate words and construct consistent sentence structures. In order to accomplish effective and efficient authoring, we further proposed and designed two types of authoring support tools: a sentence diagnostic tool that identifies unapproved variations of verbs and sentence structures, and a template-driven writing tool that helps writers compose controlled sentences by completing canonical case patterns.

Keywords: controlled lexicon building, technical authoring, descriptive analysis, variation management, grammatical case, automotive domain

Semantic Relations in the Thesaurus of English Idioms: Corpus-based Study

G. Giztova[1], L. Ismagilova[2]

[1]Kazan Federal University, Russia
[2]Kazan Federal University, Russia

This paper deals with the principles of constructing an Ideographic Dictionary of English Idioms (Thesaurus) based on corpus data. Idioms in the dictionary are arranged by their figurative meaning rather than alphabetically. The need for a new type of dictionary is motivated by the fact that at present there is no a corpus-based dictionary of English idioms built on a thesaural principle. Ideographic description of idioms enables a reader to find the biggest possible amount of idiomatic word combinations of the language that express the given concept. The basic entry of the Thesaurus is called a taxon, consisting of a conceptual descriptor used as a label of a taxon, and a group of idioms expressing the respective taxon. English Web text corpus 2013 (enTenTen13) is used as an empirical basis of the study. The analysis of corpus data presents a range of syntactic patterns, idiom variation, synonymous and polysemous idioms which cannot be retrieved from the existing idiomatic and monolingual dictionaries of the English language, since they fail to register all meanings of an idiom. Today, as lexicography is experiencing "the corpus revolution" (Hanks, 2012), this is a question of key importance. The use of corpora provides additional possibilities for compiling the idiom list and structuring entries.

Keywords: thesaurus, idioms, corpus, variation, synonymy

CROATPAS: A lexicographic resource for Croatian verbs

Costanza Marin[1], Elisabetta Ježek

Department of Humanities, University of Pavia, Pavia, Italy

This paper revolves around CROATPAS (Marini & Ježek 2019), a digital lexicographic resource for Croatian verbs able to frame verbal polysemy and metonymic shifts, which is currently being developed at the University of Pavia. Just like its Italian sister resource T-PAS (Ježek et al. 2014), CROATPAS is a corpus-derived collection of verb argument structures whose argument slots have been manually annotated using a specific set of semantic labels called Semantic Types. At the moment, the resource contains 101 verb entries linked to 457 different verb senses (called *patterns*) and over 22,000 annotated corpus lines (Marini & Ježek 2020). The possible applications of CROATPAS are endless. However, given the status of Croatian as an *under-resourced* and Less Commonly Taught Language, this paper focuses on its potential as a language teaching tool, putting forward some hypothetical vocabulary and grammar teaching suggestions. Even though CROATPAS is still in its prime, its user-friendly interface, bilingual nature and focus on verb semantics bode well for its future as a tool for the teaching of Croatian as a Foreign Language.

Keywords: Croatian, semantic resource, verb, language teaching, Less Commonly Taught Language

Frame Semantics in the Specialized Domain of Finance: Building a Termbase to Aid Translation

V. Pilitsidou[1], V. Giouli[1,2]

[1] *National and Kapodistrian University of Athens, Greece*
[2] *Institute for Language & Speech Processing, ATHENA RC, Greece*

Frame semantics (Fillmore 1977, 1982, 1985) is one of the most important developments for lexicography in the 20th century. The semantic frames approach to lexicon building and semantic representation of meaning at word and phrase level – or even beyond – has been the focus of research in computational linguistics and in Natural Language Processing. The present paper is aimed at describing completed work for the creation of a domain-specific frame-semantic lexicon in Greek (EL) and its alignment to the English (EN) FrameNet. Building on Fillmore's Frame Semantics (Fillmore 1977, 1982, 1985) and on the example set by the FrameNet project (Baker et al. 1998), we developed a bilingual EL-EN lexical resource in the financial domain based on corpus evidence. Our motivation was two-fold: (a) to better account for the semantics of the specialized lexicon – especially the verbs and predicative nouns of the financial domain, and (b) to make cross-lingual alignments at the word level in a way that is meaningful for the translation process.

Keywords: frame semantics, FrameNet, frame, financial domain, translation, terminology, terminological resource

Δημιουργία ηλεκτρονικής λεξικογραφικής βάσης για το περιθωριακό λεξιλόγιο της ΝΕ: αρχικός σχεδιασμός

Κ. Χριστοπούλου[1,3], Ι. Γ. Ξυδόπουλος[2,3]

[1] *Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Ελλάδα*
[2] *Πανεπιστήμιο Πατρών, Ελλάδα*
[3] *Ελληνικό Ανοικτό Πανεπιστήμιο, Ελλάδα*

Στην εργασία αυτή, αφού αρχικά αναφερθούμε στους λόγους που μας οδήγησαν στη δημιουργία μίας ηλεκτρονικής λεξικογραφικής βάσης για στοιχεία του περιθωριακού λεξιλογίου της νέας ελληνικής (βλ. εν. 1), παρουσιάζουμε τα λεξιλόγια που θα φιλοξενούνται στη βάση (βλ. εν. 1.1) και τα τεχνικά της χαρακτηριστικά (βλ. εν. 1.2). Εκτενώς θα αναλύσουμε τη μεθοδολογία που ακολουθούμε στο αρχικό στάδιο της έρευνας αλλά και τους λόγους που επιλέξαμε να σχεδιάσουμε τη βάση σε εφαρμογή ιστού βλ. (εν. 2.1). Ειδικότερα, εστιάζουμε στη μακροδομή της βάσης και τη μικροδομή των λημμάτων (βλ. εν. 2.2 & 2.3). Παρουσιάζουμε μία πρώτη μορφή της μικροδομής των λημμάτων, αναλύοντας τις πληροφορίες που θα εμφανίζονται (π.χ. μορφολογικές, φωνητικές, σημασιολογικές, παραδείγματα χρήσης κ.ά.). Στην ενότητα 2.4 καταγράφουμε τους λόγους που θα αξιοποιήσουμε σώματα κειμένων για την τεκμηρίωση των πληροφοριών της μικροδομής. Τέλος, στην ενότητα 3 αναφερόμαστε στις καινοτομίες που θα παρουσιάζει η ηλεκτρονική λεξικογραφική βάση για τα στοιχεία του περιθωριακού λεξιλογίου.

Λέξεις-κλειδιά: περιθωριακό λεξιλόγιο νέας ελληνικής, ηλεκτρονική λεξικογραφική βάση εφαρμογή ιστού, σχεδιασμός, μικροδομή, σώματα κειμένων, Sketch Engine

Crowdsourcing Pedagogical Corpora for Lexicographical Purposes

Tanara Zingano Kuhn[1], Branislava Šandrih Todorović[2], Špela Arhar Holdt[3], Rina Zviel-Girshin[4], Kristina Koppel[5], Ana R. Luís[1], Iztok Kosem[6]

[1] CELGA-ILTEC, University of Coimbra, Portugal
[2] Faculty of Philology, University of Belgrade, Serbia
[3] University of Ljubljana, Slovenia
[4] Ruppin Academic Centre, Israel
[5] Institute of the Estonian Language, Estonia
[6] University of Ljubljana, Slovenia
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 exhibiting structural problems, and 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

Τα σταλθέντα ή τα σταλμένα μηνύματα;» – απολιθώματα των αρχαίων μετοχών στα σύγχρονα λεξικά και στα σώματα κειμένων

Άννα Ιορδανίδου

Τμήμα Επιστημών της Εκπαίδευσης και Κοινωνικής Εργασίας, Πανεπιστήμιο Πατρών, Ελλάδα
A.Iordanidou@patras.gr

Η παρούσα εργασία αναφέρεται στις αρχαιοελληνικές μετοχές παθητικού αορίστου που δεν εντάσσονται στο ρηματικό σύστημα της νέας ελληνικής, π.χ. *εκτελεσθείς, εξαχθείς, κλαπείς, λεχθείς, προβλεφθείς*. Στο ρηματικό σύστημα της αρχαίας ελληνικής οι μετοχές δήλωναν χρόνο (ενεστώτα, μέλλοντα, αόριστο, παρακείμενο), ενώ στη νέα ελληνική οι ελάχιστες που έχουν απομείνει στον ενεστώτα και στον παρακείμενο δηλώνουν κατεξοχήν τρόπο ενέργειας (εξακολουθητικό – συντελεσμένο). Ουσιαστικά πρόκειται για την ενεργητική επιρρηματική μετοχή ενεστώτα σε *-οντας* (ή γερούνδιο, σύμφωνα με ορισμένους μελετητές), π.χ. *επιλέγοντας*, την παθητική μετοχή ενεστώτα σε ορισμένα μόνο ρήματα, π.χ. *επιλεγόμενος*, και την παθητική μετοχή παρακειμένου, π.χ. *επιλεγμένος*, η οποία σε αρκετές περιπτώσεις έχει λειτουργία ουσιαστικού ή επιθέτου. Η έρευνα στα κυριότερα σύγχρονα ελληνικά λεξικά (ΛΚΝ, ΛΝΕΓ και ΧΛΝΓ) αναδεικνύει διαφορετικές πρακτικές καταγραφής, ως επιθέτων ή ουσιαστικών στο ΛΚΝ και ως μετοχών παράλληλα με τις μετοχές παθητικού παρακειμένου στα ΛΝΕΓ και ΧΛΝΓ σε συγκεκριμένα ρήματα. Εξετάζεται αν και σε ποιον βαθμό οι λεξικογραφικές αυτές πρακτικές μπορούν να υποστηριχθούν από τα δεδομένα της γλωσσικής χρήσης, όπως προκύπτει από αναζήτηση σε σώματα κειμένων.

Λέξεις-κλειδιά: τυποποίηση· γλωσσικό κύρος· γλωσσική χρήση· σώματα κειμένων· νεοελληνικά λεξικά

www.euralex2020.gr

Abstracts of Papers

Lexicography and Language Technologies

Augmented Writing and Lexicography: A Symbiotic Relationship?

Henrik Køhler Simonsen

Copenhagen Business School, Denmark

We live in an age of disruption and technological innovations, and lexicography as a scientific discipline and practice is witnessing a fundamental paradigm shift, cf. also (Fuertes-Olivera 2016), who talks about a "Cambrian Explosion", (Simonsen 2016), who discusses the need for a new "Lexicographic Business Model" and (Tarp 2019), who refers to the paradigm shift in lexicography as "Tradition and Disruption in Lexicography". Like many other disciplines, lexicography is operating within the framework of the "Fourth Industrial Revolution", cf. (Schwab 2015), and lexicography seems to be facing many fundamental challenges. One of these challenges is Augmented Writing (AW), cf. (Banks 2019; G2.com 2019; Marconi 2017 and Simonsen 2020a, 2020b), who discuss AW and how it affects journalism, communication and lexicography, respectively.

The objective of this article is to discuss AW in a lexicographical perspective and discuss to what extent the two disciplines may form a value-adding symbiotic relationship. Based on empirical data from a test of 32 AW technologies, the article discusses this question and presents a number of theoretical considerations on how AW and lexicography might develop a symbiotic relationship drawing on (Colson 2019; Fadel et al. 2017; Liew 2013; Tarp 2019 and Simonsen 2020a, 2020b).

Keywords: Augmented Writing, Writing Assistants, Lexicographically Augmented Writing

Learning dictionary skills from Greek EFL coursebooks: How likely?

Th. Dalpanagioti

Aristotle University of Thessaloniki, Greece

This paper presents a review of the dictionary-oriented material which is included in the EFL coursebooks used in Greek state secondary education. The aim of the study is to determine whether, to what extent and what kind of dictionary skills can be developed through the mainstream coursebooks under review. To this end, based on the relevant literature, a checklist of dictionary skills is first designed to serve as an evaluation tool for examining the coursebooks. Findings reveal an overall limited number of dictionary-oriented exercises, their random distribution across proficiency levels and the underrepresentation of basic receptive and productive dictionary skills. Therefore, since the coursebooks reviewed are not characterized by a thorough or informed treatment of dictionary skills, we may conclude that learning dictionary skills from coursebooks is rather unlikely. To address this gap, we offer suggestions as to how teaching materials can be modified or enriched with a view to developing learners' dictionary-using competence in a systematic way.

Keywords: dictionary skills, learners' dictionaries, coursebook evaluation, TEFL

Audio Recordings in a Specialised Dictionary: A Bilingual Translating and Phrase Dictionary of Medical Terms

Silga Sviķ, Karina Šķirmante

Ventspils University of Applied Sciences, Ventspils, Latvia

The present climate of insufficient funding is having an impact on the development of new dictionaries. New developments should adopt the principle of starting with existing language material that could be made available for publishing in a modern electronic form.
This study reviews an electronic bilingual translation and phrase dictionary of medical terms – designed as a mobile application by a collaboration of several researchers from two higher education institutions – Riga Stradins University (RSU) and the faculties of Translation Studies (FTS) and Information Technologies (FIT) of Ventspils University of Applied Sciences (VeUAS). The dictionary mainly systematizes the study materials acquired from RSU's special study courses in Latvian and English that
were reviewed and supplemented during the development of the dictionary.
The need for such a dictionary was verified through a survey that was carried out prior to the implementation of this project. The successful development of the specialised dictionary benefitted considerably from the prior experience in the development of electronic dictionaries (mobile applications) of the VeUAS researchers and the subject expertise of the medical specialists from RSU. As well as giving a description of the functionality of the dictionary, the article provides an insight into the implementation of the dictionary project, describes the database model used in the dictionary and gives a detailed description of one of the functions of the specialised electronic dictionary – audio recordings, how they are created and added to the dictionary.

Keywords: Audio Recordings, Specialised Dictionary, Mobile Application, Medical Terms

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

Ducassé Mireille

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

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

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

Evaluation of Verb Multiword Expressions discovery measurements in literature corpora of Modern Greek

Vivian Stamou[1], Artemis Xylogianni[2], Marilena Malli[2], Penny Takorou[2], Stella Markantonatou[1]

[1] Institute for Speech and Language Processing, Athena Research Center, Greece
[2] Department of French Language Literature, University of Athens, Greece
vivianstamou@gmail.com, artemis.xylo@gmail.com, mallimariaeleni@gmail.com, pennytak07@gmail.com, stiliani.markantonatou@gmail.com

We report on issues concerning the use of association measures and linguistic knowledge (Part-of-Speeh sequences) with the environment MWETOOLKIT (Ramisch et al., 2010) for discovering all types of verb multiword expressions (VMWE) in corpora of Modern Greek (MG) literature. "MWE discovery" refers to detecting new MWEs in a corpus for lexicographic purposes (Constant et al. 2017). We are interested in boosting lexicographic work with (semi-)automatic facilities, in particular, the development of the VMWE database IDION (Markantonatou et al. 2019).

Keywords: verb multiword expression discovery, association measures, lexicography

A Typology of Lexical Ambiforms in Estonian

Ene Vainik, Geda Paulsen

Institute of the Estonian Language, Tallinn, Estonia

The present study aimsto elaborate an overall outline of the areas that give rise to PoS ambiguity in Estonian. The analysis is based ona database consisting of ca 3500 ambiguous units. Our goal is to map the problematic areas, analyse the processes behind the lexical versatility, and provide a typology of ambiguous forms (the ambiforms) for the lexicographic use. The proposed typology isbased on bi-and unidirectional PoS combinations. As aresultof the analysis,we show howthe lexicalconfluence relationsexhibit a network-like interaction of the traditional PoS categories. The typology of ambiforms is expected to have both theoretical and practical implications –from the perspective of the former, the topic of lexical ambiguity will be set in the modern linguistic and lexicographic frame, and from the angle of applicability, the results will support the lexicographers as the creatorsof lexicographic (root) databases and the developers of language technology systems analysing corpus data.

Keywords: parts of speech, lexical ambiguity, Estonian language

Towards automatic definition extraction for Serbian

Ranka Stanković[1], Cvetana Krstev[1], Mihailo Škorić[1], Rada Stijović[2], Nebojša Vasiljević[3]

[1] *University of Belgrade, Serbia*
[2] *SASA Institute for the Serbian Language, Serbia*
[3] *Loop Foundation, Serbia*

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, by reference and by synonyms) having different syntactic and lexical features. The research corpus consists of 61,213 definitions for 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. Presently developed 21 models cover 57% of dictionary definitions, and for 83% of them the full extent was recognized. The analysis showed 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

Dictionnaire des francophones - A New Paradigm in Francophone Lexicography

Kaja Dolar[1], Marie Steffens[2], Noé Gasparini[3]

[1] CREE, Inalco, Paris, France
[2] Department of Languages, Literature and Communication, Utrecht University, The Netherlands
[3] Institut International Pour La Francophonie (2IF), Lyon, France

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

Keywords: Dictionnaire des francophones, professional dictionaries, collaborative lexicography, general dictionary, linguistic variation of the French language, plurality of norms.

Making dictionaries visible, accessible, and reusable: the case of the Greek Conceptual Dictionary API

V. Giouli, N.F. Sidiropoulos

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

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

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

XD-AT: A Cross-Dictionary Annotation Tool

Meritxell Gonzàlez, Charlotte Buxton, Roser Saurí

Oxford University Press, United Kingdom

Linking lexical datasets to each other is a key strategy for expanding and enriching their content with additional data from other resources. However, different resources show significant differences in the degree of granularity of the lexicographic information. Thus, while extending more coarse-grained datasets with content from fine-grained ones seems a feasible task, the other way around cannot be tackled directly. For this reason, linking datasets at the level of meaning rather than word level is essential. But also, for the same reason, word alignment at the level of meaning is a challenging task not yet solved. Within this context, we created XD-AT, a web-based annotation tool aimed to assist humans to annotate linked sense pairs across dictionaries. In this work, we focus in XD-AT's main functionalities, capabilities and potential extensions, such as reusability and adaptability. For example, although XD-AT has been implemented to classify the type of relationship between linked senses from an English monolingual dictionary and the English side of bilingual ones, XD-AT can also be extended into a more general annotation tool for marking up any type of cross-dictionary mappings at the sense level.

Keywords: annotation tool, sense linking, dictionary mark-up, meaning overlap

Principled Quality Estimation for Dictionary Sense Linking

J. Grosse, R. Saurí

Oxford University Press, United Kingdom

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

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

IdeoMania and Gamification add-ons for App Dictionaries

V. Caruso, J. Monti, Alessia Andrisani, B. Beatrice[1], F. Contento, Z. De Tommaso, F. Ferrara, A. Menniti

University of Naples 'L'Orientale', Italy

The paper presents the main features of a lexicographic mobile game developed by a group of students during a coding course. The app is a learning resource for Italian idioms based on pictorial strategies and theoretical assumptions from phraseological research. After illustrating the app main features, we introduce the pedagogical methodology (*Challenge Based Learning* or *CBL*) used during the course and its specific improvements for supporting Humanities students in learning coding topics by feeling engaged in an app development project. *CBL* is in fact a flexible learning framework which can be used to improve students' skills in the electronic lexicography field. As an example, *IdeoMania* is addressed to idiom learning by introducing innovative solutions for lexicographic applications, such as gamification elements.

Keywords: Lexicographic Apps, Lexicography Learning Programs, Lexicography and Language Technologies, Phraseology and Collocation, Reports on Lexicographical and Lexicological Projects

Sign Language Corpora and Dictionaries: A Four-dimensional Challenge

Anna Vacalopoulou

Institute for Language and Speech Processing, Athena R.C., Athens, Greece

This paper is an analysis of the main challenges in developing sign language resources such as corpora and dictionaries. Although difficulties in data collection and processingarecommon with those in similar projects for vocal languages, there are extra complicationsthat seem to be unique to the creation of resources for sign languages. These, more language specific, problems could be categorised under three general headings: (a) linguistic obstacles, (b) financial obstacles, and (c) social obstacles. Most of the challenges in studying and describing any sign languagespring from the nature of these languagesthemselves, this is why this nature isbriefly described. Instead of dealing with the typical two-dimensional, linear representation of the linguistic message, researchers have to cope with a more complex and dynamic medium involving elements includinghand position and movement, eye gaze, facial expression as well as head and body movement.All these, among others, make the acquisition and processing of signed materialmore expensiveand time-consuming.Finally, the activity of building and exploiting sign languageresources can alsobe held back by social factors, including choice of informants, communication barriers and prejudice.

Keywords: sign language lexicography, multimodal lexicography, sign language corpora, sign language resources, Greek Sign Language

License to use: ELEXIS survey of licensing lexicographic data and software

Iztok Kosem[1,2], Bob Boelhouwer[3], Sanni Nimb[4], Miloš Jakubíček[5], Carole Tiberius[3], Simon Krek[2]

[1] *Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*
[2] *Jožef Stefan Institute, Ljubljana, Slovenia*
[3] *Dutch Language Institute, Leiden, Netherlands*
[4] *Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark*
[5] *Lexical Computing CZ s.r.o., Brno, Czech Republic*

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 not available or are 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, where one of the main aims is 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 some types of lexicographic data. Using a case study it is demonstrated how a more open approach to sharing data can benefit for the community of a particular language, and the ELEXIS community in general.

Keywords: licensing, survey, data, ELEXIS, Dictionary Matrix, lexicographic resources, dictionary, corpus

Towards Automatic Linking of Lexicographic Data: the case of a historical and a modern Danish dictionary

Sina Ahmadi[1], Sanni Nimb[2], Thomas Troelsgård[2], John P. McCrae[1], Nicolai H. Sørensen[2]

[1] *Insight Centre for Data Analytics, National University of Ireland, Galway*
[2] *Society for Danish Language and Literature (DSL), Copenhagen, Denmark*

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

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

Verbal Multiword expressions: a systematic study on the fixedness degree, application to Modern Greek and French

Mathieu Constant[1], Aggeliki/Angeliki Fotopoulou[2]

[1] *Université de Lorraine, Nancy, France*
[2] *Institute for Language and Speech Processing, Athena RC, Athens, Greece*

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 compositionality/fixedness, modelling and encoding MWEs

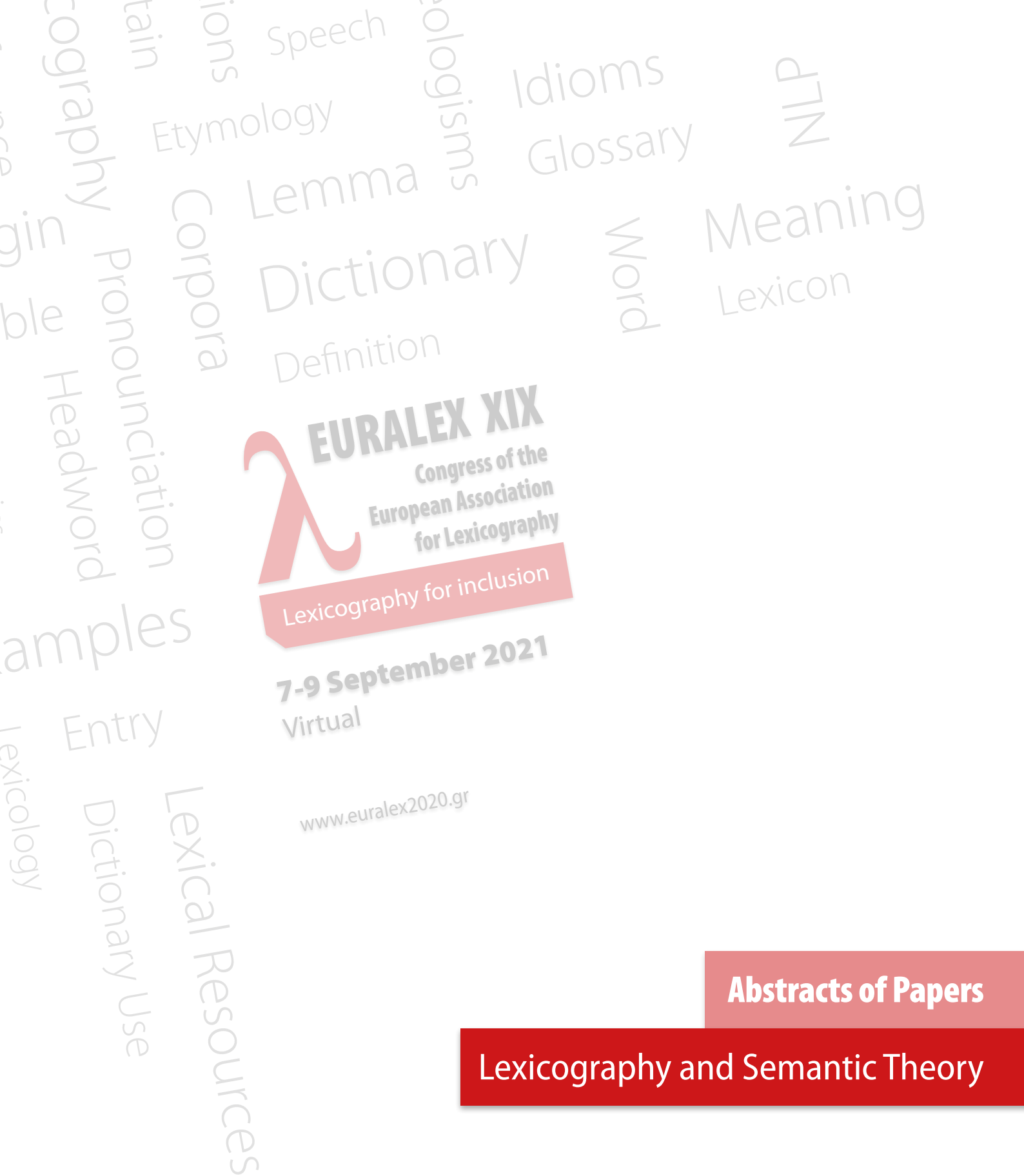
Interlinking Slovene Language Data

Thierry Declerck

Multilinguality and Language Technology, DFKI GmbH, Saarbrücken, Germany
Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Vienna, Austria

We present our current work consisting in interlinking linguistic information included in different types of Slovenian language resources. The types of resources we deal with are a lexical data base (which deals mainly with collocations), a morphological lexicon, a WordNet lexicon and a terminological data base. We first transform the encoding of the original data into the OntoLex-Lemon format, and on the base of this harmonization we can interlink the various types of information included in the different resources. This exercise can lead to a partial merging for the information that is shared across the different resources.

Keywords: Slovenian Language Data, Interlinking, OntoLex-Lemon



Abstracts of Papers

Lexicography and Semantic Theory

Building a Paralympic, frame-based dictionary – towards an inclusive design for *Dicionário Paraolímpico* (Unisinos/Brazil)

Rove Chishman[1], Bruna da Silva[1], Nardes dos Santos[1], Aline Sandra de Oliveira[11], Ana Flávia Souto de Oliveira[3], Larissa Moreira Brangel[1], Gilles-Maurice de Schryver[2]

[1] Applied Linguistics Graduation Program, Unisinos University, São Leopoldo, Brazil
[2] Department of Languages and Cultures, Ghent University, Ghent, Belgium
[3] Department of Classical Languages & Linguistics, Federal University of Santa Maria, Santa Maria, Brazil

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

Definitions of the Oxford English Dictionary and Explanatory Combinatorial Dictionary of I. Mel'čuk

Tinatin Margalitadze

Ivane Javakhishvili Tbilisi State University, Georgia
tinatin.margalitadze@tsu.ge

The Oxford English Dictionary (OED) was an innovative dictionary from many points of view. The paper focuses on one of such innovative features of the OED, namely the method of description of word meaning. One of the ambitions of the OED team was 'to show more clearly and fully than has hitherto been done, or even attempted, the development of the sense or various senses of each word from its etymology and from each other'. For this purpose, the OED editors described semantic structures of English words, mechanisms of development of transferred senses from different semantic components of word meaning. This approach transformed the OED definitions into a very valuable source for the study and investigation of semantic structures of English words.

I. Mel'cuk's theory has a considerable impact on the development of methodology of semantic description of different languages. This theory, from my point of view, is also interesting as it has returned to the lexicographic practice and further elaborated long forgotten great ideas of the OED editors, and particularly James Murray. The paper discusses some parallels between the OED semantic theory and the Explanatory Combinatorial Dictionary (ECD) of I. Mel'čuk.

Keywords: the OED; semantic theory, I. Mel'čuk, the ECD, polysemous models

Τοπωνύμια της ελληνικής και η σχέση τους με τη νεοελληνική γλωσσική εικόνα του κόσμου

O.B. Bobrova

Lomonosov Moscow State University, Moscow, Russia

Από ψυχογλωσσικής άποψης ο γλωσσικός πολιτισμός κάθε λαού αποτελεί ένα σύνολο γνώσεων, ιδεών, συνειρμών (associations) και υποδηλώσεων (connotations), αντιλήψεων και ομαδικών αναμνήσεων που βρίσκουν την έκφρασή τους σε αντίστοιχα γλωσσικά μέσα. Σπουδαίο μέρος της γλωσσικής εικόνας του κόσμου αποτελούν σίγουρα τα χωρικά στοιχεία της (περιοχές, χώρες, πόλεις κ.λπ.) που διαμορφώνουν την "πολιτιστική γεωγραφία" κάθε γλώσσας και έχουν γίνει ήδη αντικείμενο ψυχογλωσσικών μελετών.
Το παρόν εξετάζει τα χωρικά στοιχεία του νεοελληνικού γλωσσικού πολιτισμού τα οποία είναι συνδεδεμένα με όλων των ειδών τόπους και τοποθεσίες: πόλεις (*Σπάρτη*, *Μέκκα*), χώρες (*Αμερική*, *Βαβυλωνία*), όρη (*Γολγοθάς*, *Όλυμπος*), ποταμούς (*Μαίανδρος*, *Λήθη*), λίμνες (*Πρέσπα*) κ.ά. (*Θερμοπύλες*, *Βαστίλη*). Κατά τη σημασιολογική ανάλυση των τοπωνυμίων αναλύονται οι ιδιαιτερότητες της μεταφορικής χρήσης τους, καθώς και ο ρόλος τους στο νοητικό λεξικό του φυσικών ομιλητών της νέας ελληνικής. Τα αποτελέσματα της ανάλυσης μπορούν να χρησιμεύσουν στην επικαιροποίηση και εμπλουτισμό των υπαρχόντων λεξικών της νέας ελληνικής, καθώς και στη συλλογή και περιγραφή στοιχειών που αποτελούν μέρος του "πολιτιστικού λεξικού" της.

Keywords: τοπωνύμια, γλωσσική εικόνα του κόσμου, μετωνυμία, στερεότυπη μεταφορά, ψυχογλωσσολογία

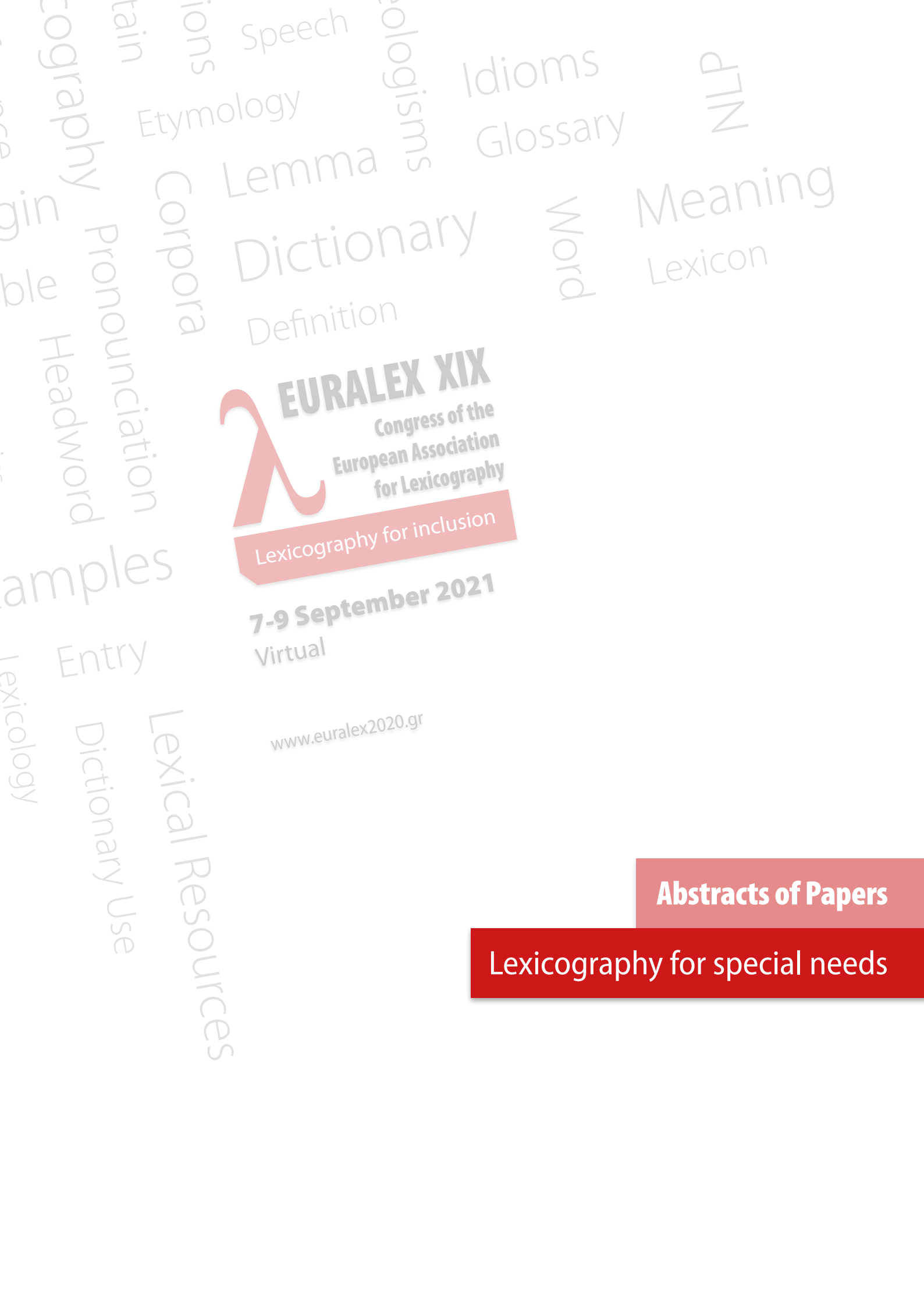
Intensifiers/moderators of verbal multiword expressions in Modern Greek

Magda Mexa[1], Stella Markantonatou[2]

[1] *Department of Philology, University of the Peloponnese, Greece*
[2] *Institute for Language and Speech Processing/Athena R.C., Greece*

We present a comprehensive view of the expression of degree modification of Modern Greek (MG) verbal multiword expressions (VMWEs) with the use of lexical elements that are not part of the VMWE. Our research draws on about 550 natural examples, retrieved from the web, for 63 VMWEs denoting ANGER, SURPRISE, AGONY, FRIGHT, ANXIETY and LOVE. Three general categories of modifiers of this type were recognized: (i) lexical elements that display intensifying or attenuating/mitigating functions as a result of grammaticalization or emphatic stress, (ii) the definite and indefinite article, intensifying *και* 'and' and, (iii) lexical elements expressing levels of gradable properties. The lexical elements in the first two categories seem to apply with a much wider VMWE population than (most of) the items of the group (iii) which is the only group of degree modifiers that seems to need to be recorded in a VMWE lexicon.

Keywords: Verbal Multiword Expressions; degree modification; lexicography



Abstracts of Papers

Lexicography for special needs

The design of an explicit and integrated intervention program for pupils aged 10-12 with the aim to promote dictionary culture and strategies

Z. Gavriilidou[1], E. Konstantinidou[2]

[1] Democritus University of Thrace, Greece
[2] Democritus University of Thrace, Greece
zoegab@otenet.gr, evi1990@hotmail.gr

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

Αρχές για τη δημιουργία ενός εξειδικευμένου λεξικού για ποιητικούς νεολογισμούς: μελέτη περίπτωσης στην ΟΔΥΣΕΙΑ του Νίκου Καζαντζάκη

Νίκος Μαθιουδάκης

Πανεπιστήμιο Γρανάδας
nikosmathious@gmail.com

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, στοιχείο της ποιητικής γραμματικής τους και 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 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, περιγράφονται διεξοδικά τη μακροδομή και τη μικροδομή του λεξικού, σημειώνοντας ενδεικτικά παραδείγματα λημμάτων των καζαντζακικών ποιητικών νεολογισμών.

Λέξεις-κλειδιά: ποιητικοί νεολογισμοί, γλώσσα της λογοτεχνίας, νεοελληνική λογοτεχνία, εξειδικευμένη λεξικογραφία, εξειδικευμένα λεξικά, ΟΔΥΣΕΙΑ· Καζαντζάκης

www.euralex2020.gr

Abstracts of Papers

Lexicography for Specialised Languages, Terminology and Terminography

Introducing Terminologue: a cloud-based, open-source terminology management tool

Michal Boleslav Měchura, Brian Ó Raghallaigh

Fiontar & Scoil na Gaeilge, Dublin City University, Dublin, Ireland

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

Issues in linking a thesaurus of Macedonian and Thracian gastronomy with the Langual system

Katerina Toraki, Stella Markantonatou, Anna Vacalopoulou, Panagiotis Minos, George Pavlidis

Institute for Language and Speech Processing, Athena R.C., Athens, Greece

In the project GRE-Taste: Taste of Greece, we have been developing a trilingual (Greek, English and Russian) thesaurus of food served in restaurants in Eastern Macedonia and in Thrace. For this purpose, have designed a web thesaurus development environment, and we have defined facets and subfacets corresponding to major categories: foods (as ingredients and as dishes), drinks, food sources (and parts thereof), places of origin, preparation methods, functions, state and nutrition. For each concept, the preferred (most common) and non-preferred (synonyms and hidden) terms are entered, as well as nutritional, cultural and other types of information as separate fields and the relationships among concepts (e.g. between a dish and its ingredients, cooking methods or place of origin). In this paper, we discuss the manner of implementing Langual thesaurus for coding foods and the issues involved in the process, such as confusing descriptions and the absence of Greek dishes. We make a suggestion for the enrichment of the Langual thesaurus towards an outcome that could ensure harmonization and interoperability among different applications. We also make a proposal towards resolving Greek terminology problems encountered in the description and classification of foods and other gastronomic concepts.

Keywords: multilingual thesauri, culinary terminology, culinary lexicography, Langual thesaurus, food classification, food description

Lexicography Redefined: Suggestions for Theoretical Recalibration

Henrik Køhler Simonsen[1], Patrick Leroyer[2]

[1] Department of Management, Society and Communication, Copenhagen Business School, Copenhagen, Denmark
[2] School of Communication and Culture, Aarhus University, Aarhus, Denmark

Lexicography has changed radically over the past 20 years and numerous scholars have discussed in a vast number of theoretical contributions whether lexicography is a science of its own and how it should be defined. But has the whole idea of lexicography, the way we see it in the first place, also changed? Is lexicography all about dictionaries one way or another? Can it be understood differently?

In this light, the purpose of our paper is to propose a broad understanding spurred by the closing remark in Adamska-Salaciak's article on lexicography and theory as follows 'theoretical lexicography in its present form is unlikely to offer any such theoretical perspective', cf. (Adamska-Salaciak 2018:14). We do not wish to continue the somewhat tautological discussion of whether lexicography is a science or not, and bring in yet another definition. Instead, we intend to take Adamska-Salaciak up on her call for further theory development and introduce a reconceptualization of lexicography founded on a social-constructivist position paving the way to a broad understanding.

In our discussion, we have drawn on established, seminal lexicographic theory, but reconceptualization requires a break with current views. Consequently, we have also drawn on theories discussed in (Simonsen 2012), (Christensen 2017), (Fadel et al. 2015), (Leroyer & Simonsen 2018a), (Leroyer & Simonsen 2018b), (Liew 2013), (Osterwalder & Pigneur 2010), (Osterwalder et al. 2014) and (Weill & Woerner 2018) etc. Elaborating on the model of (Verlinde et al. 2010) and (Simonsen 2012), we explain how what we call 'lexicographic meaning-construction processes' are at the heart of lexicography. In this light, we present a seven-faced model showing how current and novel elements of lexicographic theory interplay and can be reinterpreted.

Keywords: social-constructivist position, reconceptualization, lexicographic meaning construction processes, seven-faced model

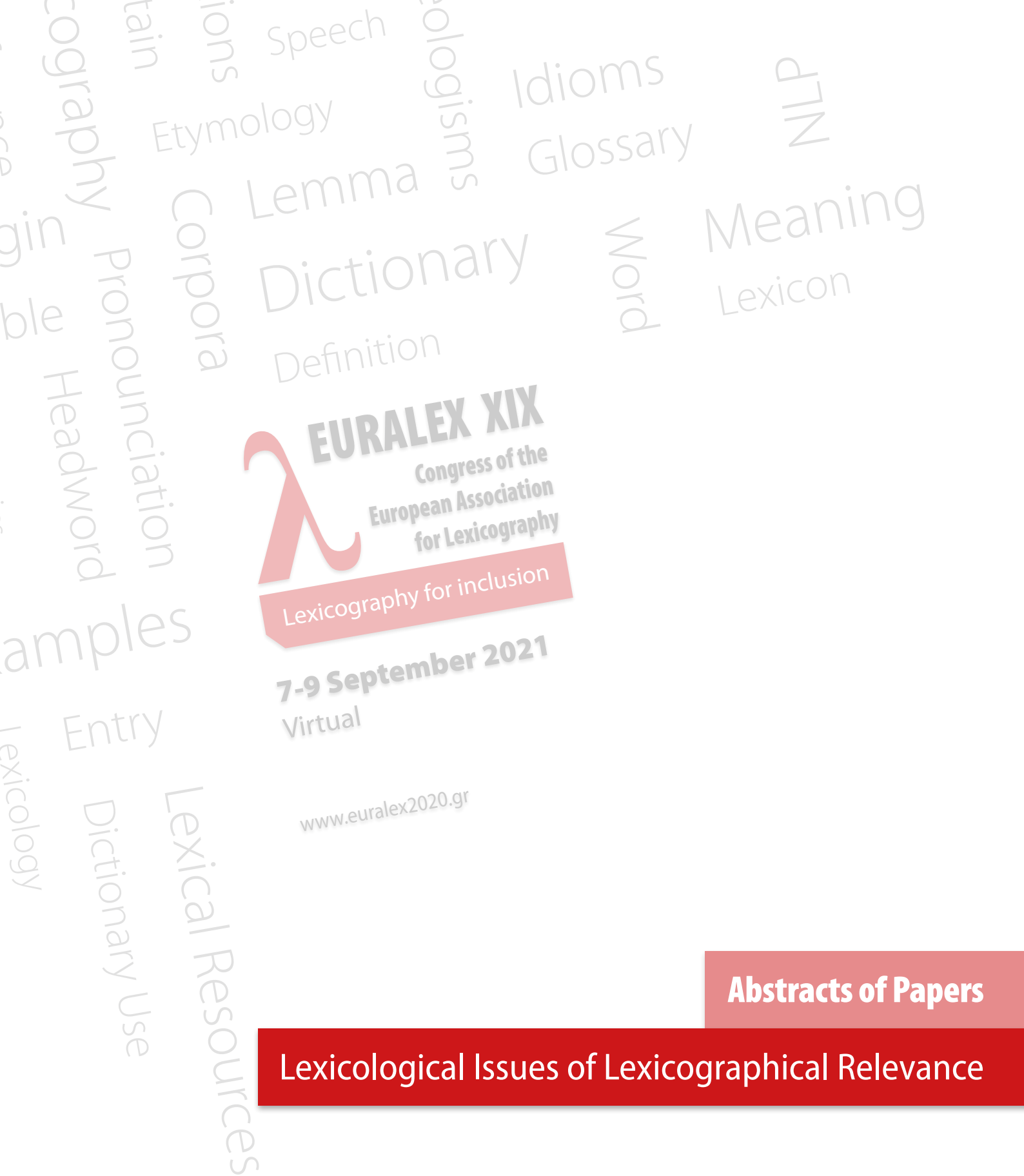
Revisiting polysemy in terminology

Marie-Claude L'Homme

Université de Montréal, Canada

For many, the success of specialized communication is achieved when it is devoid of ambiguity. However, polysemy is quite common in specialized corpora and needs to managed when compiling domain-specific resources. In this paper, we show that polysemy affects many lexical items in specialized texts and review specific cases of polysemy, some of which are seldom discussed in terminology literature. We also show how different types of polysemy can be handled in terminological resources. Methods include: 1. accounting for meaning distinctions using well known tests in lexical semantics; 2. representing links and differences between meanings with lexical relations and labelled argument structures. We also explain how Frame Semantics (Fillmore (1982) and the methodology used in the FrameNet project (Ruppenhofer et al. 2016) can provide a broader perspective on meaning distinctions in specialized fields. Methods are applied to examples found in the English versions of two terminological dictionaries in the fields of computing and the environment.

Keywords: terminology, polysemy, predicative units, alternation, terminological resource, semantic frame



Abstracts of Papers

Lexicological Issues of Lexicographical Relevance

Derivational blends in the Speech of Greek heritage speakers: a corpus-based lexicological approach

Lydia Mitits, Zoe Gavriilidou

Democritus University of Thrace, Komotini, Greece

Found in situations of language contact between Greek and English, Greek heritage speakers living in the US, Canada, Australia, etc. produce loanblends, which combine an English stem e.g. *fence* and a Greek affix e.g. -ι, as in *fénsi* 'fence'. These loanblends are very frequent contact-induced formations that have become part of the Heritage Speakers' everyday language usage. This study analyses fifty (50) such loanblends found in the Greek Heritage Language Corpus, which contains data from Greek Heritage Speakers living in Chicago, US, tests the borrowability scale constraint and the unmarked gender hypothesis for loanwords, and discusses the lexicographic protocol for the compilation of an online dictionary of loanblends of Greek Heritage Speakers.

Keywords: loanwords, loanblends, Greek Heritage Language, borrowability scale, gender assignment, unmarked gender

www.euralex2020.gr

Abstracts of Papers

Phraseology and Collocation

Οι φρασεολογισμοί-κατασκευές της νέας ελληνικής γλώσσας: μια λεξικογραφική προσέγγιση

Elizaveta Onufrieva

Φιλολογική Σχολή του Πανεπιστημίου Λομονόσοφ της Μόσχας, Μόσχα, Ρωσία

Στην παρούσα μελέτη εξετάζονται οι υπάρχοντες τύποι της λεξικογραφικής περιγραφής των νεοελληνικών φρασεολογισμών-κατασκευών – παραγωγικών φρασεολογικών μοντέλων με ένα ή περισσότερα μεταβλητά συστατικά στοιχεία – κενά slots. Η ανάλυση των τρόπων με τους οποίους οι νεοελληνικοί φρασεολογισμοί-κατασκευές περιγράφονται στα μονόγλωσσα ελληνικά λεξικά αναδεικνύει την απουσία στην ελληνική λεξικογραφία μιας ενιαίας προσέγγισης στην περιγραφή των φρασεολογισμών αυτού του τύπου, καθώς και την ύπαρξη ορισμένων προβλημάτων που συνδέονται με παραγωγικά φρασεολογικά μοντέλα γενικότερα. Ένα από αυτά τα προβλήματα είναι η ένταξη των φρασεολογισμών-κατασκευών στα λεξικά σε συμπληρωμένη μορφή ως εντελώς παγιωμένων εκφράσεων χωρίς καμία αναφορά στην ύπαρξη ενός μεταβλητού συστατικού στοιχείου. Η ανάλυση των δεδομένων του σώματος κειμένων δείχνει ότι τέτοια λεξικογραφική περιγραφή των φρασεολογικών-κατασκευών σε μερικές περιπτώσεις δεν ανταποκρίνεται στη γλωσσική πραγματικότητα. Τα συμπεράσματα της παρούσας μελέτης υποδεικνύουν την ανάγκη ανάπτυξης νέων προσεγγίσεων στη φρασεογραφία σε ό,τι αφορά την περιγραφή των φρασεολογικών μονάδων αυτού του τύπου, καθώς και την αναγκαιότητα σύνταξης ενός ειδικού λεξικού.

Keywords: φρασεολογία, φρασεολογισμοί-κατασκευές, παραγωγικά φρασεολογικά μοντέλα, λεξικογραφία

Le Traitement des Proverbes dans les Dictionnaires Explicatifs Roumains du XIXe Siècle

M. Aldea

Université Babeş-Bolyai de Cluj-Napoca, Roumanie

Cette étude s'intéresse au traitement des proverbes dans deux dictionnaires explicatifs roumains du XIXe siècle, *Lexiconul de la Buda* [Lexicon de Buda] (LBe) et *Vocabularu romano-francesu* [Vocabulaire roumain-français] (VRF). En nous appuyant sur plusieurs exemples de proverbes puisés dans ces deux ouvrages lexicographiques, nous examinons la manière dont ces proverbes sont enregistrés dans la structure même des articles, la place occupée par le mot vedette dans la structure du proverbe, les variations lexicales et les éventuels changements de sens, de même que la circulation des proverbes au cours de presque un demi-siècle, leur typologie et les proverbes correspondants dans les autres langues enregistrées dans notre corpus, telles que le latin, le hongrois, l'allemand ou le français, et, en même temps, dans d'autres langues romanes. Au terme de cette analyse, nous pouvons constater la dynamique de la langue roumaine, sa tendance à employer des formules figées, des mots appartenant au lexique fondamental et aussi bien que des mots récents, des mots empruntés aux langues néolatines pour rendre ces proverbes.

Mots-clés: proverbe, le Lexicon de Buda, Ion Costinescu, Vocabulaire roumain-français, dictionnaires explicatifs roumains, traitement lexicographique

The interaction of argument structures and complex collocations: role and challenges for learner's lexicography

Laura Giacomini, Paolo DiMuccio-Failla, Eva Lanzi

Heidelberg University, Heidelberg, Germany

This contribution focuses on the status of complex collocations in pattern-based learner's dictionaries, reporting on findings of the ongoing corpus-based project *Pattern-based learner's lexicography* (Hildesheim University/ Heidelberg University). After comparing recursively built complex collocations with argument-related complex collocations, the paper concentrates on the latter type and its functions. On the one hand, complex collocations displaying argument complementarity efficiently support the identification and formulation of sense patterns. On the other hand, they can serve different purposes within the microstructure of a pattern-based dictionary, namely as semantic types of sense patterns or as lexicographic items in a subordinate treatment unit. Argument-related complex collocations are phraseological lexicalisations of the conceptual scenes provided by sense patterns, and are therefore of key importance to language learners. The challenges related to the extraction of complex collocations from corpora are also addressed in the paper, and proposals are made for improving time efficiency, coverage, and quality of extracted candidates in future research.

Keywords: learner's lexicography, sense pattern, complex collocation, argument structure, cognitive lexicography, lexicogrammar

Abstracts of Papers

Reports on Lexicographical and Lexicological Projects

New developments to Elexifinder, a discovery portal for metalexicographical literature

David Lindemann[1], Laura Giacomini[2], Christiane Klaes[3]

[1] *Jožef Stefan Institute, Ljubljana, Slovenia*
[2] *Universität Heidelberg, Heidelberg, Germany*
[3] *Universität Hildesheim, Hildesheim, Germany*

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

The MorfFlex Dictionary of Czech as a Source of Linguistic Data

Barbora Štěpánková, Marie Mikulová

Charles University, Prague, Czech Republic

In this paper we describe MorfFlex, the Morphological Dictionary of Czech, as an invaluable resource for exploring the formal behavior of words. We demonstrate that MorfFlex provides valuable and rich data allowing to elaborate on various morphological issues in depth, which is also connected with the fact that the MorfFlex dictionary includes words **throughout the** whole vocabulary, including non-standard units, proper nouns, abbreviations, etc. Moreover, in comparison with typical monolingual dictionaries of Czech, MorfFlex also captures non-standard wordforms, which is very important for Czech as a language with a rich inflection. In the paper we also demonstrate how particular information on lemmas and wordforms (e.g. variants, homonymy, style information) is marked and structured. The dictionary is provided as a digital open access source available to all scholars via the LINDAT/CLARIAH-CZ language resource repository. It is available in an electronic format, and also in a more human-readable, browsable and partly searchable form.

Keywords: Morphology, Dictionary, Czech, Lemma, Wordform, Tag

To discriminate between discrimination and inclusion: a lexicographer's dilemma

S. Petersson, E. Sköldberg

University of Gothenburg, Sweden

The overall theme of this paper is the balance between descriptive adequacy and discrimination in dictionaries. More specifically, the purpose is to describe the process of revising dictionary articles related to the grounds of discrimination, in the forthcoming edition of the Swedish monolingual dictionary *Svensk ordbok utgiven av Svenska Akademien* [The Contemporary Dictionary of the Swedish Academy], which is expected to be published in 2020. The focus of the article is on the semantic fields related to sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation and age. Updates of the lemma list, based on a more diverse data set, are presented. Furthermore, revisions of definitions and linguistic examples, motivated by the new data and principles of inclusion, are shown. We also discuss usage labels of negatively charged words and explore cross-references and their role in facilitating non-discriminatory word choices. Moreover, methodological questions are raised, and the role of corpora and other data gathering methods are considered.

Keywords: critical lexicography, Swedish, the Swedish law of discrimination

The New Online English-Georgian Maritime Dictionary Project Challenges and Perspectives

Anna Tenieshvili

Batumi State Maritime Academy, Georgia

The New Online English-Georgian Maritime Dictionary project (NEGMD) comprises two main issues: compilation of dictionary itself and the creation and adoption of Georgian maritime terminology. Optimizing Georgian maritime terminology would greatly add to the development of the whole maritime field of Georgia and it will be especially important for the education of highly qualified seafarers in our country, who will be occupied at sea or on shore-based maritime jobs. The aim of our report on the NEGMD is to present the current status of the project, to show its specifics and the guiding principles during the compilation of the entries of this dictionary and also to present reasons illustrating the urgency of this project.
The creation of an optimal Georgian maritime terminology and the compilation of the NEGMD will increase the motivation of students of maritime education and training (MET) institutions in Georgia to study, will improve the quality of the curriculum at maritime higher education institutions of Georgia and will contribute to adoption of Georgian Maritime terminology in the field of Maritime Education and Training and also in other maritime fields in Georgia.

Key words: dictionary compilation, dictionary entry, corpora, term derivation, term definition

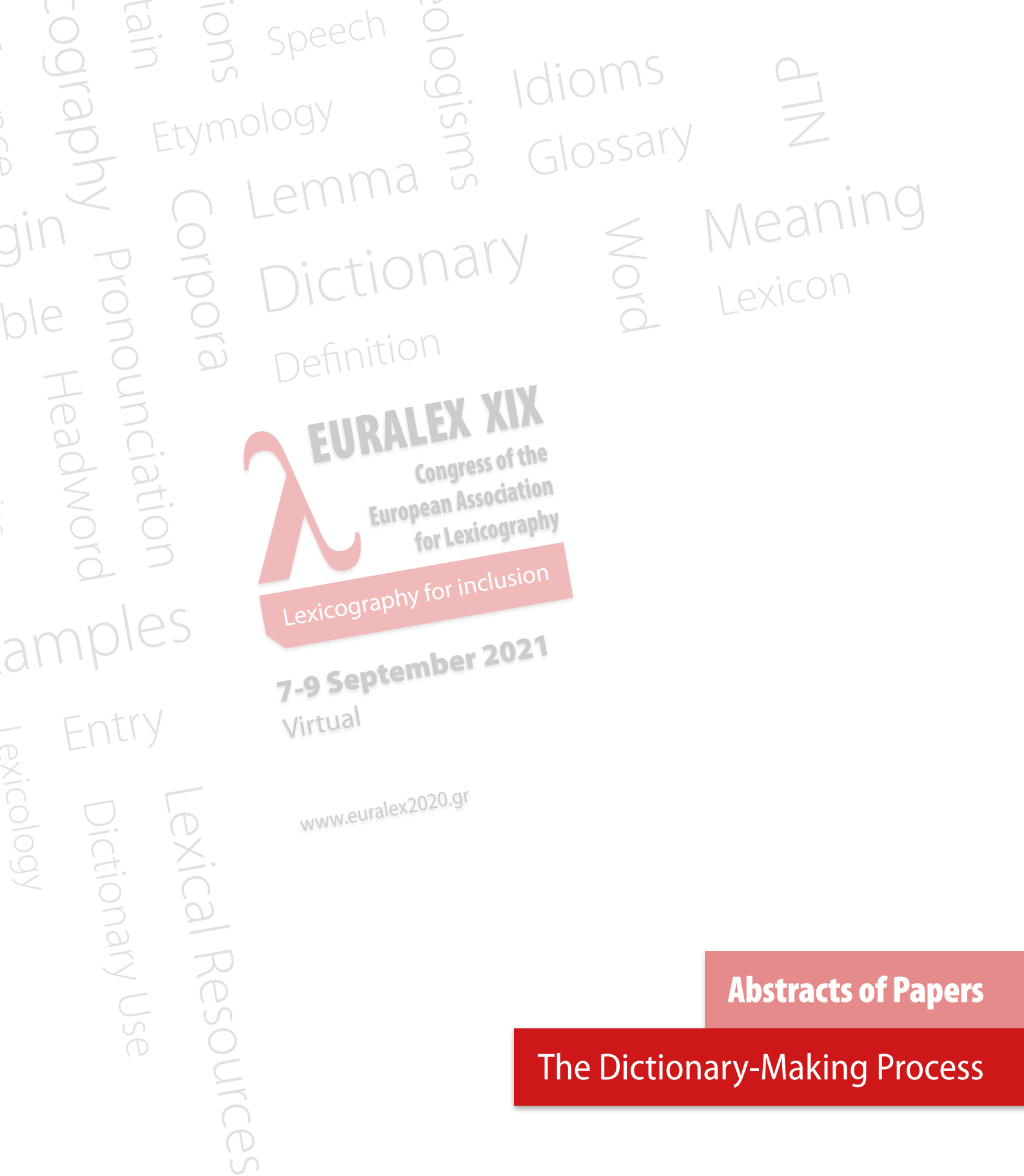
Inventory of New Romanian Lexemes and Meanings Attested on the Internet

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

Romanian Academy, Bucharest, Romania

This article presents a project that monitors the new lexemes and meanings attested on the Internet for the Romanian language and records them in a descriptive dictionary. This project tries to capture the dynamics of the language in the smallest details (e.g. the loan adaptation process) and to update the lexicographic inventory. The Internet is the best source for this purpose. The article begins with the definition of the term *new word* in the sense of this project, and with the characterization of a descriptive dictionary compared to a normative one. Then the method of selecting new words is described which is of random type, i.e. the words are selected by lexicographers from everyday life, without going through a predetermined volume of texts and are registered provided they have 10 attestations from different sources on the Internet. The article also provides a description of some technical aspects and the structure of the dictionary entries. Some solutions to the problems encountered, the first results and how to continue the project are also discussed.

Keywords: new words, Internet, descriptive dictionary, lemma variants, random selection



Abstracts of Papers

The Dictionary-Making Process

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

M. A. Barrios

Complutense University, Spain

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

Keywords: e-dictionary, lexical inheritance, Lexical Functions

"Game of Words": Play the Game, Clean the Database

Špela Arhar Holdt[1,3], Nataša Logar[2], Eva Pori[3], Iztok Kosem[3]

[1] *Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia*
[2] *Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia*
[3] *Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia*

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

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

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

J. Ježovnik, K. Kenda-Jež, J. Škofic

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

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

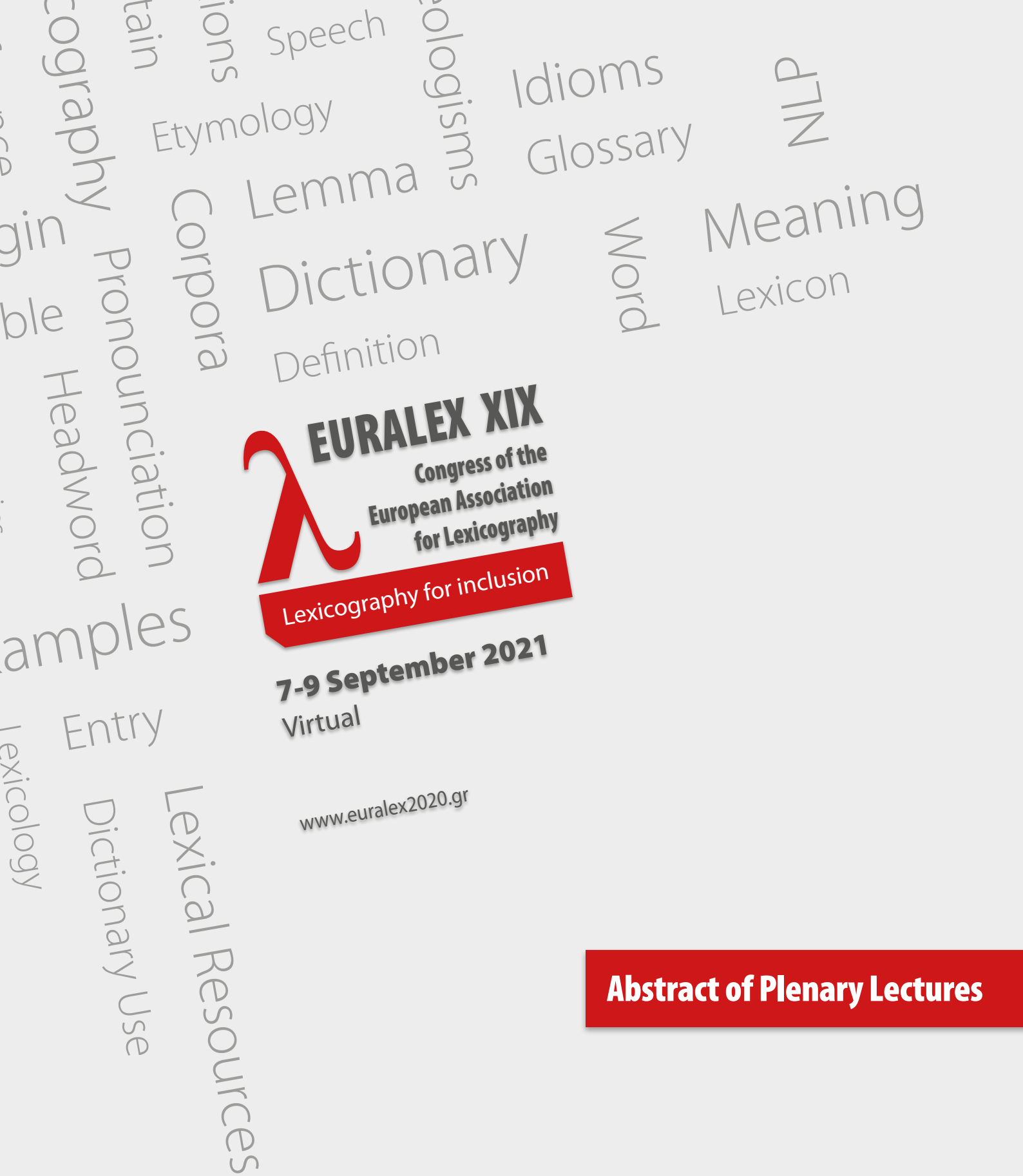
Keywords: Slovene dialects, Slovenian, Franček, lexicography, children's dictionary, dialect data

A limited defining vocabulary and the syntax of definitions

Mariusz Kamiński

Foreign Language Centre at the University of Applied Sciences, Nysa, Poland

This paper aims to challenge the view that a limited defining vocabulary makes definitions syntactically complex, wordy, and convoluted. The paper compares definitions in several dictionaries in terms of occurrence of syntactic constructions that are likely to pose a challenge for less proficient learners.



Abstract of Plenary Lectures

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

Anna Anastassiadis-Symeonidis

Université Aristote de Thessaloniki, Greece
ansym@lit.auth.gr

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

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

Ch. Charalambakis

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

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

Dictionaries and Morphology

Janet DeCesaris

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

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

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

Danie Prinsloo

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

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

Writing with dictionaries Lexicographic support for writing

Robert Lew[1], Ana Frankenberg-Garcia[2]

[1]Department of Psycholinguistic Studies, Poland
[2]University of Surrey, United Kingdom
rlew@amu.edu.pl, a.frankenberg-garcia@surrey.ac.uk

The two main uses of dictionaries are (1) to help users understand a (usually written) text; and (2) to help in the production of well formed (written) prose. These two uses, or functions, have also been termed reception and production; and the dictionaries designed to support with these uses in mind passive and active dictionaries, respectively. The two primary uses call for radically different dictionary content and presentation, and there is no doubt that, of the two activities, it is text production (writing) that is the more challenging to the language user and lexicographer alike. Dictionaries for production have constituted a vanishingly small genre of dictionaries, and practical (and financial) considerations usually resulted in dictionaries that were aimed towards both reception or production, but primarily the former. A prominent exception was the *Idiomatic and Syntactic English Dictionary* (Hornby et al. 1942). These days, with the progressing digitization of dictionaries, separate dictionaries (or versions thereof) for reception and production is a much more realistic option. In my talk, I want to focus on lexicographic support for writing, starting with monolingual dictionaries and thesauri, through bilingualized and bilingual dictionaries, and ending with more task-focused lexicographic tools now known as writing assistants.

Index of Authors

Roman

A

B

C

D

E

F

G

H

I

J

K

EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-9 September 2021

Virtual

www.euralex2020.gr