Electronic lexicography in the 21st century (eLex 2021)
Post-editing lexicography

Book of abstracts

edited by Iztok Kosem
Michal Cukr

virtual, 5–7 July 2021

elex.link/elex2021

Electronic lexicography in the 21st century (eLex 2021):
Post-editing lexicography
Book of abstracts

edited by
Iztok Kosem
Michal Cukr

published by
Lexical Computing CZ s.r.o., Brno, Czech Republic

proofreading by
Paul Steed
Dean DeVos
Jan Nagel

Brno, July 2021

TABLE OF CONTENTS

EXTENDED ABSTRACTS

ABSTRACTS OF PAPERS

ABSTRACTS OF KEYNOTE TALKS

Extended abstracts

Legal Lexicography: From Paper Dictionaries to Online Tools

Tomáš Duběda

Charles University, Institute of Translation Studies, Czechia
E-mail: dubeda@ff.cuni.cz

Action Legal translation has a privileged position within the sector of non-literary translation: it deals with texts that may have serious practical implications, it is partly subject to statutory regulation (Act Nr. 357/2019); and it is the object of a specific sub-discipline called Legal Translation Studies (Prieto Ramos, 2014). Bilingual lexicographic tools in the wide sense of the term are crucial for legal translators. Unlike general dictionaries, legal dictionaries should not be mere word-lists, but relevant sources of information facilitating the translator's decision-making (Chromá, 2004). This implies some degree of encyclopaedic information (De Groot & Van Laer, 2006). Legal dictionaries including Czech are only available for major languages of European descent, and a very limited number of other languages. Only a few of them have also been released in electronic format.

Apart from bilingual dictionaries, legal translators also use a variety of online sources: institutional (IATE, ISAP), encyclopaedic (Wikipedia), corpus-based (Google Translate, Glosbe, Linguee), personal (translation memories, personal glossaries) or informal (web glossaries, resources shared in different translator communities).

Despite the undeniable benefits that electronic lexicography brings, online bilingual dictionaries are still extremely scarce in the domain of law. The first pioneering projects include Nielsen (2014, Danish – English) and Szemińska (2017, English – Polish). LEGILEX-FR, an online French-Czech and Czech-French database of legal language, has been developed since 2020 at the Institute of Translation Studies in Prague.

LEGILEX-FR is a publicly available tool conceived as a complement to the conventional French-Czech and Czech-French legal dictionary (Larišová, 2008). The database, currently containing more than 4,500 entries, makes extensive use of the advantages offered by online format:

- both a desktop and a mobile interface,
- fast and intuitive full-text search in both directions, reducing search-related cost (Van Laer, 2004),

- encyclopaedic information (term categorisation, definitions and remarks on compared law),
- remarks on usage, frequency, style etc.,
- references to legislation and other external sources,
- for each search, instantaneous access to user-defined corpus,
- easy updates.

In Van Laer's (2014) tripartite typology of legal dictionaries (word-list – explanatory – comparative), LEGILEX-FR aspires to the explanatory/comparative degrees. The database is primarily intended for professional translators and students of translation. It integrates some of the needs revealed in a poll conducted among sworn translators (Anonymous, in print), trying to combine information richness with practicality in the most optimal way.

Given the specific nature of legal language, the database contains significantly more collocations (multi-word units, e.g. personne physique, clause de non-concurrence) than single words. Special attention is paid to longer stretches of text and formulaic expressions (e.g. il est préalablement exposé ce qui suit).

Since legal terminology is closely connected to the legal system it is part of (Chromá 2004), not all equivalents are symmetrical (i.e. functioning in both directions). By providing information about the particular jurisdiction (ČR, FR, BE, CH, LU, CA, EU), the database helps translators differentiate between functional equivalents (corresponding terms that exist in both systems) and linguistic equivalents (terms created for the purpose of translation).

Keywords: online lexicography; legal translation; legal language corpora; Czech; French

References

De Groot, G.-R. – Van Laer, C. J. P. (2006): The Dubious Quality of Legal Dictionaries. International Journal of Legal Information, 34/1, s. 65–86.

Chromá, M. (2004): Legal Translation and the Dictionary. Tübingen: Max Niemeyer Verlag.

Chromá, M. (2010): Česko-anglický právnický slovník s vyvětlivkami. Voznice: Leda.

Larišová, M. (2008): Francouzsko-český, česko-francouzský právnický slovník. Plzeň: Aleš Čeněk.

Nielsen, S. (2014): Database of Legal Terms for Communicative and Knowledge Information Tools, in: Mac Aodha, M. (ed.), Legal Lexicography. A Comparative Perspective. Farnham: Ashgate.

Prieto Ramos, F. (2014): Legal Translation Studies as Interdiscipline: Scope and Evolution. Meta 59/2, pp. 233–466.

Szemińska, W. (2017): Helping the Translator Choose: The Concept of a Dictionary of Equivalents. Jazykovedný časopis 68/2, pp. 355–363.

Van Laer, C. J. P. (2014): Bilingual Legal Dictionaries: Comparison Without Precision? in: Mac Aodha, M. (ed.), Legal Lexicography. A Comparative Perspective. Farnham: Ashgate.

Websites:

Act Nr. 354/2019 on Sworn Interpreters and Sworn Translators [accessed 10 January 2021]. <https://www.zakonyprolidi.cz/cs/2019-354>

ISAP – Databáze závazných termínů [accessed 10 January 2021]. <https://isap.vlada.cz/dul/zavaznet.nsf/ca?OpenView>

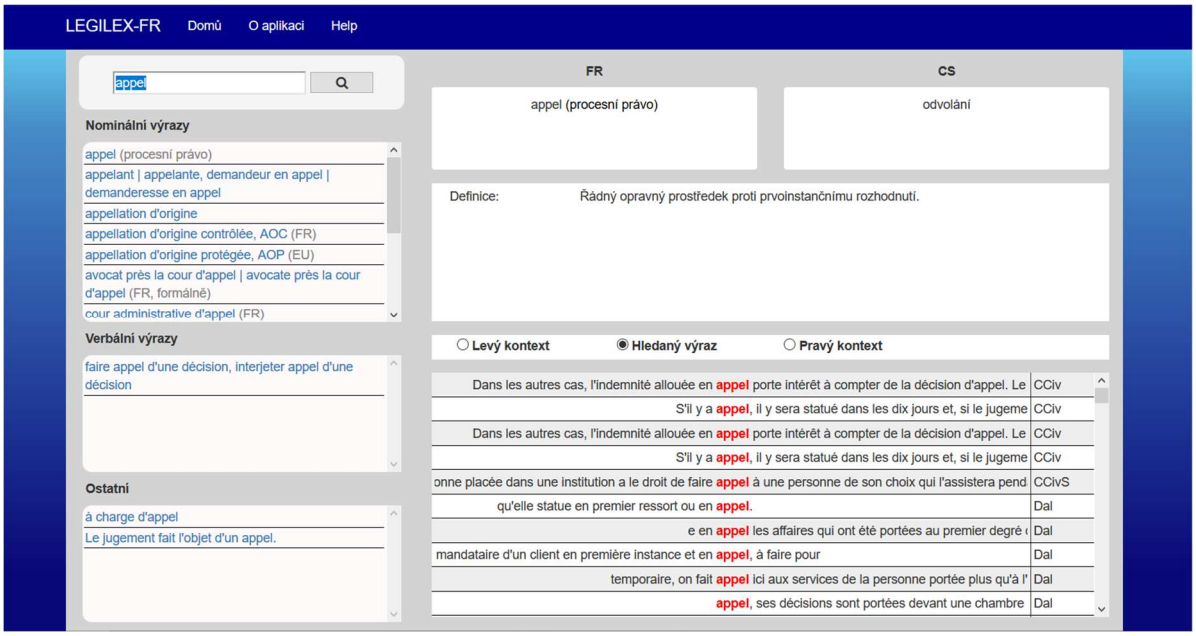
IATE – Interactive Terminology for Europe [accessed 10 January 2021]. <https://iate.europa.eu/home>

LEGILEX-FR <http://lex.ff.cuni.cz/legilex-fr>

Linguee [accessed 10 January 2021]. <https://www.linguee.com>

Glosbe [accessed 10 January 2021]. <https://cs.glosbe.com>

Wikipedia [accessed 10 January 2021]. https://cs.wikipedia.org/wiki/Hlavn%C3%AD_strana



Translating action verbs using an online dictionary based on video animations

Anne-Kathrin Gärtig-Bressan

University of Trieste

E-mail: akgaertig@units.it

Action verbs, that is, verbs that can be used to refer to concrete, observable actions, such as motion verbs or verbs expressing the positioning, modification or destruction of objects etc., are a challenge in L2-acquisition, in human and also in machine translation. The reason for this can be found in the way in which languages lexicalize actions: there are languages such as Italian or partially English that prefer extremely polysemic verbs, which in their proper meaning can refer to a large number of actions (for ex. *to put*, *to take*; *mettere*, *prendere*), but the variation of meanings doesn't correspond between languages (cfr. Moneglia & Panunzi, 2010). In addition, there are typologically different languages, such as German, that prefer very specific verbs, applicable only to one or few precise actions (cfr. Korzen, 2018, based on Talmy, 1985 & 2000; Gärtig-Bressan, 2019a), for ex. *aufspannen* (*einen Schirm aufspannen* 'to open an umbrella') or *auftrennen* 'to unpick'. Translating from a language of the first group into one of the second group requires a continuous disambiguation of polysemy (cfr. Nied Curcio, 2002).

Traditional bilingual dictionaries are not always of help: not all meanings of a polysemic verb are registered in them, and even when the different meanings are listed, it is not always easy for the user, especially if he is a beginner, to select the right one.

An alternative is offered by the IMAGACT ontology developed under the direction of Massimo Moneglia at the Universities of Florence and Siena and the CNR Pisa (cfr. Moneglia et al., 2012, 2014; Panunzi et al., 2014; Pan et al., 2018). This freely accessible online ontology of action verbs was developed corpus-based for the languages English and Italian and thus contains the denotation for a big variety of actions (1010) that are frequently referred to linguistically. Each action is represented by a short video or animation so that semantic paraphrases and other means of verbal disambiguation are not necessary. The ontology can be accessed as an onomasiological dictionary via the videos, but also semasiologically via the verb in a given language. By now, the database contains a total of 15, genealogically and typologically different languages (besides Italian and English, among others German, Polish, Chinese, Japanese, Hindi, Arabic).

The proposed paper presents a study on the usefulness of the ontology for translation from L1 Italian to L2 German. Around 20 Italian university students of German as L2 with an average level of B1-B2 were asked to translate 20 simple Italian sentences with polysemic

action verbs into German. Half of the students were allowed to use traditional bilingual dictionaries for this task, the other half worked with IMAGACT. The initial hypothesis was that the IMAGACT group would select the correct German verb more often with the help of the videos, while the other group might achieve better results in terms of correct conjugation and construction, because the microstructure of IMAGACT is extremely reduced and does not provide explicit morphosyntactic information.

The first hypothesis could be confirmed with strong influence: While for the verbs translated using IMAGACT, the appropriate German equivalent was chosen in over 80% of the cases, for the verbs translated using a traditional dictionary this was the case in only 50%. There were only slightly more errors in the IMAGACT group when it came to conjugating the German verb and using it correctly in the sentence.

The work with the new resource was predominantly evaluated positively by the students. The lack of information on conjugation and valency as well as the restriction of the ontology to verbs were criticized.

Keywords: IMAGACT ontology; L2-translation Italian – German; action verbs; bilingual and multilingual lexicography; experimental study

References

Gärtig-Bressan, A.-K. (2019a). Aktionsverben im inter- und intralingualen Vergleich: Die IMAGACT-Ontologie und ihre Erweiterung um Deutsch. *Linguistik online*, 94(1), pp. 19–43.

Gärtig-Bressan, A.-K. (2019b). I verbi generali italiani come sfida nella traduzione verso il tedesco L2 e l'ontologia IMAGACT come supporto. *Rivista internazionale di tecnica della traduzione*, 21, pp. 133-155.

IMAGACT. Accessed at: www.imagact.it (31 May 2021)

Korzen, I. (2018) L'italiano: una lingua esocentrica. Osservazioni lessicali e testuali in un'ottica tipologico-comparativa. In: I. Korzen (ed.) *La linguistica italiana nei Paesi Nordici*. Pisa: Pacini (=*Studi Italiani di Linguistica Teorica e Applicata* 47 (1)), pp. 15–36.

Moneglia, M. & Panunzi, A. (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In E. Cresti & I. Korzen (eds.) *Language, Cognition and Identity. Extensions of the endocentric/exocentric language typology.* Firenze: FUP, pp. 27–45.

Moneglia, M. et al. (2012). The IMAGACT Cross-linguistic Ontology of Action. A new infrastructure for natural language disambiguation. In N. Calzolari et al. (eds.) *Proceedings of the 8th International Conference on Language Resources and Avaluation* (LREC 2012). Paris: ELRA, pp. 2606–2613.

Moneglia, M. et al. (2014). The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In N. Calzolari (ed.) *Proceedings of LREC'14*. Reykjavik: ELRA – European Language Resources Association, pp. 3425–3432.

Nied Curcio, M. L. (2002). La difficoltà della polisemia nell'acquisizione del tedesco come L2. *Studi italiani di linguistica teorica e applicata*, 31(1), pp. 133–149.

Pan, Y. et al. (2018). IMAGACT4ALL. Una ontologia per immagini dell'azione per l'apprendimento del lessico verbale di base delle lingue seconde. In A. De Meo & M. Rasuo (eds.) *Usare le lingue seconde. Comunicazione, tecnologia, disabilità, insegnamento*. Milano: AItLA, pp. 129–148.

Panunzi, A. et al. (2014). Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology. In A. Abel & C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15-19 July 2014, Bolzano/Bozen. Bolzano/Bozen: EURAC research, pp. 1163–1170.

Talmy, L. (1985). Lexicalization Patterns: Semantic Structure in Lexical Forms. In T. Shopen (ed.) *Language Typology and Syntactic Description*. Cambridge: The University Press, vol. III, pp. 57–149.

Talmy, L. (2000). *Toward a Cognitive Semantics*. Vol. 2: *Typology and process in concept structuring*. Cambridge, Mass./London: The MIT Press.

A bilingual dictionary from a parallel corpus linked at the lexical level

Tarrin Wills

University of Copenhagen, Denmark
E-mail: tarrin@hum.ku.dk

Bilingual and historical dictionaries can be produced from translated parallel corpora either automatically or manually. The automatic methods (e.g. using SketchEngine - Baisa et al., 2015) require alignment of corpora at the level of the paragraph or sentence, the smallest feasible level where a one-to-one correspondence can occur between two corpora in the same order. The manual methods (such as the new *Lexicon of the Nordic Medieval Laws* - Love et al., 2020) involve dictionary entries largely based on manually excerpting lexical equivalents between the two corpora, showing the usage of the word in the context and as interpreted by the translator, based on their understanding of the whole text.

Although many if not most words in a text have a one-to-one correspondence with the translation, the word order is almost inevitably not the same, meaning that they cannot be linked at the lexical level without reordering. Corpus linguistic tools assume that texts come in a fixed order at all levels of their linguistic structure. Linking two corpora at a very detailed level, however, requires a data model that permits one-to-one linking of words, for example, but with an alternative ordering in the translated version.

The problem of linking text and translation at the lexical level can be overcome to a certain extent with appropriate tools and methods The editing and translation project, 'Skaldic Poetry of the Scandinavian Middle Ages' (skaldic.org), uses a data model that allows for linking translations at the lexical level and reordering them appropriately for the target language (English). Visual tools are provided for those working with the data to facilitate this process. The editor-translators are encouraged to include all lexical elements in the translation, within the limits of the idiom of the target language, a common practice anyway for scholarly translations of historical texts. Those entering the data produce the closest feasible match between the text and translation at the lexical level.

The MSCA-funded Lexicon Poeticum (lexiconpoeticum.org) project has lemmatised the resulting corpus. These processes, when taken together, are sufficient to produce automatic dictionary entries that list the contextual translations for each word. These give an overview of the usage of the word in all contexts as interpreted by the editor-translator. Further information linked to the texts and words from the original project can also be incorporated into the resulting entries, including the source materials, annotations and other semantic analyses. This is sufficient information for most users of the lexicon. The

final dictionary will therefore require only further editing in cases where different usages are not encompassed by the translations and/or the usage requires further explanation than the contextual translations.

This method requires suitable digital tools; a manageable corpus; source and target languages that are remotely related at least; and compatible practices of editing and translation in the preparation of the corpora. Within these parameters, the Lexicon Poeticum project demonstrates that it is possible to create useful lexicographic resources automatically, based solely on translation and lemmatising..

Keywords: historical lexicography; digital humanities; translation studies; Old Norse

References

Vít Baisa, Barbora Ulipová, and Michal Cukr, ‘ Bilingual Terminology Extraction in Sketch Engine ’. In Ninth Workshop on Recent Advances in Slavonic Natural Language Processing, the Czech Republic, December 2015, pp. 61–67.

Clunies Ross et al. (eds), *Skaldic Poetry of the Scandinavian Middle Ages* (Brepols, Turnhout). https://skaldic.org.

Tarrin Wills (ed.), *Lexicon Poeticum* . https://lexiconpoeticum.org.

Jeff Love et al. (eds.). A Lexicon of Medieval Nordic Law. Cambridge, UK: Open Book Publishers, 2020. https://doi.org/10.11647/OBP.0188 and https://www.dhi.ac.uk/lmnl.

Scribbling in the digital margins: annotating and extending published lexicographic works in Evoke

Sander Stolk

Universiteit Leiden

E-mail: mailto:s.s.stolk@hum.leidenuniv.nl

Lexicographic works available on the Web have a lot to offer their users: their remotely accessible functionality ranges from browsing entries and performing searches to, in some cases, drawing statistics from the lexical facts contained within. A function that we seldom see in publications of dictionaries since the move from ink to internet, however, is that of fashioning personal copies.[1] Where owners of a printed copy are able to annotate and expand their copy, circulate the result amongst peers, and stimulate a dialogue, users of electronic editions often have to resort to private bookkeeping in a separate file or (in some cases)[2] use a public commenting system that requires moderation and may well end up cluttering the website. This paper presents an alternative.

The web application Evoke[3] implements a novel approach that provides its users with ways to navigate and analyse thesauri, but also to annotate and tag lexical facts and thereby creating customized copies. User additions in Evoke, unlike in many existing solutions, do not place an additional burden on the host of the dictionary: no user accounts need managing, no hosting of user content is necessary, and no content moderation is needed. User data is not stored online but, instead, kept in the user's internet browser.[4] Moreover, as any additions reference rather than contain the original content, the functionality in Evoke ensures users still abide by licenses that prevent users from downloading original dictionary content.

Evoke provides users with full control over their annotations and tags: they can make file backups, share these, and review ones made by others when opened on the website of the original electronic dictionary. Thus, researchers are able to engage in open science, are enticed to interact directly with the dictionary, and remain engaged with the original website of the lexicographic work for viewing their additions in unison with that which

[1] No such functionality is available, for instance, in well-known works such as the OED Online (2021). URL: https://www.oed.com ; or The Historical Thesaurus of English, 2nd ed (version 5.0), eds. C. Kay et al. (2021). https://ht.ac.uk .

[2] A public commenting system can be seen in use in the Historical Thesaurus of Scots, ed. S. Rennie (2017). URL: https://scotsthesaurus.org .

[3] Evoke (2018). URL: http://evoke.ullet.net . Demonstration: http://evoke.ullet.net/demo.

[4] This approach deviates from those that require online hosting, including the annotation tool *hypothes.is*, URL: http://hypothes.is ; and the framework *SOLID*. URL: https://solidproject.org .

they annotated. As the annotations are digital, explicit, and interpretable alongside the dictionary content, their form is highly suitable to utilize in any additional service that the publishers or lexicographers of the work may provide (be it paid or at no charge, be it advanced analyses or further refinement). In its design and the functionality it offers, then, Evoke navigates concerns of both users and publishers in providing functionality for scribbling in the digital margins of its electronic lexicographic works.

Keywords: software; annotation; tagging; thesaurus; web storage; linked data

The usefulness of non-examples in online dictionaries

Anna Dziemianko
Adam Mickiewicz University
E-mail: danna@ifa.amu.edu.pl

Non-examples showing typical grammatical errors with (high-frequency) headwords are present in major e-MLDs (e.g., OALD, LDOCE, MEDO). Unfortunately, little is known about their actual usefulness. The aim is to determine whether non-examples in online dictionaries affect error correction accuracy and time as well as the immediate and delayed retention of correct usage. Four questions are posed:

Q1. Do non-examples help to correct grammatical errors?

Q2. Does error correction time depend on the presence of non-examples?

Q3. Is the immediate and delayed retention of correct grammatical structures affected by non-examples?

Q4. Are error correction accuracy, time and usage acquisition conditioned by example distribution in entries?

A four-part online experiment (a pre-test, a main test, immediate and delayed post-tests) was built around 18 sentences showing incorrect use of English nouns, verbs, adjectives, adverbs and determiners. The pre- and post-tests checked the subjects' ability to correct the errors without dictionaries. In the main test, error correction was based on monolingual online dictionary consultation. Two dictionary versions were created: one offered regular encoding examples, the other additionally showed non-examples in red. In each version, examples useful for error correction occupied entry-initial, medial and final positions. The post-tests checked the ability to correct errors immediately after exposure and 2 weeks later.

196 learners of English (B2 in CEFR) participated in the study. 102 accessed the online dictionary with non-examples, and 94 consulted the version without non-examples.

2 x 3 repeated measures GLM ANOVAs were conducted for each dependent variable. Tukey HSD tests were used to investigate significant differences.

Example types had a significant effect on error correction accuracy ($F=5.76$, $p=0.037$, partial $\eta^2=0.366$). Non-examples (40.90%) helped to correct over 50% more errors than

regular ones (26.77%; 40.90*100/26.77=152.78). The position of examples played no statistically significant role (F=2.69, p=0.09, partial η^2=0.212).

Figure 1. Error correction accuracy by example type

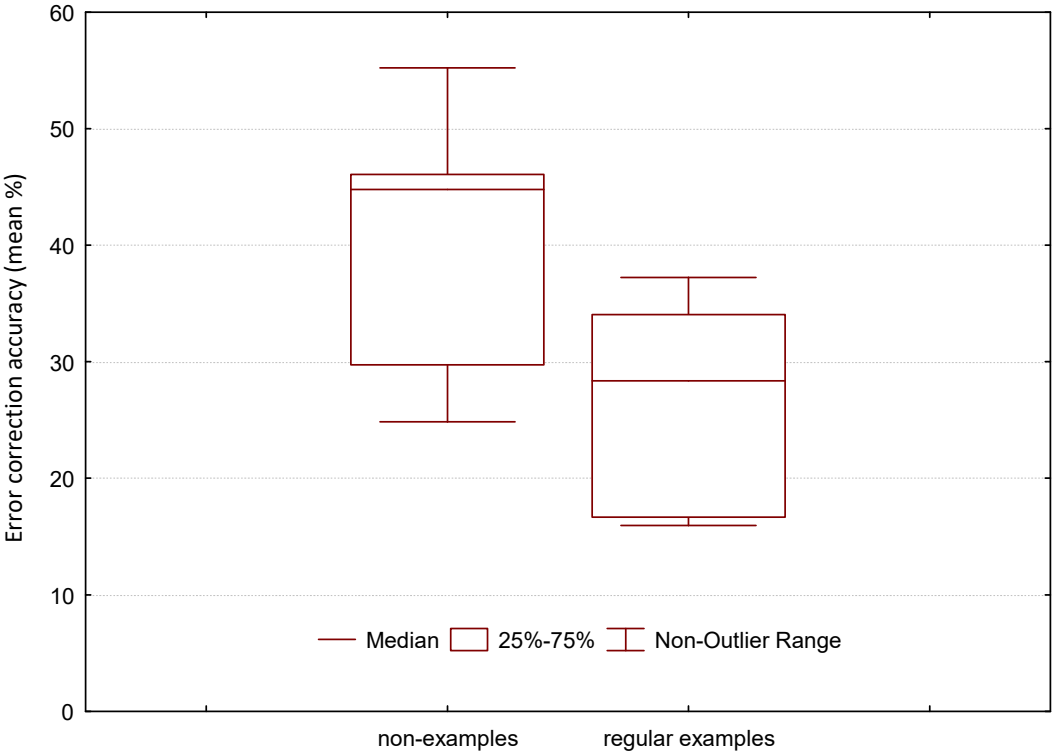
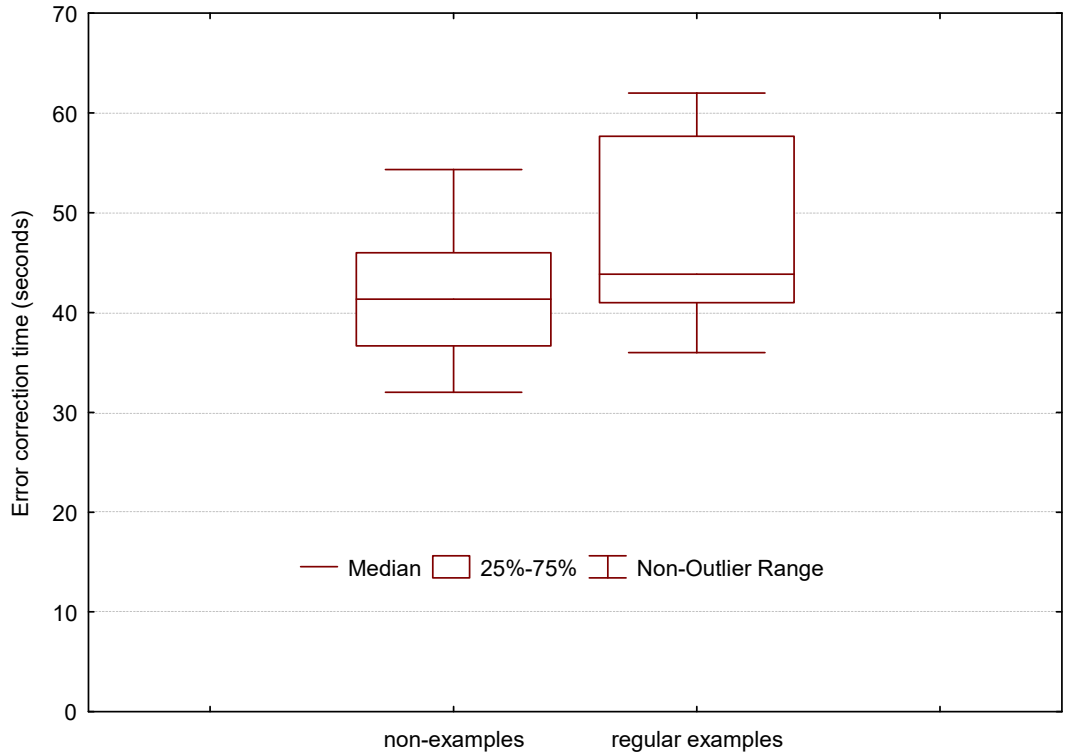


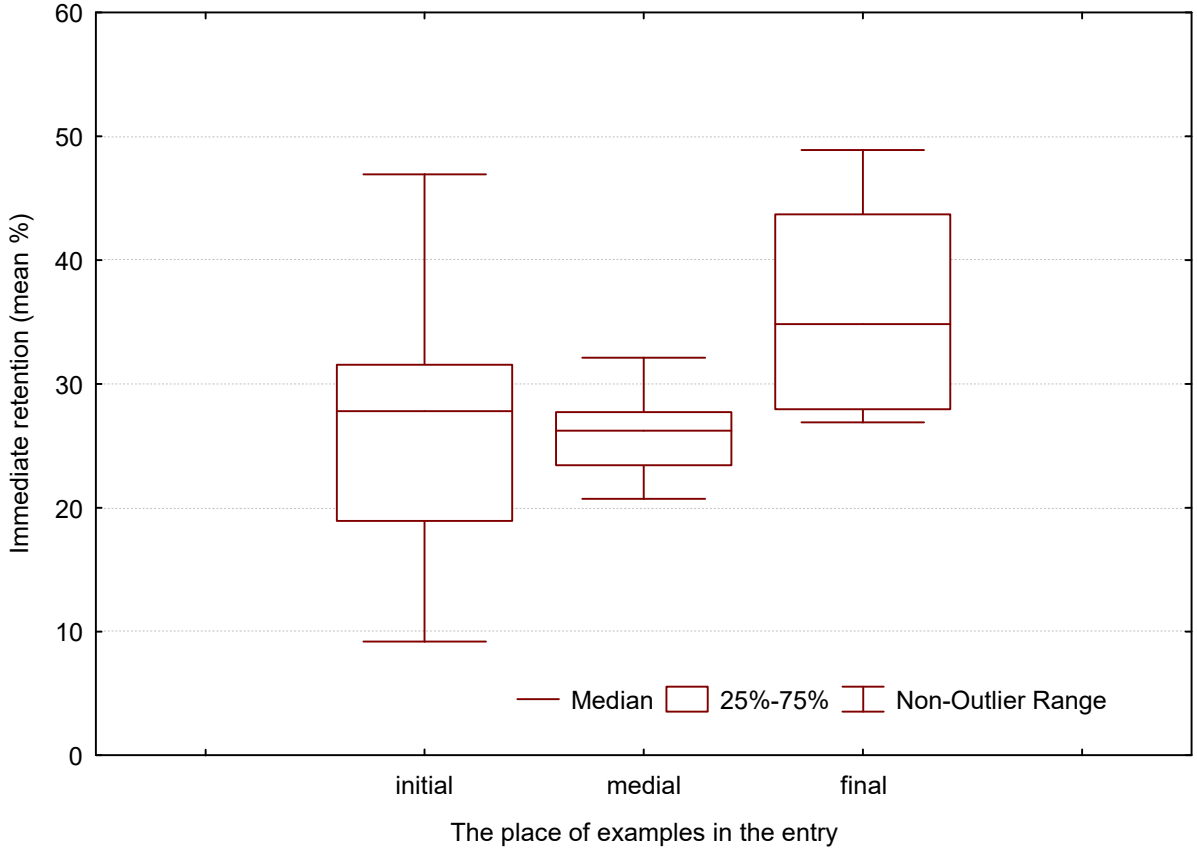
Figure 2. Error correction time for a test item by example type



Error correction time was not dependent on example types (F=1.09, p=0.32, partial η2=0.098) or distribution (F=0.43, p=0.65, partial η2=0.042) (Figure 2).

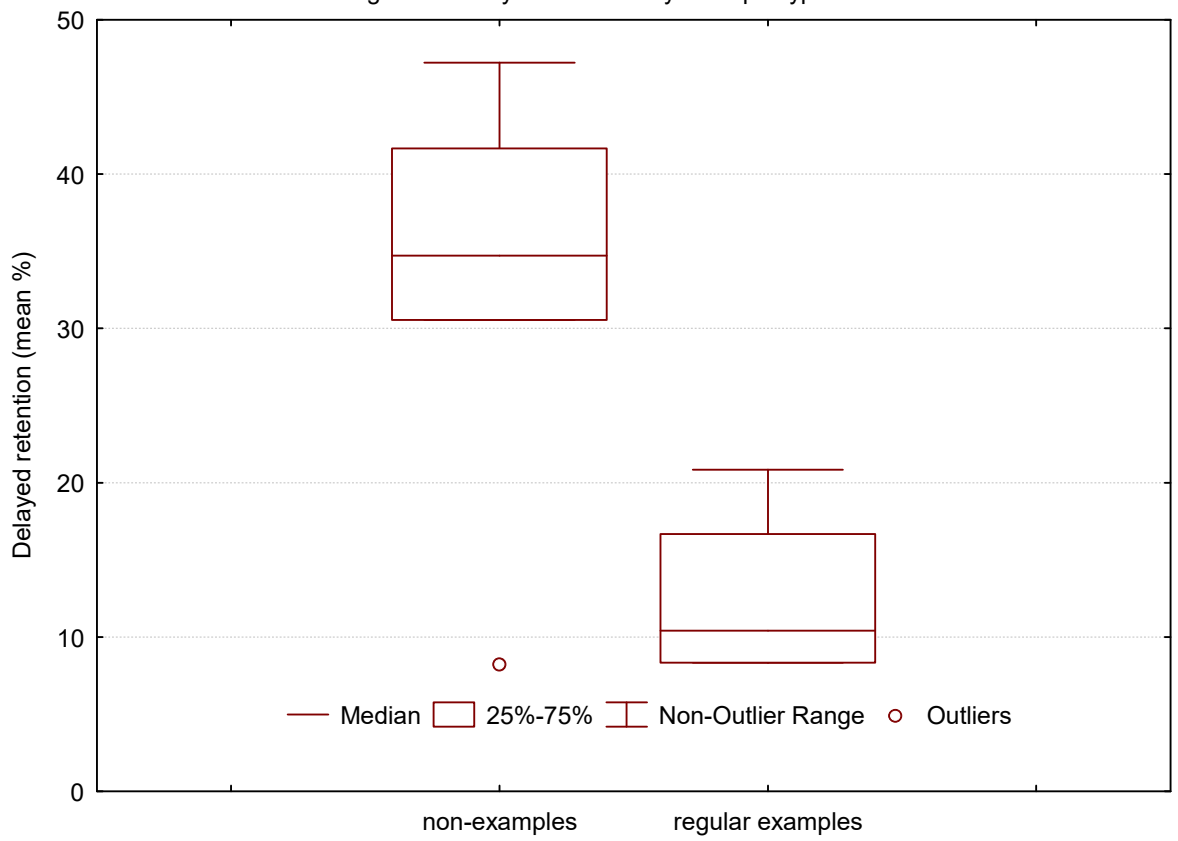
Non-examples did not affect the ability to correct errors immediately after exposure (F=2.92, p=0.12, partial η2=0.226), but the position of examples did (F=4.24, p=0.03, partial η2=0.298). Entry-final examples (36.19%) contributed about 40% more to the immediate retention of correct grammatical structures than those in the middle of the entry (26.07%, 36.19*100/26.07=138.79, p=0.04), and one third more than entry-initial ones (27.04%), but the latter difference was not significant (36.19*100/27.04=133.84; p=0.07).

Figure 3. Immediate retention by example position



Non-examples had a significant and strong effect on usage acquisition in the long run (F=11.62, p=0.01, partial η^2=0.537). The subjects exposed to them 2 weeks earlier corrected from memory over 160% more errors (32.87%) than the others (12.50%; 32.87*100/12.50=262.96). Example distribution did not play a significant role (F=0.39, p=0.68, partial η^2=0.038).

Figure 4. Delayed retention by example type



Non-examples increase error correction accuracy (Q1), but do not speed up task performance (Q2). They do not develop the ability to correct errors immediately after exposure, but help enormously to retain correct usage in the long run (Q3). Example distribution influences only the immediate retention of correct structures; entry-final examples help to remember them best (Q4). Research limitations and implications are discussed in the full paper.

Keywords: online dictionaries; online dictionary use; non-examples; examples; error correction; usage; learning; retention

Variation in idioms dictionaries: Lexicographer's work in the post-editing era

Jelena Parizoska[1], Ivana Filipović Petrović[2]

[1] Faculty of Teacher Education, University of Zagreb, Zagreb, Croatia
[2] Croatian Academy of Sciences and Arts, Zagreb, Croatia

E-mail: jelena.parizoska@ufzg.hr, ifilipovic@hazu.hr

Idioms present a major challenge in dictionary-making. In the digital era, the focus is mainly on automatic identification and extraction of idiomatic expressions (Gantar et al. 2018; Škvorc et al. 2020). On the other hand, one of the key issues facing lexicographers who compile dictionaries of idioms is variation. Corpus studies of idioms in different European languages show that they vary their lexico-syntactic structure regularly and this variation is systematic (e.g. Moon 1998; Cignoni et al. 1999; Fellbaum 2009; Omazić and Parizoska 2020). Furthermore, idioms are often creatively exploited in discourse (Naciscione 2010; Jaki 2014) and modifications, i.e. deliberate changes of an idiom's structure and/or meaning for communicative purposes, are widespread.

When it comes to idioms which display flexibility, the organization of dictionary entries fundamentally depends on the lexicographer's theoretical approach to idiomaticity in general and variation in particular. The aim of this paper is to highlight four practical issues of post-editing work involved in the lexicographic treatment of idiom variation, especially for morphologically complex languages: 1) finding all the conventionalized forms of an idiom in a corpus, 2) identifying the commonest variation, 3) distinguishing variations from modifications and 4) finding the most common patterns of modification for individual expressions. We will use entries in the Online Dictionary of Croatian Idioms (under development) as illustrative examples. In this dictionary, items with common lexis which have different grammatical forms are treated as variations (rather than individual expressions) and are listed in a single entry. This is in line with the cognitive linguistic view that variations present the same event in different ways. For instance, *biti u škripcu* 'be in a corner', *dovesti* koga *u škripac* 'back someone into a corner' and *izvući* koga *iz škripca* 'get someone out of a corner' all refer to variant ways of being, getting into or getting out of a problematic situation. Special boxes at the end of some entries include modifications and show what types of changes are typically made to an idiom when it is used creatively. For example, in the expression *za dva koplja iznad* (lit. two spears above 'a cut above') the word *dva* 'two' is regularly replaced by other numerals in order to show how much better someone or something is (e.g. *za sedam/deset/sto kopalja iznad* 'seven/ten/a hundred cuts above'). The reason for including modifications is that they are

constrained by an idiom's meaning and the meanings of the individual components (Langlotz 2006; Omazić 2015). In addition, modifications display some regularity. Thus, although particular lexical items replacing *dva* 'two' in *za dva koplja iznad* are not institutionalized, the modification is regular in that *dva* can be replaced by any other numeral, depending on the situation.

Overall, even though content is created automatically, post-editing work in idioms dictionaries requires a linguistic background and advanced knowledge of designing corpus queries so as to find variations and modifications. Ultimately, this means creating dictionary entries which show users that idioms are dynamic vocabulary items.

Keywords: idiom variation; online dictionary of idioms; corpus; Sketch Engine; Croatian

References

Cignoni, L., Coffey, S., & Moon, R. (1999). Idiom variation in Italian and English: Two corpus-based studies. *Languages in Contrast*, 2(2), pp. 279–300.

Fellbaum, C. (ed.) (2009). *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. London: Continuum.

Gantar, P., Colman, L., Parra Escartín, C., & Martínez Alonso, H. (2018). Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, 32(2), pp. 138–162.

Jaki, S. (2014). *Phraseological Substitutions in Newspaper Headlines: "More than Meats the Eye"*. Amsterdam, Philadelphia: John Benjamins.

Langlotz, A. (2006). *Idiomatic Creativity*. Amsterdam, Philadelphia: John Benjamins.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Naciscione, A. (2010). *Stylistic Use of Phraseological Units in Discourse*. Amsterdam, Philadelphia: John Benjamins.

Omazić, M. (2015). *Phraseology through the looking glass*. Osijek: Josip Juraj Strossmayer University of Osijek, Faculty of Humanities and Social Sciences.

Omazić, M., Parizoska, J. (2020). Sheme dinamike sile i promjenjivost glagolskih frazema [Force-dynamic schemas and variability of verbal idioms]. *Jezikoslovlje*, 21(2), pp. 179–205.

Škvorc, T., Gantar, P., & Robnik-Šikonja, M. (2020). MICE: Mining Idioms with Contextual Embeddings. arXiv preprint arXiv:2008.05759. (29 January 2021)

Taking a broad view of post-editing lexicography

Ana Frankenberg-Garcia

Centre for Translation Studies, University of Surrey, UK
E-mail: a.frankenberg-garcia@surrey.ac.uk

Can we post-edit lexicography? A quick search for verbs that collocate with *lexicography* as object of the clause in the 13-billion-word English Web 2020 corpus (Jakubíček et al. 2013) returns only 9 hits. None of those hits are for *post-edit*. The only verb shown is *specialise*. Closer inspection of the concordances generating the verb-noun collocation *specialise + lexicography* show that *specialise* is not in fact being used as a verb, but rather as a modifier, in contexts such as "her academic background is in specialised lexicography". This anecdotal example serves to illustrate three points: (1) there does not seem to be much that people do with lexicography, (2) the above automatic extraction of verbs that collocate with lexicography as object needs post-editing, and (3) post-editing lexicography is a recent concept, not attested in the English Web 2020 corpus. So what exactly does post-editing lexicography mean? This paper discusses two central strands of post-editing lexicography, taking a broad view of what it entails and the challenges that exist.

The first strand is the main focus of the eLex 2021 call for papers, i.e, the post-editing of automatically extracted lexicographic content. Even though the term post-editing is new in the field of lexicography, the idea that the role of the lexicographer has changed thanks to the development of corpora and corpus processing tools that facilitate lexicography is not new. Despite the great progress in automation anticipated in Grefenstette (1998) and put in practice through tools like Sketch Engine (Kilgarriff et al., 2014), skilled human intervention is still needed to interpret the complexity of natural languages and provide useful summaries for dictionary users (Rundell, 2002). Additionally, human expertise is essential to detect slips that evade automation and to help further improve the lexicographer's toolkit (Frankenberg-Garcia et al., 2020). As I am sure several other papers at eLex 2021 will demonstrate, this applies to various stages in e-lexicography, from the creation of corpora to the extraction and selection of good dictionary examples.

The second strand of post-editing lexicography I would like to discuss is the type of post-editing that becomes necessary for amending, adapting and extending published lexicographic contents. It is widely acknowledged that dictionaries need to evolve with time if they are to remain relevant. It follows that the changes implemented after a dictionary is published can also be regarded as a form of post-editing. These can be divided into three broad types: revision that becomes necessary because of (1) changes in language (e.g. adding new words, new senses, new collocations, etc.) (2) changes in society's views of language (e.g. revising definitions and labels), and (3) changes in technology (e.g.

adapting to new formats and the new potential brought about by technologies).

In an attempt to map this emerging topic, the paper addresses the above strands of post-editing lexicography with practical examples from real-world projects: the *Oxford Portuguese Dictionary*, the *Oxford English Dictionary*, the lexical database of academic collocations behind the *CollocAid* writing assistant (Frankenberg-Garcia et al. 2019, 2020), and the *Oxford Monolingual Portuguese Dictionary*.

The paper concludes with a discussion of the role of e-lexicography in facilitating all forms of post-editing. It argues that the need for both sophisticated lexicography skills on the part of dictionary editors, as well as an ongoing dialogue between editors and the developers of tools to assist lexicographers are central to the debate on post-editing lexicography.

Keywords: e-lexicography; updating dictionaries; lexicographer skills; post-editing

Acknowledgements

This research was funded by the UK Arts and Humanities Research Council (AHRC) (AH/P003508/1).

References

Elex 2021 Post-editing Lexicography (2021) https://elex.link/elex2021/

CollocAid (2021). Available at http://www.collocaid.uk/

Frankenberg-Garcia, A. Rees, G. and Lew, R. (2020) Slipping through the cracks in e-Lexicography. *International Journal of Lexicography*. Available at: https://doi.org/10.1093/ijl/ecaa022

Frankenberg-Garcia, A. Lew, R., Roberts, J., Rees, G. and Sharma, N. (2019) Developing a writing assistant to help EAP writers with collocations in real time, *ReCALL*, 32(2): pp 23-39. Available at: https://www.cambridge.org/core/journals/recall/article/developing-a-writing-assistant-to-help-eap-writers-with-collocations-in-real-time/91D2F3FBBA7E6DF7E46523F20E6ECCE7

Grefenstette, G (1998) The Future of Linguistics and Lexicographers: Will there be Lexicographers in the year 3000? *Proceedings of EURALEX*, 1998, Liege, Belgium. Available at: https://euralex.org/publications/the-future-of-linguistics-and-lexicographers-will-there-be-lexicographers-in-the-year-3000/

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013) The TenTen corpus family. *Proceedings of the 7th International Corpus Linguistics Conference*, Lancaster 2013, 125-127. Available at: http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

Kilgarriff, A., V. Baisa, J. Bušta, M. Jakubíček, V. Kovár, J. Michelfeit, P. Rychlý & V. Suchomel (2014) The Sketch Engine: Ten Years On. Lexicography 1: pp. 7–36.

Oxford English Dictionary (2021) Oxford: Oxford University Press. Avaialable at https://www.oed.com/

Oxford Languages Portuguese Monolingual Dictionary (2021) Oxford: Oxford University Press. Available through Google OneBox at https://www.google.com/

Oxford Portuguese Dictionary (2015). Oxford: Oxford University Press. Available at: https://premium.oxforddictionaries.com/

Rundell, M. (2002) Good Old-Fashioned Lexicography: Human Judgment and the Limits of Automation. In Correard, M.-H. (ed.), Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Grenoble: EURALEX, 138–155. Available at https://www.euralex.org/elx_proceedings/Lexicography%20and%20Natural%20Language%20Processing/Michael%20Rundell%20-%20Good%20Old-fashioned%20Lexicography%20Human%20Judgment%20and%20the%20Limits%20of%20Automation.pdf

An Evaluation of Definition Paradigms in Lexicography for Word Sense Alignment

Sina Ahmadi, John P. McCrae

Data Science Institute, National University of Ireland Galway
E-mail: firstname.lastname@insight-centre.org

Sense definitions are principal components of monolingual dictionaries describing various meanings of words in plain text. Since antiquity, there have been many theories and discussions on how to define a concept, i.e., *definiendum*, and the words and phrases which are used for this purpose, i.e., *definiens*. Durkin (2016) provides a description of such theories from historical, logical, and lexicographical points of view.

Dictionaries, as crucial resources for documenting languages, have been widely used in language technology and natural language processing. Given the increasing number of lexico-semantic resources thanks to community-driven initiatives such as Wiktionary[5] and Open Multilingual WordNet, the alignment of such resources is of importance to promote interoperability and facilitate the integration of various resources in a viable manner.

In the context of the word sense alignment task where word definitions are aligned automatically, we assume that retrieving the composing parts of sense definitions is useful to facilitate the alignment tasks. To this end, we carry out an evaluation of two analytical and relational paradigms on the MWSA English data (Ahmadi et al., 2020) containing annotated glosses of Webster's Dictionary 1913 and Princeton WordNet. The paradigms are defined as follows:

- Analytical definitions define a formal descriptive sentence consisting of four main components, namely species, verb, genus, and differentia.
- Relational definitions explain the meaning of a word in comparison to other entities, e.g., "extraneous (adjective)" defined as "not belonging to a thing" in Webster 1913.

As a preliminary study, we use a pattern-based approach as proposed by (Westerhout, 2010) where definitions are analyzed to retrieve genus and entity. This task is carried out using regular expressions with functional keywords, such as "opposite of", "belonging to" or "of or pertaining". Given definitions of a lemma with its part-of-speech in two dictionaries, in our case Webster 1913 and Princeton WordNet, two definitions are to be

[5] https://www.wiktionary.org/

aligned if they have an identical genus or entity after lemmatization.

Among the English aligned sense definitions, we select 100 definitions randomly among which 50 are alignable, i.e., specified with exact, and the rest are non-alignable, i.e., specified by none. Although non-alignable definitions are correctly classified in all cases, only 13 among the 50 other definitions are classified correctly. This indicates the poor performance of the pattern-based or symbolic approach for this task.

In addition to different phrase structures which lead to an unsimilar syntactic analysis, lexical choice determines the genus and entity of each definition. For instance, the definition of "angulation (noun)" as "the act of making angulate" and "making angular" use two semantically-related but different words "angular" and "angulate". On the other hand, descriptive phrases are missing in many definitions, as for "usurpation (noun)" defined as "wrongfully seizing" and "the act of usurping". It should also be noted that definitions are not of same granularity across resources.

Finally, we believe that such challenges which are faced in computational lexicography and natural language processing should be of interest to lexicographers and community-driven lexical content creators. This way, further computer-assisted techniques may be more efficiently integrated in the process of dictionary creation and compilation.

Keywords: electronic lexicography; natural language processing; lexical resource alignment

References

Ahmadi, S., McCrae, J. P., Nimb, S., Khan, F., Monachini, M., Pedersen, B. S., ... & Gabrovsek, D. (2020, May). *A multilingual evaluation dataset for monolingual word sense alignment*. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3232-3242).

Durkin, P. (Ed.). (2016). *The Oxford handbook of lexicography*. Oxford University Press.

Grosse, J., & Saurí, R. (2020). *Principled Quality Estimation for Dictionary Sense Linking*. In Proceedings of the XIX EURALEX conference.

Westerhout, E. (2010). *Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch*. Netherlands Graduate School of Linguistics.

Etymology in the Lanscape of Digital Lexicography

Yevhen Kupriianov[1], Iryna Ostapova[2], Volodymyr Shyrokov[2], Mykyta Yablochkov[2]

[1] National Technical University "Kharkiv Polytechnic Institute",
2 Kyrpychova str., 61002, Kharkiv, Ukraine
[2] Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine,
3 Holosiivska avenue, 03039, Kyiv, Ukraine
E-mail: eugeniokuprianov@gmail.com, irinaostapova@gmai.com, vshirokov48@gmail.com, gezartos@gmail.com

The digital environment offers completely new service tools for the researchers to manage linguistic material in a dictionary. First of all we refer to dictionary index systems. By dictionary indexation we mean a set of formalized rules and procedures to derive linguistic facts from the dictionary. These rules are implemented in the form of the user interfaces. However, it should be noted that the automatic building of the index schemes in a digital dictionary won't be effective without prior elaboration of a formal model reflecting its entire lexicographical structure.

Theoretically, every structural element of the entry shall be indexed. To achieve this goal the index elements have to be matched with the components of the dictionary formal model considered as lexicographic system. It is obvious that the advanced indexing technologies can be built only for lexicographic systems implemented in digital format.

The present paper describes the method and technology of indexing Etymological dictionary of the Ukrainian Language (EDUL).

The first volume of the dictionary was published in 1982 and the sixth saw the light in 2012. The text comprises 26,165 entries for developing of which 273 languages were involved. For each language a separate index was built with exact localization of each word or index unit. The total index volume is about 194,000 index units. At present the generated index texts are being edited.

For language indexing of the dictionary, there have been developed a special computer system, virtual lexicographic laboratory (VLL), which is adapted to the structure of the EDUL and intended for automatic building of the indices. The VLL created for EDUL provides the access to the entire dictionary text regardless of the publication time of particular volume and gives the opportunities for making various digital interpretations of etymological information.

To create the VLL the following steps have been made:

1. Development a formal model of EDUL lexicographic system.

2. Preparation of the EDUL digital text and identification of its meta-language signs marking the elements of its lexicographic system structure (the texts of all EDUL volumes were converted into HTML format and unified with respect to the file structure and sign system).

3. Elaboration of database structure corresponding to the structure of EDUL lexicographic system taking into account dictionary meta-language markers.

4. Automatic conversion of EDUL text into lexicographic database with elaborated structure.

The basic functions of VLL are as follows:

1) Access to the dictionary through the headword list and displaying of the entry structural elements in a tree form;

2) Editing any structural element of a dictionary entry;

3) Building a dictionary entry of defined structure;

4) Automatic indexing by language (or a set of languages defined by the user).

The dictionary of Ukrainian language (SUM-20; sum20ua.com) lacks etymological information element in its structure. A new edition of SUM-20 is being compiled in DWS (Digital Writing System). The integration with EDUL would enrich lexicographic structure of SUM-20 and give additional possibility of tracing word meaning changes.

Keywords: etymology, database structure, formal model, language index, virtual lexicographic laboratory

Construction of multilingual parallel corpora of literary texts using automatic alignment and crowd sourcing facilities

Galina Kedrova, Sergey Potemkin

Lomonosov Moscow State University, Russia, Leninskie gory 1, Moscow, Russia
E-mail: kedr@philol.msu.ru

Multilingual parallel corpora of literary texts are widely used both for training of translators and translation practice (Mosavi Miangah & Mohammadi Dehcheshmeh, 2012). Furthermore, literary texts' parallel corpora are also appreciated as the acquitted main source of intercultural communication and nations cultural heritage transfer as valuable part of Digital Literary Studies, and Digital Humanities in general (Ganascia, 2015). However, construction and annotation of literary texts' parallel corpora is a very challenging procedure for insufficiency of the results of automatic segmentation and annotation realized with existing NLP resources (Zanettin, 2017). One of the main challenges is the alignment annotation that universally involves some degree of manual intervention. To solve this problem in our project Multilingual parallel corpus of translations of A.P. Chekhov's works (Benko & Potemkin) we've combined the technology of automatic text annotation (Potemkin & Kedrova, 2010) and manual checking and correction of its results by the students of the Philological Faculty of Lomonosov Moscow State University within the framework of annual computer practice. In total we engaged there were more than 150 students studying from several major European languages departments to verify alignment of paragraphs and sentences in different translations of A.P. Chekhov's stories into English, German, French, Spanish, Italian, Swedish, Norwegian, Dutch, Portuguese and Finnish. As a result of the work done, each sentence / phrase in the translation texts was unambiguously aligned with the original textual fragment. Detailed analysis of the mismatches in alignment revealed that languages of translation differed for types and inventory of automatic alignment's errors, as well as with unintentional errors or deliberate rephrasing produced by the translators, e.g. omission of some phrases or addition of a new text to explain the meaning of the preceding paragraphs, etc. Practical result of the work was language dependent inventory of potential translation-resistant linguistic issues and a compendium of translational techniques used by various translators to transmit semantics and stylistics of the source text, which we expect plan to be replenished in the future. The developed method can also be applied to translations from a foreign language into Russian.

Keywords: Multilingual parallel corpora; Corpus of A. Chekhov's works, automatic alignment, errors in automatic alignment, crowdsourcing

References

Mosavi Miangah, T., & Mohammadi Dehcheshmeh, M. (2012). The effect of using parallel corpora on translation quality: a case study. Translation Studies, 9(36), 97-112.

Ganascia, J.-G. (2015). The Logic of the Big Data Turn in Digital Literary Studies. Frontiers in Digital Humanities. Vol. 2:7. DOI: http://dx.doi.org/10.3389/fdigh.2015.00007

Zanettin, F. (2017). Issues in Computer-Assisted Literary Translation Studies. Intralinea. Special Issue: Corpora and Literary Translation Issues in Computer-Assisted Literary Translation Studies.

Benko, V. & Potemkin, S. Corpus and dictionary of the language of A.P. Chekhov [Electronic resource] URL: http://www.philol.msu.ru/~serge/Chekhov/ (date of visit 01.13.2021).

Potemkin S. B. & Kedrova G. E. (2010). Alignment of un-annotated parallel texts // Proceedings of the II International Congress on Corpus Linguistics, Corunna, Spain, 2010.

Prospects on a workbench for the detection and analysis of semantic change: Towards automating the lexicographic process

Maike Park[1], Dominik Schlechtweg[2]

[1] Leibniz-Institute for the German Language, Mannheim

[2] Institute for Natural Language Processing, University of Stuttgart

E-mail: park@ids-mannheim.de, dominik.schlechtweg@ims.uni-stuttgart.de

A continuously growing interest in the practical application of natural language processing tools outside of their „playground“ has lead to a number of collaborative works between computational linguists and lexicographers or dictionary makers in the past years. So far, methods for the automated detection of new vocabulary and change of meaning have been applied to the task of identifying (new) lemma candidates (Falk et al., 2014; McCracken, 2015; Wanner et al., 2017; Klosa & Lüngen, 2018; Sørensen & Nimb, 2018; Waszink, 2019) and frequent semantically changed words (i.a. Cook et al., 2013; Fišer & Ljubešić, 2019) for the compilation or extension of dictionaries.

Approaches to the detection of infrequent new meaning and genuinely novel senses of a word are still limited though. On the one hand, both corpus and computational linguistic methods rarely succeed when it comes to the detection of infrequent words or meanings, because they tend to rely on frequency measures. On the other hand, most lexicographers and dictionary makers have neither the (human) means nor adequate tools to inspect infrequent candidates, because the results are oftentimes presented in the form of roughly edited lexical data, such as word lists, collocation pairs etc. (sorted by frequency), instead of systematically processed material that is easy to work with. These issues usually lead to the analysis of high-frequency candidates by one or two lexicographers, based on manually drawn corpus samples of occurrences of a new word or meaning.

Our approach integrates recent advances in computational linguistics into the lexicographic process in order to help with the detection of semantic change and to enhance inter-subjectivity of lexicographical decisions by developing a system that combines controlled human annotation (Schlechtweg et al., 2018) with theoretical work on lexical semantic change (Blank, 1997) and computational detection methods (Hamilton et al., 2016; Devlin et al., 2019): DURel is a freely available online annotation tool that uses human annotations of sentence pairs of a word to form sense clusters and visualize them over time, allowing lexicographers to investigate diachronic semantic changes, divergences of senses between language varieties or registers, vagueness of meaning or polysemy.

We will present results of first trial runs on German diachronic corpus data (Kurtyigit et

al., 2021) and discuss the potential of our approach to aid lexicographers in making or extending dictionaries.

Keywords: visualization; lexical semantic change; computational lexicography; natural language processing; lexicographic workflow; word sense disambiguation

Acknowledgements

Dominik Schlechtweg was supported by the Konrad Adenauer Foundation during the conduct of this research. The system's beta version was implemented by Annalena Streichert, Anne Reuter, Enrique Waldo Medina Castaneda and Lukas Theuer Linke.

References

Blank, A. (1997). *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Niemeyer: Tübingen.

Cook, P., Lau, J.H., Rundell, M., McCarthy. D., Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word senses. In *Kosem et al. (eds.) Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 49-65.

Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics", pp. 4171-4186.

Falk, I., Bernhard, D., Gerárd, C. (2014). From Non Word to New Word: Automatically Identifying Neologisms in French Newspapers. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 4337-4344.

Fišer, D., Ljubešić, N. (2019). Distributional modelling for semantic shift detection. In *International Journal of Lexicography 32 (2)*, pp. 163–183.

Hamilton, W., Leskovec, J., Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2116–2121.

Klosa, A., Lüngen, H. (2018). New German words: detection and description. In Čibej et al. (eds.) *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubliana, Slovenia: Ljubljana University Press, Faculty of Arts,

pp. 559-569.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., Schulte im Walde, S. (2021). Lexical Semantic Change Discovery. *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.* Available at: https://arxiv.org/abs/2106.03111 [Last accessed: 23.06.2021]

McCracken, J. (2015). Using machine learning for semi-automatic expansion of the Historical Thesaurus of the Oxford English Dictionary. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 211-235.

Schlechtweg, D., Schulte im Walde, S., Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 169-174.

Sørensen, H., Nimb, S. (2018). Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. In *Čibej et al (eds.) Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts.* Ljubliana, Slovenia: Ljubljana University Press, Faculty of Arts, pp. 819-826.

Wanner, L., Ferraro, G., Moreno, P. (2017). Towards Distributional Semantics-Based Classification of Collocations for Collocation Dictionaries. In *International Journal of Lexicography 30 (Vol. 2).* Oxford: Oxford University Press, p. 167-186.

Waszink, V. (2019). Neologisms in an Online Portal. *Abstract from the Globalex Workshop on Lexikography and Neologisms (GWLN 2019).*

Websites:

https://www.ims.uni-stuttgart.de/data/durel-tool
https://www.ims.uni-stuttgart.de/data/wugs

Gamifying the path to corpus-based pedagogical dictionaries

Tanara Zingano Kuhn[1], Rina Zviel-Girshin[2], Špela Arhar Holdt[3], Branislava Šandrih Todorović[4], Carole Tiberius[5], Ana Luis[1], Kristina Koppel[6], Danka Jokić[7], Iztok Kosem[3,8]

[1] CELGA-ILTEC, University of Coimbra, Portugal
[2] Ruppin Academic Center, Israel
[3] Centre for Language Resources and Technologies, University of Ljubljana, Slovenia
[4] University of Belgrade, Faculty of Philology, Serbia
[5] Institute for the Dutch Language, Netherlands
[6] Institute of the Estonian Language, Estonia
[7] Univesity of Belgrade, Serbia
[8] Jožef Stefan Institute, Slovenia
E-mail: akgaertig@units.it

The use of crowdsourcing techniques in lexicography has been gaining increased attention over the past few years. One of the most known methods for collecting information from the crowd is through crowdsourcing platforms such as Amazon Mechanical Turk, CrowdFlower or Pybossa, where participants contribute voluntarily or receive financial compensation for their contribution. Another possible way of crowdsourcing data involves the use of games, and this method is also known as Games with a Purpose (von Ahn, 2006). GWAPs were often designed to annotate or clear language data for the creation of various lexical infrastructures, for example JeuxDeMots (Lafourcade, 2007), Phrase Detectives (Poesio et al., 2013), Wordrobe (Venhuizen et al., 2013), ZombiLingo (Guillaume, 2016), Game of Words (Arhar Holdt et al., 2020; Kosem et al., 2020).

Nonetheless, the use of gamification, and crowdsourcing in general, in lexicography is still limited. The goal of our paper therefore is to report on a game that a group of researchers within the COST Action European Network for Combining Language Learning with Crowdsourcing Techniques (EnetCollect; CA16105; https://enetcollect.eurac.edu) has been developing in order to obtain example sentences to be used in pedagogical dictionaries of five different languages, namely, Dutch, Estonian, Serbian, Slovene, and Portuguese. With this game, we aim to learn what the crowd considers to be inappropriate for language learning material, and then use these sentences as a dataset to first train a binary machine learning model that will be able to automatically classify sentences as appropriate or inappropriate. Afterwards, we intend to train a multi-class classifier that would be able to perform fine-grained annotation of inappropriate sentences, according to the reason of their inappropriateness. The results will be used to create pedagogical corpora that can

serve, among other purposes, as a source of examples for pedagogical dictionaries.

Initially, experiments with the crowdsourcing platform Pybossa were performed, using automatically extracted sentences from web corpora of four different languages (Dekker et al, 2019; Zingano Kuhn et al., 2019). In those experiments, the crowd was asked to select the sentences that they considered to be offensive. The results indicated that a more straightforward, exact task for the crowd should be formulated and additional elements should be added to increase user involvement. We then decided to develop a multi-modes game in which players not only inform which sentences they consider to be inappropriate to language learning purposes, but also provide the reason for such a choice. For this, players have to indicate in which category or categories the selected example fits, ranging from sensitivity-related content to structural problems. The development of this game involves three stages, namely, data preparation, game preparation, and machine learning preparation, each one composed of a series of steps, and with defined expected outputs. Accordingly, each stage presents a great number of challenges and decision-making, from specific tasks to prepare the data, to definition of the game logic and computational development, to name but a few. In this paper, we share challenges faced, solutions found, and lessons learned, in order to provide a stepping stone to other colleagues willing to adopt this method in their lexicographical projects.

Keywords: crowdsourcing, corpora, gamification, pedagogical dictionaries

References

Arhar Holdt, Š., Logar, N., Pori, E., Kosem, I. "Game of Words": play the game, clean the database. (2020). In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (eds.). Lexicography for inclusion: EURALEX XIX: Congress of the European Association for Lexicography: 7-11 September 2021, Ramada Plaza Thraki, Alexandroupolis, Greece : proceedings book. Vol. 1, (EURALEX proceedings, ISSN 2521-7100). 2020 ed. [Poznań: European Association for Lexicography], pp. 41-49. https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-Preview.pdf.

Dekker, P., Zingano Kuhn, T., Šandrih, B., Zviel-Girshin, R., Arhar Holdt, Š., Schoonheim, T. (2019). In I. Kosem and T. Zingano Kuhn (eds.). *Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts.* Sintra, Portugal, 1-3 October 2019. Brno: Lexical Computing CZ s.r.o., pp. 84-85.

Guillaume, B., Fort, K., Lefebvre, N. (2016). Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. *Proceedings of the International Conference on Computational Linguistics (COLING).* Accessed at: https://hal.inria.fr/hal-01378980/ [07/05/2020]

Kosem, I. F. Martelli, R. Navigli, M. Jakubíček and Jelena Kallas (2020) *ELEXIS*

Deliverable D4.3 Crowdsourcing module. Available at: https://elex.is/wp-content/uploads/2020/03/ELEXIS_D4_3_Crowdsourcing_module.pdf

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *SNLP'07: 7th International Symposium on Natural Language Processing, Dec 2007*. Pattaya, Chonburi, Thailand, 7.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L. and Ducceschi, L. (2013). Phrase Detectives: Utilizing Collective Intelligence for Internet-scale Language Resource Creation. *ACM Transactions on Interactive Intelligent Systems 3*, pp. 3:1–3:44.

Venhuizen, N. J., Basile, V., Evang, K, Bos, J. (2013). Gamification for word sense labeling. *IWCS 2013*. Potsdam, Germany, pp. 397–403.

von Ahn, L. (2006). Games with a Purpose. *Computer* 39, 6 (June 2006), pp. 92–94. DOI:https://doi.org/10.1109/MC.2006.196.

Zingano Kuhn, T., Dekker, P., Šandrih, B., Zviel-Girshin, R., Arhar Holdt, Š., Schoonheim, T. (2019). Crowdsourcing corpus cleaning for language learning resource development. In S. Bibauw. *EUROCALL Conference 2019 "CALL and Complexity": Book of abstracts*. Louvain-la-Neuve: European Association of Computer Assisted Language Learning, 2019, p. 159.

How useful are writing assistants to researchers with English as a Second Language? A review of existing tools

Gustavo Zomer, Ana Frankenberg-Garcia

University of Surrey, United Kingdom

E-mail: g.zomer@surrey.ac.uk, a.frankenberg-garcia@surrey.ac.uk

Writing papers for publication is a cognitively demanding task, which can be even more demanding for researchers writing in English as a second language (Flowerdew, 2019). The use of tools to support writing can help these researchers (henceforth L2-English researchers) produce better texts, and they can at the same time reduce the cognitive burden of writing. Although there are various types of aids to writing available, their implementation varies across different writing assistants. Moreover, some writing assistant features seem more relevant to L2-English researchers than others, and there are further tools and resources which could be useful but are not integrated to text editors. This study presents a review of existing writing assistants from the perspective of the extent to which they can support L2-English researchers.

Although there are other reviews of writing assistants available (e.g. Tarp et al., 2017; Strobl et al., 2019), they tend to provide a generic overview, without detailing all the features each tool offers and without assessing the relevance of those features to a specific target audience. Another limitation is that reviews of tools pertaining to a fast-evolving field such as this one are in need of constant updating. In the present analysis, we review a total of 38 writing assistants available in 2020. They were selected based on a systematic online search using terms such as "grammar checker", "spell checker", "writing assistant", "proofreading tools", and alternative wordings of those terms. Only tools that were openly available for testing were considered.

The review focused on a systematic comparison of different functionalities (e.g. spelling check, grammar check, synonyms, predictive writing, readability score, lexical density analysis, corpus integration, and so on) and on the relevance of those features to assisting L2-English researchers. This was undertaken by developing a taxonomy with nine broad parameters of comparison and a total of 57 features within those parameters, and then pasting a text with different types of writing problems into the text editor of each assistant in order to assess how they performed.

The results show that the writing assistants under review deal mainly with error correction, offering only a limited number of features targeted at improving academic

writing and addressing the specific needs of writers using English as a second language.

The contributions of this study are two-fold. First, the methodology proposed and its underlying taxonomy can be useful to future reviews evaluating the fast-growing number of writing assistants becoming available. Second, the results of the present review point towards areas in need of improvement when it comes to developing writing assistants specifically aimed at helping L2 researchers..

Keywords: Writing assistants; L2 English; Academic English; EAP; L2 writing

References

Flowerdew, J. (2019). The linguistic disadvantage of scholars who write in English as an additional language: Myth or reality. Language Teaching, 52(2), pp. 249-260.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A. and Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. Computers & Education, 131, pp. 33-48.

Tarp, S., Fisker, K. and Sepstrup, P. (2017). L2 writing assistants and context-aware dictionaries: New challenges to lexicography. Lexikos, 27, pp. 494–521.

Old Russian Plant Names Dictionary: technical aspects

Kira I. Kovalenko, Valeria B. Kolosova

Institute for Linguistic Studies, Russia

E-mail: kira.kovalenko@gmail.com, chakra@eu.spb.ru

Any dictionary compilation so far has been extremely time consuming and sometimes tedious activity. Nowadays computer technologies make it easier and reduce the routine actions. Starting a new historical dictionary – the Dictionary of Russian phytonyms of the 11th-17th centuries – we decided to use the benefits of modern methods and visualize some lexicographic data.

At the time, Old Russian plant names are represented in such multivolume lexicographical dictionaries as the Old Russian language Dictionary (11th–14th cc.), the Russian language of the 11th–17th cc. Dictionary, and the Quotidian Russian of Muscovite Rus' of the 16th–17th cc. Dictionary. But these dictionaries have a number of drawbacks: 1) a number of phytonyms for various reasons are not included, even if they are represented in the card Index; 2) there are errors in the identification of phytonyms; 3) only the first fixation is indicated, without specifying the further fate of the word in the language; 4) there are no references to other names of the same plant; 5) the dictionaries are still being compiled, and plant names stated on the last letters of the alphabet are unavailable (Kovalenko et al., 2018].

As the first stage of the Old Russian plant names dictionary compilation, the *PhytoLex* database was created (https://phytolex.iling.spb.ru). At the moment it contains more than 16,000 records of plant names with citations and all the information about their sources (Kolosova et al., 2018a, 2018b]. As the dictionary is planned initially in electronic form, its size is unlimited and may include not only words that have been in use for a long time, but also those whose existence in Russian was short, which was typical, for example, for a number of latinisms widely used in the documents of the Apothecary Chancery (Olekhnovich, 2018).

Word entries of the electronic plant names dictionary are being compiled in *Lexonomy* — a cloud-based, open-source platform for writing and publishing dictionaries (Měchura, 2017). The data from the *PhytoLex* database is downloaded as a *csv* file, then processed by a programme in *Phython,* which sorts citations by time and returns lexicographical data in the TEI markup. The compilers' task is to pick up the most representative citations for word entries. Various Python libraries allow visualizing lexicographic data and creating statistics. It gives an opportunity to fix not only the time of the word appearance in Russian, but also the time of its disappearance in case the phytonym did not take root in the language and was replaced by a synonym. All that helps to extend the boundaries of

the traditional dictionary and visualize the Russian botanical terminology development.

Keywords: Old Russian language; plant names; phytonyms; lexicography; dictionary compilation; visualization

References

Dictionary of Quotidian Russian of Muscovite Rus of the 16th–17th cc. [Slovar' obihodnogo russkogo jazyka Moskovskoj Rusi XVI–XVII vv.] St. Petersburg, 2004—.

Dictionary of the Old Russian language (11th–14th cc.) [Slovar' drevnerusskogo jazyka (XI–XIV vv.)]. Moscow, 1988—.

Dictionary of the Russian language of the 11th–17th cc. [Slovar' russkogo jazyka XI–XVII vv.] Moscow, 1975—.

Kolosova V., Zaytseva K., Kovalenko K. PhytoLex – the Database of Russian Phytonyms: from Idea to Implementation // SlaviCorp 2018. 24–26 September 2018. Charles University, Prague. Book of Abstracts. 2018b. P. 88-90. [https://slavicorp.ff.cuni.cz/wp-content/uploads/sites/144/2018/09/SlaviCorp2018_Book_of_Abstracts.pdf]

Kolosova V., Zaytseva K., Kovalenko K. PhytoLex: conception and technical implementation // El'Manuscript 2018a. 7th International Conference on Textual Heritage and Information Technologies. Vienna and Krems, Austria, 14-18 September 2018. Abstracts, Participants, Programme. P. 49-50.

Kovalenko K.I, Kolosova V.B., Shchekin A.S. Historical dictionaries as sources of the PhytoLex — Russian phytonyms database (XI–XVII cc.) [Istoricheskie slovari kak istochniki bazy dannyh russkih fitonimov PhytoLex (XI-XVII vv.)]. In: Rossijskaya akademicheskaya leksikografiya: sovremennoe sostoyanie i perspektivy razvitiya. St. Petersburg, 2018. P. 253-362.

Měchura M.B. 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, 19-21 September 2017, Leiden, The Netherlands (https://www.lexonomy.eu/docs/elex2017.pdf)

Olekhnovich O.G. Vliyanie latinskogo yazyka na formirovanie russkih fitonimov (na primere "Dokumentov Aptekarskogo prikaza" XVII v.). In: Sovremennaya nauka: aktual'nye problemy teorii i praktiki. Seriya «Gumanitarnye nauki». 2018. 12-3. P. 110-113.

Revising *Dicionário Olímpico*: A log-files analysis of a Brazilian frame-based online dictionary

Bruna da Silva[1,2], Rove Chishman[1], Gilles-Maurice de Schryver[2,3]

[1] Unisinos University, Brazil
[2] Ghent University, Belgium
[3] University of Pretoria, South Africa
E-mail: bruna.dasilva@ugent.be, rove@unisinos.br, gillesmaurice.deschryver@ugent.be

Although still underused in Digital Lexicography, log-files analysis has great potential for empirical dictionary-use research through free implicit feedback and unobtrusive monitoring (de Schryver & Joffe 2004). In the context of revising and improving Dicionário Olímpico (Chishman and colleagues 2016), a log-files analysis using AWStats (2000-2019) was designed to investigate user behaviour wrt patterns of navigation and, more specifically, page views. *Dicionário Olímpico* is an online dictionary based on the theoretical foundations of Frame Semantics (Fillmore 1982) that describes the lexicon of the 40 Summer Olympic sports. The present work reports on the main outcomes of a pilot analysis and points to future perspectives for broadening the research scope to include all the available data. For the pilot analysis we had, amongst others, only access to the top 1,000 pages visited per month. Also, the usage data collected for this experiment corresponds to just three months (August, September, and November), the months with the highest average unique visitor numbers, for the years 2016-2019. Bearing in mind *Dicionário Olímpico*'s lexicographic structure (i.e., the fact that the tool presents three main levels with information on sports, scenarios, and words), the results of this investigation are divided into three categories.

Regarding the SPORTS level, the analysis revealed that not all 40 sports pages appear in the top 1,000 viewed pages' list. At the same time, the total number of viewed sports decreases throughout the four-year period, while the number of views per sport that is consulted increases. Exceptionally popular is rhythmic gymnastics, whose page-view numbers are 2.5 times higher than the second sport in the list, being volleyball. Results concerning the SCENARIOS level indicate that only around half of the total number of 780 scenarios appear in the top 1,000 viewed pages; but also that the total number of viewed scenarios remains stable over time. For the WORDS level, the logs indicate that users only explore a small part of them: on average 508 per month, out of a total of 3,930 words. In general, the results from this initial exploration point to the following major finding: most dictionary data is not (frequently) seen by users. This begs the question: Do users not see that data because they do not need it, or do they not see it because they fail

to find it? Lexicographers secretly hope that all their work may be of use at some point in time, so we shall assume for now that users simply failed to navigate to 'all' the dictionary contents. Therefore, even though it is still necessary to broaden the scope of the research in order to be more conclusive on what the search patterns really imply, the current main preliminary finding leads to the conclusion that strategies need to be developed to make more of the dictionary contents findable. Ways to encourage users to consult unexplored parts of the dictionary could be to include a "sport/scenario/word of the day" function. Another finding is that as users follow the hierarchical dictionary structure inherent to the Frame Semantics approach (sports > scenarios > words), page-view numbers decrease. On the one hand, one could claim that this is to be expected, considering that there are more scenarios than sports and more words than scenarios. On the other hand, there could be something fundamentally wrong with the navigation method itself, making users lose patience before reaching the word level. Currently, the full lists of usage logs regarding the entire period the dictionary has been available online, including data from 2020 and 2021, are being analysed, and this for all the pages viewed, not just the top 1,000 per month, for a selection of months. In order to answer the question about the usefulness of the inherent dictionary structure, particular attention will also go to a study of the referrals from search engines: do these typically refer to pages with sports, scenarios or words?

A limitation of log-files studies in general concerns the gathering of qualitative data that could reveal to what extent users are satisfied or dissatisfied with the dictionary's content. Adding ways to collect explicit feedback, such as via online feedback forms, e-mail invites, user interaction buttons, or even emoticon-based Likert scales – cf. de Schryver & Joffe (2004), Klosa & Gouws (2015), Liu (2017), and Efthimiou et al. (2019), respectively – could contribute to fill this gap. The dictionary interface is currently being redesigned to allow for such qualitative feedback as well.

Keywords: digital lexicography; online dictionary; sports terminology; log files; usage research.

References

AWStats. (2000-2019). AWStats – Open source log file analyzer for advanced statistics. Available from https://awstats.sourceforge.io/

Chishman, R. and colleagues. 2016. Dicionário Olímpico. São Leopoldo: Unisinos. Available from http://www.dicionarioolimpico.com.br/

de Schryver, G.-M. & D. Joffe. 2004. On how electronic dictionaries are really used. In: Williams, G. & S. Vessier (eds). Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004, pp. 187–96. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Efthimiou, E., S.-E. Fotinea, T. Goulas, A. Vacalopoulou, K. Vasilaki & A.-L. Dimou.

2019. 'Sign Language technologies and the critical role of SL resources in view of future Internet accessibility services', Technologies 7(1), pp. 1–21.

Fillmore, C. J. 1982. 'Frame Semantics'. In: The Linguistic Society of Korea (ed.), Linguistics in the Morning Calm, pp. 111–37. Seoul: Hanshin Publishing Co.

Klosa, A. & R. H. Gouws. 2015. 'Outer features in e-dictionaries', Lexicographica: International Annual for Lexicography 31, pp. 142–72.

Liu, X. 2017. 'Multimodal exemplification: The expansion of meaning in electronic dictionaries', Lexikos 27, pp. 287–309.

About users needs of a lexicographic tool for Academic Writing in Spanish

Margarita Alonso-Ramos, Eleonora Guzzi

Universidade da Coruña, Spain

E-mail: margarita.alonso@udc.es, eleonora.guzzi@udc.es

In recent years we have built a lexicographical tool oriented to the production of academic texts in Spanish, – the *Herramienta de Ayuda a la Redacción de textos Académicos*, henceforth, HARTA–based on a corpus of academic texts (Alonso-Ramos et al., 2017, García-Salido et al., 2018). HARTA focuses on collocations (e.g. *extraer una conclusion* 'to draw a conclusion') and on what we call *formulas* (*formulemes* in Mel'čuk, 2015). By *formulas* we mean expressions such as *dicho de otra manera* ('to put it differently'), or *hay que destacar* ('it must be stressed'), which are not registered in Spanish dictionaries, but, however, they are widely used in academic discourse. In accordance with the current trends in lexicography, we designed a combined dictionary-corpus tool (Paquot, 2012; Asmussen, 2013; Verlinde & Peeters, 2012) in the belief that, in many cases, user queries are more easily answered by showing examples of a given lexical combination, rather than by offering a whole lexicographic description. The current interface of HARTA includes a different treatment for collocations and formulas, because we claim that users will employ different strategies to look them up: in the first case, users query for a specific collocate that combines with a base (e.g. looks for the verb which combined with *conclusion* means 'to conclude', i.e. *to draw*); in the second case, users look for an expression which fulfils a given discourse function (e.g. expressions such as *in other words*, *put it differently*, *that is to say*, etc. are all reformulative expressions).

This distinction, as well as the ease of corpora as a way to solve lexicographic needs, have to be proven with users. Even though HARTA is still an ongoing project, we think it is worthwhile to test it with university students as well as professional users in order to verify if it meets their needs. Since the interface was put on line, we submitted the tool to a pilot test with a small number of informants (university students with Spanish as L1). Preliminary results show that informants do not take advantage of all the information included in HARTA and they tend to use the tool as an "answer key": they verify if their choice appears in HARTA, but they do not confirm if their choice fits well in the task. These first results lead us to rethink how to evaluate if HARTA can meet the needs of students while writing. Since users are not always aware of their own needs (Tarp, 2009), we must begin by identifying them. With this aim, we set up an experimental study with a more qualitative approach where informants' task is to propose alternative lexical combinations first without any help of lexical resources and later using HARTA. We

combine screen recording in conjunction with a thinking-aloud task (Müller-Spitzer et al., 2018). In the full version of our presentation, we will provide a detailed analysis of the results of the study.

Keywords: academic writing; collocations; formulas; dictionary-corpus tool; user needs

References

Alonso Ramos, M., García-Salido, M., & Garcia, M. (2017). Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pp. 571-586.

Asmussen, J. (2013). Combined Products: Dictionary and Corpus. In R. Gouws, U. Heid,W. Sheweickard & H. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography.* Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin/Boston: De Gruyter Mouton, pp. 1081–90.

García-Salido, M., Garcia, M., Villayandre, M., & Alonso-Ramos, M. (2018). A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In Calzolari, N. (Conference chair), Choukri, K., Cieri, C., Declerck, T., Goggi, S., Koiti Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. & Tokunaga, T. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 260- 65, Miyazaki, Japan.

Mel'čuk, I. (2015). Clichés, an Understudied Subclass of Phrasemes. *Yearbook of Phraseology*, 5, pp. 35–50.

Müller-Spitzer, C., M.J. Domínguez, M. Nied, I. M, Silva, & S. Wolfer, (2018), Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources, *Lexikos* 28, pp. 287-315.

Paquot, M. (2012). The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. In S. Granger & M. Paquot (eds.) *Electronic Lexicography.* Oxford University Press, pp. 163–185.

Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos*, 19(1), pp. 275–296.

Verlinde, S. & Peeters, G. (2012). Data access revisited: the Interactive Language Toolbox. In S. Granger & M. Paquot (eds.) *Electronic lexicography.* Oxford: Oxford University Press, pp. 147–162.

Dictionary of Core Academic Vocabulary based on Czech Academic Word List Akalex

Dominika Kováříková[1], Oleg Kovářík[2]

[1] Czech National Corpus, Charles university, Panská 7, 11000 Prague

[2] Datamole Inc., Banskobystrická 11, 16000 Prague 6

E-mail: dominika.kovarikova@ff.cuni.cz, oleg.kovarik@gmail.com

Several lists of academic words and academic phrases have emerged over the past two decades (Coxhead, 2000; Paquot, 2010; Gardner & Davies, 2014; Morley, 2014), most of them English-oriented. Similarly to the well renowned lists (listed above), the Czech Akalex list (Kováříková & Kovářík, 2021) of about 1,000 single-word and multi-word expressions is based on frequency and distribution criteria. The material for the academic word list is a representative corpus of written Czech SYN2015 divided into a subcorpus of academic texts (11 million words) and a reference subcorpus of fiction and journalism texts (83 million words). Only words that are at least 3x more common in academic than non-academic texts have been included in the Akalex list; all entries must be relatively frequent in academic texts (at least 20 instances per million words); and all words must be attested and evenly distributed in at least 20 of 24 academic disciplines available in SYN2015. Relatively simple criteria produced outstanding and convincing results comparable to other lists of academic words. The criteria were chosen to produce a list similar in extent to the Academic Keyword List (930 words and MWUs; Paquot, 2010) for comparison purposes.

Based on the Akalex academic word list, an online dictionary of core Czech academic vocabulary is being produced, which, in addition to the headwords themselves, will contain other information relevant to target users, namely frequency information, meaning (for lower-level users), the most common academic collocations and synonyms. In addition, the dictionary will also include translation equivalents in English for professional academic writers. The future plan is to provide equivalents for other languages, so that this list can serve students and academics of various philological disciplines.

For compiling the dictionary, we use several online corpus tools available at the Czech National Corpus web page. The database of translation equivalents Treq (Škrabal & Vavřín, 2017) is used to search for relevant translations and synonyms, and the application Word at a Glance (Machálek, 2020) provides a basic overview of the searched word including collocations and similarly used words.

The online core dictionary of academic Czech is designed to serve two main purposes. Firstly, it is a practical tool to facilitate the challenging process of writing academic texts

for both students and professionals. Secondly, it can be used in teaching academic skills to undergraduate students and in teaching (academic) Czech as a second language. The target users are therefore university students and professionals in all academic fields.

Keywords: core academic vocabulary, academic word lists, academic lexicography, corpus tools

References

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2), pp. 213–238.

Gardner, D., Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics*, 35(3), pp. 305–327.

Kováříková, D., Kovářík, O. (2021): *Akalex: Czech Academic Word List*. Prague: Institute of the Czech National Corpus. Available at: www.korpus.cz/akalex.

Machálek, T. (2020). Word at a Glance: Modular Word Profile Aggregator. In: *Proceedings of LREC 2020*, pp. 7011–7016.

Morley, J. (2014): *Academic Phrasebank: A compendium of commonly used phrasal elements in academic English in PDF format*. Manchester: University of Manchester.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New-York: Continuum.

Škrabal, M. – Vavřín, M. (2017): Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii*, 99(2), pp. 245–260.

Accessible Lexicography: A free online school dictionary of Greek accessible for visually-impaired senior elementary children

Zoe Gavriilidou, Apostos Garoufos

Democritus University of Thrace, Greece

E-mail: zoegab@otenet.gr, apogaroufos@gmail.com

This paper reports on the compilation of a monolingual online Greek school dictionary targeted at children aged between 9 and 12 years old which is accessible for visually-impaired children.

School dictionaries, like the print dictionary *Το πρώτο μου Λεξικό* (My first dictionary), for children aged 6-9 or *Το Λεξικό μας* (Our Dictionary) for children aged 9-12, compiled by the Greek Ministry of Education in the frame of the reform of school curricula, are dictionaries designed to be used by school children and adapted to their mental, linguistic, cultural, and encyclopedic development (Tarp & Gouws, 2012) and their use in classroom may facilitate language learning, vocabulary acquisition, reading comprehension or writing skills. However, their use in schools is not always successful, either because they do not address the exact needs of specific target groups of school-age children, or because pupils are not strategic dictionary users (Gavriilidou, 2013; Gavriilidou et al., 2020) and lack important reference skills that would allow them to make quick and successful searches in the dictionary (Tarp, 2011; Chadzipapa et al., 2020).

Furthermore, in the case of visually-impaired children which are educated in mainstream primary schools in Greece such print school dictionaries are not accessible, and no other adapted lexicographic materials exist. Additionally, ICT supported tools or e-learning resources are lacking in Greek schools to support learning of this specific population, so there is an urgent need for the creation of tools and resources tailored for the needs of children with visual impairments that would, however, be cost-effective, given that the costs of creating such materials is often economically not justifiable due to a small number of visually-impaired children.

To address the above mentioned needs and help remove barriers to learning of visually-impaired children, we are in process of compilation of the *Online school-dictionary of Greek* (OSDG). The OSDG, when completed, it will comprise 5000 lemmas and is the first online School dictionary which, additionally, is accompanied by a novel application which allows the visually impaired dictionary user, with the assistance of a keyboard shortcut which converts it in a Braille basic keypad, to easily enter the online dictionary

and search for any entry by using only six buttons on the keyboard as (s)he does with a Braille typewriter. In this way, the user is not obliged to use or learn the QWERTY keyboard. An audio reading of the looked-up entry is provided.

The OSDG is **user oriented** and **corpus-based.** The headword selection was based on word frequency of entries in a special corpus containing all school books of 4th to 6th Grades of Greek elementary schools. This corpus also ensured the age appropriateness of the definitions. Finally, to ensure user friendliness and age-appropriate cognitive load, the microstructure of the dictionary includes information about the Part of Speech, synonyms or antonyms, phraseology and word families.

Keywords: school dictionary; pedagogical Lexicography; visually-impaired children; accessible lexicography

References

Chadjipapa, E., Gavriilidou, Z., Markos, A. & A. Mylonopoulos (2020), The effect of gender and educational level on dictionary use strategies adopted by upper-elementary and lowersecondary students attending Greek schools, *International Journal of Lexicography*, https://doi.org/10.1093/ijl/ecaa012

Efthymiou, A., Dimos, Η., Mitsiaki, Μ. & Ι. Antipa, (2002), *An illustrated Dictionary for Α΄, Β΄, C΄ Classes of elementary School: My first Dictionary,* Athens [In Greek].

Gavriilidou, Z., 2013. 'Development and Validation of the Strategy Inventory for Dictionary Use (S.I.D.U.).' *International Journal of Lexicography* 26.2, pp. 135-153.

Gavriilidou, Z., Mavrommatidou, S. & A. Markos (2020), The effect of gender, age and career orientation on digital dictionary use strategies, *International Journal of Research Studies in Education*, 9(6), pp. 63-76. https://doi.org/10.5861/ijrse.2020.5046

Kapsalis, G., Paschalis, A., Tsialos, S. & D. Goulis, (2020) Our dictionary for *Δ΄, Ε΄, F΄ Classes of Elementary School*, Athens [In Greek].

Tarp, S. (2011), Pedagogical Lexicography: Towards a New and Strict Typology Corresponding to the Present State-of-the-Art, *Lexicos*, 21(1), pp. 217-231.

Tarp, S. & R. Gouws, (2012) School Dictionaries for First- Language Learners, *Lexicos*, 22, pp. 333-351.

Learner lexicography from a phraseological perspective - Converting a corpus-derived phraseme list into an electronic EFL reference tool[6]

Anna Fankhauser

Osnabrück University, Neuer Graben 40, 49069 Osnabrück

E-mail: anna.fankhauser@uni-osnabrueck.de

The significance of phraseological items (i.e. relatively fixed multiword units or formulaic language (cf. Stein, 2007)) for English as a Foreign Language (EFL) learners is widely accepted: Research suggests that formulaic language promotes nativelikeness and fluency in learners' language production and fosters the efficiency of language processing and acquisition in the context of foreign language learning (cf. Pawley & Syder, 1983, Wray, 2002, Nation & Shin, 2007, Martinez & Schmitt 2012). Yet, we still lack a systematic approach to integrating relevant phraseological items into EFL teaching and learning.

The current paper presents an extensive corpus study addressing this problem by systematically defining a core of relevant British English and American English phraseological items and outlines how the findings of the study are edited, presented and visualized in order to allow for their integration into EFL teaching and learning.

As a first step towards establishing a corpus-derived core of high- and mid-frequency phrasemes, large corpora of spoken British and American English were compiled. Subsequently, 2 to 6-gram lists were generated with the Sketch Engine and items above a threshold level previously set according to Nation's findings about vocabulary size for L2 learners were extracted (cf. Nation, 2006, Schmitt, 2008). An additional set of pedagogically relevant criteria was established for manually selecting the phrasemes to be incorporated into the list.

To make the results of the corpus study available for EFL teaching and learning, the corpus-derived list was converted into a reference format applicable to individual situations of language production and reception. The lexicographic editing of the list involved re-grouping the hitherto frequency-ordered items according to their functional and pragmatic characteristics to which end a new model of functional phraseme categorization based on Burger (2015) was developed. After categorizing the items into referential (approximately 3,700 items), structural (approximately 700 items) and communicative (approximately 3,800 items) phrasemes (cf. Burger, 2015) and further assigning them to pedagogically

[6] Following Donalies (1994) and Burger (2015), the term "phraseme" is used as an umbrella term referring to phraseological expressions as defined in the abstract's first paragraph.

relevant subcategories such as routine formulae (speech acts including greeting, agreeing and prompting) and speech formulae (pragmatic and discourse markers) or denominating and descriptive formulae, additional information categories were added. These include structural descriptions of items, part of speech equivalents, context information, topical fields, degree of idiomaticity, as well as corpus examples and context-specific translations. For example, the expression *go behind sb.'s back*, was categorized as a referential phraseme and assigned to the functional subcategory of 'denominating events/actions/developments'. Additionally, the topical information 'interpersonal/family relations' and the category 'figurative idiom' were included in the item's description. Items displaying multiple functions were listed in each of the categories to which they could be assigned. While the functional categorization is chosen as the default setting of the electronic reference list, anyone involved in the learning process (e.g. teachers, learners, material designers, etc.) can re-order the items according to any of the information categories to ensure the list's maximum usability for varying situations of usage and for the concrete focus of a specific learning unit.

EFL teachers and learners are thus not only provided with a systematic core of useful phraseological items but also with a lexicographic reference tool that facilitates individualized learning output.

Keywords: Learner Lexicography; English as a Foreign Language; Corpus-Based Phraseology; EFL Reference Tool

References

Burger, H. (2015). *Phraseologie. Eine Einführung am Beispiel des Deutschen* (5th ed.). Erich Schmidt Verlag.

Donalies, E. (1994). Idiom, Phraseologismus oder Phrasem? Zum Oberbegriff eines Bereichs der Linguistik. *Zeitschrift für germanistische Linguistik: deutsche Sprache in Gegenwart und Geschichte 22*(3), pp. 334-349.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), pp. 299–320.

Nation, P., & Shin, D. (2007). Beyond single words: the most frequent collocations in spoken English. *ELT Journal 62*(4), pp. 339-348.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes 63*(1), pp. 59-82.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and native-like fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication*, pp. 191-226. Longman.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research 12*(3), pp. 329-363.

Stein, S. (2007). Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In H. Burger (Ed.), *Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung, vol. 1*, pp. 220-236. De Gruyter.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Multiword expressions in the Ekilex data model

Arvi Tavast, Jelena Kallas, Margit Langemets, Kristina Koppel
Institute of the Estonian Language, Tallinn, Estonia
E-mail: arvi.tavast@eki.ee, jelena.kallas@eki.ee, margit.langemets@eki.ee, kristina.koppel@eki.ee

Language units presented in dictionaries form a continuum of length and complexity from simple headwords to definitions. Unlike the running text of paper dictionaries, structured lexical databases need to assign each data element to its place in the data model, based on user needs and theoretical considerations. Current initiatives dealing with these topics include Ontolex-Lemon, TEI Lex-0, LMF, ELEXIS and LEXIDMA, as well as Ekilex (Tavast et al., 2018; Koppel et al., 2019; Tavast et al., 2020).

For the purposes of this paper, we define a multiword expression (MWE) as a language unit that contains other language units of the same dictionary, whether its properties are predictable from its components or not, including derivatives, compounds, collocations, phrasemes, multi-word terms and idiomatic expressions. The breadth of this definition is characteristic of its object: commonly used classifications of MWEs are theory-dependent, fuzzy and not mutually exclusive. A particular MWE can drift between categories as the language develops, or even belong to more than one category simultaneously, adding complexity and subjectivity to the lexicographer's decision process. Example: *punane vein* 'red wine' used to be a collocation for *vein*, but has now been included as its own headword.

In this paper, we discuss how the data structures for MWEs can support this process by creating a closer fit between the continuous nature of language and its discrete representation in dictionaries.

Since the start of production use of Ekilex in late 2019, we have undertaken a series of data model migrations in the general direction of simplicity. A recurring pattern has been to reduce the number of specialised entities, previously considered indispensable due to ingrained habits from earlier dictionary writing systems. As it turns out, changes in lexicographic processes are systematically slower than anticipated. We have learned the hard way that initial resistance from users is not always a valid reason for preferring a traditional solution over a simpler one, because user preferences will eventually catch up.

Presentation of MWEs is a case in point. Initial design of Ekilex did dispense with specialised structures for most types of previously distinguished MWEs, presenting them as regular words with the type specified as a parameter instead. Two types retained their own entities, collocations (see for details Kallas et al., 2015) and usage examples, based on the idiosyncratic structure of their presentation.

The overwhelming complexity of this solution, combined with its theoretical fragility, has

now led us to reconsider the previously downvoted argument that these MWEs are difficult to distinguish from regular words. Just like words, they can have meanings, definitions, equivalents, registers, usage examples of their own, etc., and a particular MWE can simultaneously be a headword itself and be listed as a MWE in the article of some of its components.

As a result, we are moving towards a single universal structure for all language units. We'll present how this simplifies the data model and the lexicographic process, while also fitting the dictionary closer to language reality.

Acknowledgements

The creation and development of the portal was funded by the Digital Focus programme of the Ministry of Education and Research (2018–2021) and by the EKI-ASTRA programme (2016–2022). The creation of the dictionary and terminology database Ekilex was funded by the EKI-ASTRA programme (2016–2022). Software development has been provided by OÜ TripleDev. The research received funding from the European Union's Horizon 2020 research and innovation programme, under grant agreement No 731015.

Keywords: multiword expressions, dictionary writing system, data model

References

ELEXIS. Accessed at: https://elex.is (07 February 2021)

LEXIDMA: *Lexicographic Infrastructure Data Model and API.* Accessed at: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma (07 February 2021)

LMF ISO 24613:2008. Accessed at: https://www.iso.org/standard/37327.html (07 February 2021)

Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E,. Langemets, M., Michelfeit, J., Tuulik, M., Viks, Ü. (2015). Automatic generation of the Estonian Collocations Dictionary database. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 1—20.

Koppel, K., Tavast, A., Langemets, M., Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pereira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.) *Proceedings of the eLex 2019 conference. 1-3 October 2019, Sintra, Portugal.* Brno: Lexical Computing CZ, s.r.o., pp. 434—452.

Ontolex-lemon. Accessed at: https://www.w3.org/2016/05/ontolex/ (07 February 2021)

Tavast, A., Langemets, M., Kallas, J., Koppel, K. (2018). Unified Data Modelling for

Presenting Lexical Data: The Case of EKILEX. In Čibej, J., Gorjanc, V., Kosem, I., Krek, S. (eds.) *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana, 17-21 July 2018.*. Ljubljana University Press, Faculty of Arts, 749−761.

Tavast, A., Koppel, K., Langemets, M., Kallas, J. (2020). Towards the superdictionary: layers, tools and unidirectional meaning relations. In Gavriilidou, Z, Mitsiaki, M, Fliatouras, A. (eds.) *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I.* Alexandroupolis, Greece: Democritus University of Thrace, 215−223.

TEI Lex-0. Accessed at: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html# (07 February 2021)

What programmers want: avoiding recursion in dictionary schemas

Michal Měchura

Natural Language Processing Centre, Masaryk University
E-mail: valselob@gmail.com

One thing we often see in dictionary schemas is that they allow some form of recursive embedding, in other words, containing objects of one type inside objects of the same type. Typical examples are subsensing (a sense contains other, more specialized senses) and subentrying (for example when the entry for hole contains a subentry for black hole). Recursive embedding is a distracting complication for various digital agents (= software tools that process dictionary entries) such as dictionary writing systems or programs that extract data from dictionaries. From the perspective of someone who writes software tools for processing dictionary entries, it would be more convenient if each entry had only a flat list of senses and if no subentries existed. In this paper we will (1) argue that the needs and preferences of software agents and their writers should not be dismissed as secondary and should be taken seriously, and we will (2) propose an alternative, more IT friendly data-modelling pattern for dictionaries in which phenomena such as subsensing and subentrying are re-engineered as relations.

Subsensing

Broadly speaking, dictionary schemas allow subsensing either by recursion (instances of sense can contain other instances of sense) or by subtyping (instances of sense can contain instances of subsense). Both varieties of subsensing have the effect that the same kind of information (definitions, example sentences etc.) ends up being located at different depths inside the entry.

Subsensing is a device used by lexicographers for a valid purpose, to model sense relations or to ease the navigation of large entries. Our proposal is not abolish subsensing, but to change how we encode it. We propose to move away from hard-coding the sense-to-subsense hierarchy through embedding, and instead to encode it as sense relations. We keep the list of senses flat, each sense receives a unique ID, and we record the is-a-subsense-of relations in a stand-off fashion, as pairs of IDs. For presentation (= for displaying the entry to an end-user) we can use the stand-off relations to dynamically reconstruct a tree-structure with senses and subsenses. For all other processing, we have a flat list of senses with no embedding.

Subentry

Normally, a dictionary entry is headed by a headword, and then the rest of the entry describes that headword. But this orderly pattern can be broken by things which override the headword. A typical cause is the presence of a multiword subentry. Its presence somewhere in the body of an entry changes the object of description. While traversing the tree structure of an entry, the software agent reaches a point where, from that point onwards, we are no longer describing the headword we started with and we are describing something else instead. The challenge for a programmer is to keep track of these switches in what is being described.

Our proposal is to take those entry-internal elements that override the headword out of the entries and promote them to the status of entries. The fact that they should be shown inside other entries as subentries, will be encoded relationally, through unique IDs. For purposes other than presenting the entry to end users, we have a flat list of entries, and there are no subentries inside entries.

Conclusion

In our paper we will analyze subsensing and subentrying on real-world examples from several major born-digital dictionaries. We will show how hierarchical embedding can be re-engineered as ID-to-ID relations, without loss of expressivity. We will discuss what implications such a change would have for dictionary authoring (dictionary writing systems) and for dictionary viewing (stylesheets, navigation).

Also, we will compare our approach to earlier attempts by other authors to re-cast tree-structured dictionaries as networks, graphs or relational databases. The purpose of these other attempts has often been to re-invent what a dictionary is. The purpose of our re-modelling attempt, on the other hand, is not to invent new kinds of dictionaries. Our purpose is to produce a more IT-friendly dictionary encoding scheme for the kinds of dictionaries that are commonly authored today. IT professionals will find it easier (and themselves more willing) to work with entries in this format if they can count on the fact that complications of recursion and embedding are never going to arise.

Keywords: subsenses; subentries; dictionary encoding; recursion; embedding; nesting

A Multi-Word Thesaurus for 30+ Languages

Ondřej Herman[1], Miloš Jakubíček[2], Pavel Rychlý[1]

[1] Faculty of Informatics, Masaryk University

[2] Lexical Computing

E-mail: ondrej.herman@sketchengine.co.uk, milos.jakubicek@sketchengine.co.uk, pary@fi.muni.cz

This paper elaborates on a new development implement in Sketch Engine, a leading corpus management system (Kilgarriff et al., 2014), focusing on making a distributional thesaurus for multi-word units available. Since 2006 Sketch Engine features thesaurus for single-word units, calculated using the information obtained from Sketch Engine's word sketches (Rychlý, & Kilgarriff, 2007). In 2012, a extension to the word sketch concept has been introduced towards handling multi-word sketches (Kilgarriff et al., 2012). Since then, the single-word thesaurus basically waited to catch up the multi-word development, which is now presented.

Sketch Engine is a leading text corpus management system which as of 2019 includes several hundreds of preloaded corpora, monolingual as well as parallel ones, available to its users, who can also create their own corpora, have them annotated (part-of-speech tagged, lemmatized etc.) and contrast them against the preloaded ones. In 2010, Sketch Engine started the so-called TenTen series of web corpora (Jakubíček, 2013), aiming at building a corpus of ten billion words (1010, thus "TenTen") for as many languages as possible. Targeting ten billion words was not a random choice: by 2010 we had a corpus of that size for English and it clearly showed that it allows many of the Sketch Engine features that work well with a one billion word corpus and single-word units, to work well also on multi-word units. Also, given the Zipfian distribution observed in natural language, it was clear that making the corpora bigger is the only possible way that would allow us to research further on the issues of multiword expressions.

On top of the word sketches a distributional thesaurus has been part of Sketch Engine since 2006, facilitating an efficient algorithm which was tracktable on multi-billion word corpora (Rychlý & Kilgarriff, 2007). The thesaurus is using word sketches for computing the similarity score: it basically compares word sketch collocations for every pair of words in the corpus and the similarity relates to the fraction of shared collocates between these two words, taking the collocation weights as given by logDice into account. The new multi-word extension of the thesaurus uses the multi-word sketches as its backbone. The calculation starts by dumping the whole word sketch database and discovering multi-word sketches (i.e. two and more words connected with a word sketch relation) with a minimum frequency of 100 (less frequent items are unlikely to have any salient thesaurus items). These items form a new multiword thesaurus lexicon in addition to single-word items, and

are subject to the normal thesaurus calculation. We compare the multi-word thesaurus based on word sketches with a multi-word thesaurus based on word embeddings.

Keywords: Sketch Engine; word sketches; multi-word unit; thesaurus

References

Jakubíček, M. et al. (2013). The tenten corpus family. In: 7th International Corpus Linguistics Conference CL, pp. 125-127.

Kilgarriff, A. et al. (2014). The Sketch Engine: ten years on. Lexicography, 1.1, pp. 7-36.

Kilgarriff, A. et al. (2012). Finding multiwords of more than two words. Proceedings of EURALEX 2012.

Rychlý, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 41-44.

Measuring User Workload in e-Lexicography with the NASA Task Load Index

Geraint Paul Rees

Universitat Rovira i Virgili, Spain

E-mail: geraintrees@gmail.com

The user-friendliness of dictionaries and other lexicographic resources, in other words, the ease with which they can be used to look up information, has been an important preoccupation in lexicography research for at least the last 50 years. Several studies have examined the impact that innovations in access structure have had on the relative success of dictionary look-ups as well as the effort needed on the part of the user when performing these look-ups. A few of these studies, such as Lew et al. (2013) which used eye tracking to study users' selection of senses in bilingual dictionaries, have employed methods from the field of Human Computer Interaction (HCI) research. This study continues in this vein. The HCI construct of Task Load has been used to examine workload in a range of applications from nuclear power plants to websites. Here it is used to examine the workload involved in looking up information on English collocations in an academic English writing task. More specifically, the NASA Task Load Index (Hart & Staveland, 1988) (NASA-TLX) is used to compare 106 advanced L2 English students' perceptions of the workload involved in looking up collocation information using ColloCaid (Frankenberg-Garcia et al., 2019, 2020) - an integrated text editor and collocation dictionary – with their perceptions of the workload involved using a traditional word processor and other online lexicographic resource combination for the same task.

The results show that mean perceived overall workload was markedly lower when using the writing assistant to find collocation information than when using the other resource and word processor combination. The average (self-reported) time taken to complete the task was also slightly lower when using the writing assistant. The results also give some insight into the students' preferred online lexical resources.

The study concludes with a reflection on the suitability of the NASA-TLX for e-lexicography research and invites feedback and suggestions for potential applications of this and other HCI methods in e-lexicography research.

Keywords: NASA-TLX; workload; EAP; writing assistant; user study

References

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019).

Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), pp. 23–39. https://doi.org/10.1017/S0958344018000150

Frankenberg-Garcia, A., Rees, G. P., & Lew, R. (2020). Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*. https://doi.org/10.1093/ijl/ecaa022

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload*, *1*(3), pp. 139–183.

Lew, R., Grzelak, M., & Leszkowicz, M. (2013). How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. *Lexikos*, *23*(1). https://doi.org/10.5788/23-1-1213

Lexicographic APIs: the state of the art

Michal Měchura

Natural Language Processing Centre, Masaryk University, Czechia
E-mail: valselob@gmail.com

This paper presents a review of a recently emerged trend in e-lexicography: application programming interfaces (APIs) on the Internet which provide access to lexicographic content in machine-readable formats. Many dictionary publishers have recently started providing such APIs in the hope that external third-parties will use them to build innovative applications and that this will be an additional source of revenue for the publisher.

In retrospect, the recent turn towards APIs is a logical consequence of the disruption in which the dictionary publishing industry finds itself, with conventional paper-based publishing in decline while revenues from online publishing (websites, apps) are not fully off-setting the deficits. This leads many dictionary publishers to wanting to redefine themselves as application-agnostic "content providers", as licensors of high-quality language-reference content, which third parties would pay for to re-use and re-publish in their own applications and websites. To function as an application-agnostic content provider, a dictionary publisher needs to provide convenient machine-readable access to the content. An API is one way to provide such access.

In this paper we will look at examples of lexicographic APIs recently launched by several major dictionary publishers, including but not limited to: the Lexicala API, the Oxford Dictionaries API, the PONS API, the Macmillan Dictionary API, the Cambridge Dictionaries API and the Wordnik API. We will list their commonalities and differences, evaluating their strengths and weaknesses, and setting them in the wider context of RESTful APIs on the web today. The goal of the paper is to summarize the state of the art of this newly emerged trend in e-lexicography. As part of our presentation at eLex we will unveil an online curated directory of lexicographic APIs which we are compiling and which we are going to keep on updating in the future as the lexicographic API industry evolves.

Keywords: lexicograpic API; machine-readable dictionary; dictionary licensing; REST; JSON

Elexifier: a cloud-based dictionary conversion tool

Simon Krek[1], Andraž Repar[1], Carole Tiberius[2], Iztok Kosem[1], Janez Brank[1], Tina Munda[1]

[1] Jožef Stefan Institute, Slovenia
[2] Dutch Language Institute, Netherlands
E-mail: akgaertig@units.it

In this paper, we describe the tool Elexifier (elexifier.elex.is) which is a cloud-based dictionary conversion service for conversion of legacy XML and PDF dictionaries into a standard data format based on the Elexis Data Model (defined in the ELEXIS Horizon2020 project). It takes as input an XML or PDF dictionary and produces a TEI Lex-0 or OntoLex Lemon compliant XML file in line with the specifications described in the Elexis Data Model. To transform a custom XML dictionary, users need to use the Elexifier interface to define a JSON transformation, which specifies mapping rules for transforming custom XML elements into the Elexis Data Model core elements. The transformation specification offers a rich set of options for XML element selection allowing users to transform almost any XML format. To transform a PDF dictionary, users need to annotate a sample of the PDF file which is first transformed into a flat structure using a PDF to XML conversion script. A chunk of the resulting XML file is sent to Lexonomy for manual annotation and the annotations act as training data for the machine learning algorithm. Machine learning assumes a three-level structure with pages as first level base, entries as second level base and senses as third level base. A model is constructed for each level and trained on 75% of the data annotated in Lexonomy. Afterwards, labels for each token (separate word or symbol in the dictionary) of the unlabelled data are predicted for each level. The model used is a recurrent neural network with two inputs for each input token: one-hot encoded token features (such as font, size and so forth) and LSTM-encoded token contents. The two inputs are merged and fed into a bidirectional LSTM, which then outputs a one-hot encoded label. Current results show great promise as they often exceed 90% f1 score (varies between levels and datasets) and are achieved within a short training time. In the paper, we will describe the Elexifier application and its typical workflows, and demonstrate its usefulness on a variety of use cases (XML and PDF dictionary conversion, creation of linked lexical data, sense linking) within the framework of the Elexis Horizon2020 project.

Keywords: dictionary conversion tool; Text Encoding Initiative (Lex0); sense linking; data modelling

Abstracts of papers

A workflow for historical dictionary digitisation: Larramendi's Trilingual Dictionary

David Lindemann, Mikel Alonso
UPV/EHU University of the Basque Country, Vitoria-Gasteiz, Spain
E-mail: david.lindemann@ehu.eus, mikelalon@gmail.com

In this paper, we present a workflow for historical dictionary digitisation, with a 1745 Spanish-Basque-Latin dictionary as the use case. We start with scanned facsimile images, and get to represent attestations of modern standard Basque lexemes as Linked Data, in the form they appear in the dictionary. We are also able to produce an index of the dictionary, i.e. a Basque-Spanish version, and to map extracted Spanish and Basque lexical items to reference dictionary lemma list entries. The workflow is entirely based on freely available software. OCR and information extraction are performed using Machine Learning algorithms; data exhibits and the transcription curation environment are provided using Wikisource and Wikidata. Our evaluation of a first iteration of the workflow suggests its capability to deal with early modern printed dictionary text, and to reduce manual effort in the different stages significantly.

Keywords: Historical Lexicography; Digitisation; OCR; information extraction; Linked Data

GIPFA: Generating IPA Pronunciation from Audio

Xavier Marjou
Lannion, Brittany, France
E-mail: xavier.marjou@gmail.com

Transcribing spoken audio samples into the International Phonetic Alphabet (IPA) has long been reserved for experts. In this study, we examine the use of an Artificial Neural Network (ANN) model to automatically extract the IPA phonemic pronunciation of a word based on its audio pronunciation, hence its name Generating IPA Pronunciation From Audio (GIPFA). Based on the French Wikimedia dictionary, we trained our model which then correctly predicted 75% of the IPA pronunciations tested. Interestingly, by studying inference errors, the model made it possible to highlight possible errors in the dataset as well as to identify the closest phonemes in French.

Keywords: audio; transcription; phonemes; Artificial Neural Network; dataset

Visionary perspectives on the lexicographic treatment of easily confusable words: Paronyme – Dynamisch im Kontrast as the basis for bi- and multilingual reference guides

Petra Storjohann

Leibniz-Institut für Deutsche Sprache, R5-13, 68161 Mannheim, Germany
E-mail: storjohann@ids-mannheim.de

The German e-dictionary documenting confusables *Paronyme – Dynamisch im Kontrast* contains lexemes which are similar in sound, spelling and/or meaning, e.g. *autoritär/autoritativ*, *innovativ/innovatorisch*. These can cause uncertainty as to their appropriate use. The monolingual guide could be easily expanded to become a multilingual platform for commonly confused items by incorporating language modules. The value of this visionary resource is manifold. Firstly, e-dictionaries of confusables have not yet been compiled for most European languages; consequently, the German resource could serve as a model of practice. Secondly, it would be able to explain the usage of false friends. Thirdly, cognates and loan word equivalents would be offered for simultaneous consultation. Fourthly, users could find out whether, for example, a German pair is semantically equivalent to a pair in another language. Finally, it would inform users about cases where a pair of semantically similar words in one language has only one lexical counterpart in another language. This paper is an appeal for visionary projects and collaborative enterprises. I will outline the dictionary's layout and contents as shown by its contrastive entries. I will demonstrate potential additions, which would make it possible to build up a large platform for easily misused words in different languages.

Keywords: contrastive lexicography; bilingual paronyms; easily confused words; false friends; multilingual platform

Codification Within Reach: Three Clickable Layers of Information Surrounding the New Slovenian Normative Guide

Helena Dobrovoljc[1,2], Urška Vranjek Ošlak[1]

[1] ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Novi trg 2, SI-1000 Ljubljana, Slovenia;

[2] University of Nova Gorica, School of Humanities, Vipavska 13, SI-5000 Nova Gorica, Slovenia

E-mail: helena.dobrovoljc@zrc-sazu.si, urska.vranjek@zrc-sazu.si

This paper presents how language technology tools enable the integration of different types of normative data into a single language manual. The new Slovenian Normative Guide, the central normative manual consisting of normative rules and an orthographic dictionary, is based on language problems reported by language users. The normative guide consists of normative rules, and the orthographic dictionary supplements them with additional examples. The normative guide contains not only a systematic set of basic writing rules at the vowel-letter level (orthography or spelling), but also other consensual norms of the standard language. In order to effectively meet the needs of today's users of Slovenian, it was necessary to create a new concept for the orthographic dictionary so that it could effectively accompany the normative guide. In revising the normative rules, data collected on the Language Counselling Service platform were used. The normative guide is surrounded by three digitally interconnected layers of normative information; these three resources help the user navigate through the new normative view of the Slovenian language and provide arguments and explanations for the decisions made in the revision process.

Keywords: Slovenian; normative guide; orthographic dictionary; corpora research

From term extraction to lemma selection for an electronic LSP-dictionary in the field of mathematics

Theresa Kruse, Ulrich Heid

Institute for Information Science and Natural Language Processing (IwiSt), University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

E-mail: theresa.kruse@uni-hildesheim.de, ulrich.heid@uni-hildesheim.de

We work on term extraction for a corpus-based LSP-dictionary. Our field of study is the mathematical domain of graph theory. Our working hypothesis is that mathematics lends itself to a specific approach for term and information extraction with a lexicographical purpose. We compare different methods for term extraction: The first one combines pattern-based and statistical mean implemented by Schäfer et al. (2015), the second one has been developed especially for mathematical texts using domain-specific definition patterns based on work in the tradition of Meyer (2001). Further comparisons are made with a list of term candidates which are not part of the general language lexicon used in a version of TreeTagger trained on news text (Schmid, 1994) and with the term extraction provided by Sketch Engine (Kilgarriff et al., 2014). We use manual annotation by three expert raters and inter-rater agreement with -statistics to compare and evaluate the approaches. Additionally, we qualitatively analyse the extracted results. For selecting the lemmas, we work with a German corpus of lecture notes, textbooks and papers.

Keywords: LSP-dictionaries; mathematics; pattern-based extraction; automatic creation; semantic relation

Living Dictionaries: An Electronic Lexicography Tool for Community Activists

Gregory D. S. Anderson[1], Anna Luisa Daigneault[1]

[1]Living Tongues Institute for Endangered Languages, 4676 Commercial St SE, #454.
Salem, OR 97302, USA

E-mail: gdsa@livingtongues.org, annaluisa@livingtongues.org

Living Dictionaries are comprehensive, free online technological tools integrating audio, images and other multimedia that can assist endangered and other language communities, providing a simple way to create high-quality multilingual documentation records. The platform is a progressive web application functioning within any Internet browser on any computer or mobile device, Android or iOS. If needed, Living Dictionaries can be created, managed and edited using only smartphones or tablets, which can function as complete workstations for recording and entering linguistic data and other multimedia. Living Dictionaries may be public or private and may include written entries with translations and example sentences in multiple languages and scripts, audiovisual files, parts of speech and semantic domains, morphosyntactic linguistic analysis and be tagged with other metadata. The platform is free because for almost all minority language communities the costs related to producing high-quality linguistic materials can be insurmountable. A moral imperative of the 21st century is the decolonisation and democratisation of linguistic resources. Online dictionaries should reflect the user communities, tailored to suit their needs as well as curated by citizen-linguists. Community resources have greater uptake and engagement by communities if they take a primary role in developing them.

Keywords: dictionary; language technology; endangered languages; lexicography; web application

Mudra's Upper Sorbian-Czech dictionary – what can be done about this lexicographic "posthumous child"?

Michal Škrabal[1], Katja Brankačkec[2]

[1] Charles University, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic
[2] Institute of Slavonic Studies of the Czech Academy of Sciences,
Valentinská 1, 110 00 Praha 1, Czech Republic
E-mail: michal.skrabal@ff.cuni.cz, brankatschk@slu.cas.cz

Jiří Mudra, among his numerous selfless activities, was a Czech *doyen* of Sorbian studies. He had been working for decades on an Upper Sorbian-Czech dictionary but, unfortunately, had not finished his work on it at the time of his death. Presently, we are considering completing Mudra's project. The material collected by Mudra is undoubtedly valuable for us, providing us with a launchpad for further work; still, it is necessary to challenge it with the current data and a modern lexicographic approach. The paper presents the proposed individual methods aimed at finishing the main body of the dictionary.

Every lexicographer works with the data and tools available in his or her time – and Mudra was certainly no exception. There is, therefore, no reason to maintain exaggerated reverence towards his dataset where it is in apparent conflict with the current language reality. The aim is not to foster Mudra's cult, but to acknowledge his admirable initiative and enthusiasm. The best way to do so is to complete his dictionary with all the possibilities currently offered to us and make it available – as the first academic dictionary in this language combination – to Czech users.

Keywords: Upper Sorbian-Czech dictionary; completion; Jiří Mudra

Enriching a terminology for under-resourced languages using knowledge graphs

John P. McCrae[1], Atul Kr. Ojha[1], Bharathi Raja Chakravarthi[1], Ian Kelly[2], Patricia Buffini[2], Grace Tang[3], Eric Paquin[3], Manuel Locria[3]

[1]ADAPT Centre, Data Science Institute, NUI Galway, Ireland

[2] ADAPT Centre, Dublin City University, Ireland

[3] Translators without Borders

E-mail: john@mccr.ae, {atulkumar.ojha,bharathiraja.asokachakravarthi}@nuigalway.ie, ian.anthony.kelly@gmail.com, patricia.buffini@adaptcentre.ie, {grace,ericpaquin,manuel}@translatorswithoutborders.org

Translated terminology for severely under-resourced languages is a vital tool for aid workers working in humanitarian crises. However there are generally no lexical resources that can be used for this purpose. Translators without Borders (TWB) is a non-profit whose goal is to help get vital information, including developing lexical resources for aid workers. In order to help with the resource construction, TWB has worked with the ADAPT Centre to develop tools to help with the development of their resources for crisis response. In particular, we have enriched these resources by linking with open lexical resources such as WordNet and Wikidata as well as the derivation of a novel extended corpus. In particular, this work has focused on the development of resources for languages useful for aid workers working with Rohingya refugees, namely, Rohingya, Chittagonian, Bengali and Burmese. These languages are all under-resourced and for Rohingya and Chittagonian there are only very limited major lexical resources available. For these languages, we have constructed some of the first corpora resources that will allow automatic construction of lexical resources. We have also used the Naisc tool for monolingual dictionary linking in order to connect the existing English parts of the lexical resources with information from WordNet and Wikidata and this has provided a wealth of extra information including images, alternative definitions, translations (in Bengali, Burmese and other languages) as well as many related terms that may guide TWB linguists and terminologists in the process of extending their resources. We have presented these results in an interface allowing the lexicographers to browse through the results extracted from the external resources and select those that they wish to include in their resource. We present results on the quality of the linking inferred by the Naisc system as well as qualitative analysis of the effectiveness of the tool in the development of the TWB glossaries.

Keywords: under-resourced languages; terminology; linking; natural language processing; knowledge graphs

The ELEXIS System for Monolingual Sense Linking in Dictionaries

John P. McCrae[1], Sina Ahmadi[1], Seung-Bin Yim[2], Lenka Bajcetic[2]

[1] Data Science Institute, NUI Galway, Ireland

[2] Austrian Academy of Sciences

E-mail: john@mccr.ae, sina.ahmadi@insight-centre.org, seung-bin.yim@oeaw.ac.at, lenka.bajcetic@oeaw.ac.at

Sense linking is the task of inferring any potential relationships between senses stored in two dictionaries. This is a challenging task and in this paper we present our system that combines Natural Language Processing (NLP) and non-textual approaches to solve this task. We formalise linking as inferring links between pairs of senses as exact equivalents, partial equivalents (broader/narrower) or a looser relation or no relation between the two senses. This formulates the problem as a five-class classification for each pair of senses between the two dictionary entries. The work is limited to the case where the dictionaries are in the same language and thus we are only matching senses whose headword matches exactly; we call this task Monolingual Word Sense Alignment (MWSA). We have built tools for this task into an existing framework called Naisc and we describe the architecture of this system as part of the ELEXIS infrastructure, which covers all parts of the lexicographic process including dictionary drafting. Next, we look at methods of linking that rely on the text of the definitions to link, firstly looking at some basic methodologies and then implementing methods that use deep learning models such as BERT. We then look at methods that can exploit non-textual information about the senses in a meaningful way. Afterwards, we describe the challenge of inferring links holistically, taking into account that the links inferred by direct comparison of the definitions may lead to logical contradictions, e.g., multiple senses being equivalent to a single target sense. Finally, we document the creation of a test set for this MWSA task that covers 17 dictionary pairs in 15 languages and some results for our systems on this benchmark. The combination of these tools provides a highly flexible implementation that can link senses between a wide variety of input dictionaries and we demonstrate how linking can be done as part of the ELEXIS toolchain.

Keywords: sense linking; lexicography; natural language processing; linked data; tools

MOR*Digital*: The Advent of a New Lexicographic Portuguese Project

Rute Costa[1], Ana Salgado[2], Anas Fahad Khan[3], Sara Carvalho[1,4], Laurent Romary[5], Bruno Almeida[1,6], Margarida Ramos[1], Mohamed Khemakhem[7], Raquel Silva[1], Toma Tasovac[8]

[1] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal

[2] Academia das Ciências de Lisboa, Portugal

[3] Istituto Di Linguistica Computazionale 'A. Zampolli', Italy

[4] CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Portugal

[5] Inria, team ALMAnaCH, France

[6] ROSSIO Infrastructure, Portugal

[7] Arcascience, France

[8] BCDH – Belgrade Center for Digital Humanities
E-mail: rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, fahad.khan@ilc.cnr.it, sara.carvalho@ua.pt, laurent.romary@inria.fr, brunoalmeida@fcsh.unl.pt, mvramos@fcsh.unl.pt, medkhemakhemfsegs@gmail.com, raq.asilva@gmail.com, ttasovac@humanistika.org

MOR*Digital* is a newly funded Portuguese lexicographic project that aims to produce high-quality and searchable digital versions of the first three editions (1789; 1813; 1823) of the *Diccionario da Lingua Portugueza* by António de Morais Silva, preserving and making accessible this important work of European heritage. This paper will describe the current state of the art, the project, its objectives and the methodology proposed, the latter of which is based on a rigorous linguistic analysis and will also include steps necessary for the ontologisation of knowledge contained in and relating to the text. A section will be dedicated to the various investigation domains of the project description. The output of the project will be made available via a dedicated platform.

Keywords: digital humanities; GROBID-Dictionaries; legacy dictionary; lexicography; ontologies; standards

Catching lexemes.
The case of Estonian noun-based ambiforms

Geda Paulsen[1,2], Ene Vainik[1], Ahti Lohk[1], Maria Tuulik[1]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia

[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden

E-mail: {geda.paulsen, ene.vainik, ahti.lohk, maria.tuulik} @eki.ee

The aim of this study is to test a statistic relying on corpus data, the distributional index (D-index): a statistical benchmark that helps lexicographers judge if a morphological form has been conventionalised to the degree of becoming an independent lexeme. Our focus is on the decategorisation type that originates from a case form of a noun and is directed to an adverb, adposition or adjective. The words or inflected forms corresponding to more than one word class interpretation are in this study termed ambiforms. The analysis compares the D-index levels of ambiforms categorised as nouns and another PoS. The results suggest that for the outcome to be most authentic, the noun-based ambiforms should be analysed without the decategorisation influence, i.e. the D-index analysis should be applied in the pre-PoS-disambiguation stage.

Keywords: form distribution; morphology; lexicography; language technology; Estonian

Automatic Lexicographic Content Creation for Lexicographers

María José Domínguez Vázquez[1], Daniel Bardanca Outeiriño[2], Alberto Simões[3]

[1] Universidade de Santiago de Compostela – ILG, Santiago de Compostela, Spain

[2] Universidade de Santiago de Compostela – Santiago de Compostela, Spain

[3] 2Ai, School of Technology, IPCA, Barcelos, Portugal

E-mail: majo.dominguez@usc.es, daniel.bardanca@rai.usc.es, asimoes@ipca.pt

This paper presents *Combinatoria*, a tool for the semi-automatic generation of biargumental valency patterns for nominal phrases, as well as the current development of the tool for describing the passive valency of the noun. First, we describe a set of prototypes developed as exploratory tools for this new approach, together with the lexical and syntactic resources required for the generation of nominal phrases. We will focus especially on lexical resources, their automatic retrieval, and how they assist the lexicographic team in their tasks. This is followed by a description of the tool, the data filtering process, and the presentation of the obtained results. Finally, we include a brief discussion on the usefulness of these generators not only as stand-alone plurilingual dictionaries, but also as integrated resources in other electronic tools.

Keywords: multilingual valency dictionaries; argument patterns; automatic language generation; natural language processing

LeXmart: A platform designed with lexicographical data in mind

Alberto Simões[1], Ana Salgado[2,3], Rute Costa[3]

[1] 2Ai – School of Technology, IPCA, Barcelos, Portugal

[2] Academia das Ciências de Lisboa, Lisboa, Portugal

[3] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal

E-mail: asimoes@ipca.pt, anacastrosalgado@gmail.com, rute.costa@fcsh.unl.pt

LeXmart is an open-source web platform used to support the lexicographer's work through editing, control, validation, management, and publication of lexical resources. This tool was specifically developed to facilitate the compilation of general monolingual dictionaries in which data is encoded according to the Text Encoding Initiative (TEI) schema (chapter 9). Here, we will describe the challenges of adapting LeXmart to deal with TEI Lex-0 and distinct types of lexical resources, namely *Dicionário da Língua Portuguesa* (DLP) and *Vocabulário Ortográfico da Língua Portuguesa*, lexicographic works from Academia das Ciências Lisboa, and *Dicionário Aberto*, the retro-digitised version of the Cândido de Figueiredo dictionary. This article describes the steps taken to update the LeXmart platform to deal with the TEI Lex-0 schema and describe the challenges on properly encoding these three projects while allowing the lexicographical team to work continuously. This work builds on automatic operations performed on top of the original resources. It also includes the changes made to the editor to make it capable of dealing with the encoding updates and the new types of resources.

Keywords: dictionary editing system; e-lexicography; online dictionary; TEI Lex-0

Reshaping the Haphazard Folksonomy of the Semantic Domains of the French *Wiktionary*

Noé Gasparini[1], Cédric Tarbouriech[2], Sébastien Gathier[3], Antoine Bouchez[4]

[1,3,4] Institut international pour la Francophonie - Université Jean Moulin Lyon 3, 1C avenue des Frères Lumière CS 78242 - 69372 Lyon Cedex 08 France

[2] Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse & CNRS, France

E-mail: noe.gasparini@univ-lyon3.fr, cedric.tarbouriech@irit.fr, sebastien.gathier@univ-lyon3.fr, antoine.bouchez19@gmail.com

Semantic domains are a source of headaches in dictionary projects, and one was built haphazardly in the French edition of the collaborative online project *Wiktionary* called *Wiktionnaire*. *Wiktionnaire* is a lexicographical project that started 17 years ago. It is hosted by the Wikimedia Foundation and edited by a community of volunteers that made it a mature project, but with lacunas, with semantic domains being one of these. Between January 2019 and December 2020, this nomenclature of semantic domains was transformed by a small team with complementary expertise and skills. The team consisted of four people with academic knowledge in linguistics, lexicography and information science, as well as technical skills for coding, proofreading and community management. The strategy was the following: mapping the existing terminology, comparison and extension of the list, documentation, structuring, discussions with the community, deployment, cleaning of remaining irregularities, and monitoring the changes after this process. The result of this two-year operation is a complete reshaping of a messy folksonomy into an innovative lattice nomenclature fully integrated into the *Wiktionnaire* and adopted by the community, but also used in an RDF-based dictionary reusing that data, the *Dictionnaire des francophones*. This paper outlines the context of this work on continually changing content and presents the strategy used by the team, including the major issues and choices encountered during the process.

Keywords: semantic domains; *Wiktionnaire*; *Wiktionary*; folksonomy; collaborative lexicography

An Online Tool Developed for Post-Editing the New Skolt Sami Dictionary

Mika Hämäläinen[1], Khalid Alnajjar[1], Jack Rueter[1], Miika Lehtinen[2], Niko Partanen[1]

[1] University of Helsinki, Unioninkatu 40, 00100 Helsinki, Finland

[2] University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland

[3] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Lisboa, Portugal

E-mail: firstname.lastname@helsinki.fi, firstname.lastname@oulu.fi

In this paper, we present our free and open-source online dictionary editing system that has been developed for editing the new edition of the Finnish-Skolt Sami dictionary. We describe how the system can be used in post-editing a dictionary and how NLP methods have been incorporated as a part of the workflow. In practice, this means the use of FSTs (finite-state transducers) to enhance connections between lexemes and to generate inflection paradigms automatically. We also discuss our work in the wider context of lexicography of endangered languages. Our solutions are based on the open-source work conducted in the Giella infrastructure, which means that our system can be easily extended to other endangered languages as well. We have collaborated closely with Skolt Sami community lexicographers in order to build the system for their needs. As a result of this collaboration, the latest Finnish-Skolt Sami dictionary was edited and published using our system.

Keywords: Skolt Sami; online dictionary; NLP

Finding gaps in semantic descriptions. Visualisation of the cross-reference network in a Swedish monolingual dictionary

Kristian Blensenius[1], Emma Sköldberg[2], Erik Bäckerud[3]

[1] University of Gothenburg, Gothenburg (Sweden)
[2] University of Gothenburg, Gothenburg (Sweden)
[3] The Swedish Academy Dictionary, Lund (Sweden)
E-mail: kristian.blensenius@gu.se, emma.skoldberg@svenska.gu.se, erik.backerud@svenskaakademien.se

Providing lexical information in dictionary entries by cross-referencing between semantically related headwords is very important, both from a reception-oriented and a production-oriented perspective. This study presents a survey of cross-references in a comprehensive monolingual dictionary of Swedish. It discusses cross-referencing in dictionaries in general as well as in the Swedish dictionary, focusing on the following four types of paradigmatic cross-references: SEE, COMPARE, SYNONYM, and OPPOSITE. By using data-visualisation software, the semantic network in the dictionary is overviewed in a new way. Furthermore, errors, gaps as well as other areas of improvement in the dictionary related to cross-referencing are discovered. Moreover, the relationships between the existing cross-references, how they are introduced in the dictionary and the dictionary's intended target groups are addressed. The study also reveals that the traditional lexicographic policies of the dictionary need to be adjusted to take advantage of the transition from paper to electronic publication.

Keywords: cross-references; paradigmatic relations; Swedish; lexicography; semantics

The Latvian WordNet and Word Sense Disambiguation: Challenges and Findings

Ilze Lokmane[1], Laura Rituma[2], Madara Stāde[1], Agute Klints[1]

[1] University of Latvia, Department of Latvian and Baltic Studies, Visvalža 4a, Riga, LV-1050

[2] Institute of Mathematics and Computer Science, University of Latvia, Raina bulvaris 29, Riga, LV-1050

E-mail: ilze.lokmane@lu.lv, laura.rituma@lumii.lv, stade.madara@gmail.com, agute.klints@gmail.com

The article addresses the issues of word sense disambiguation within the process of developing an electronic lexical semantic resource, the Latvian WordNet. Apart from word senses, the resource also contains semantic paradigmatic relations between these senses, and therefore sense granularity must align with the need for creating synonymous, hyponymic, meronymic and antonymic links between Latvian words, as well as external links with the Princeton WordNet.

The development of the Latvian WordNet started in 2020 and it is based on two sources: a summarising electronic dictionary Tēzaurs.lv and available corpora. Because the word senses listed in Tēzaurs.lv are not directly usable for the needs of computer linguistics due to a number of reasons, the developers of the Latvian WordNet checked and revised the senses manually based on corpus data. Thus, the work on distinguishing word senses serves two purposes: 1) creating a Latvian WordNet, and 2) improving the structure of existing entries in the dictionary Tēzaurs.lv.

The article primarily focuses on the elaboration of common criteria for distinguishing word senses. The analysis concentrates on verbs as these are the most complex part of speech from the point of view of making sense distinctions. The authors conclude that the process is based on a set of criteria that form a certain hierarchy depending on the semantic group of verbs, namely, syntactic distribution, semantic distribution, as well as the interrelation between the two, and semantic decomposition of senses. Particular attention is paid to the interrelations of superordinate senses and subsenses, from which it is possible to conclude that an absolutely uniform and consistent subsense distinction is not likely to be possible, and, therefore, in cases of uncertainty, decisions are made in favour of what is needed to develop the Latvian WordNet.

Keywords: word sense disambiguation; sense distinction; electronic lexical semantic resource; syntactic and semantic distribution; lexical decomposition

Heteronym Sense Linking

Lenka Bajcetic[1], Thierry Declerck[1,2], John P. McCrae[3]

[1]Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse 19, Wien 1010, Austria
[2] German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Saarbrücken, Germany
[3] Data Science Insitute, NUI Galway, Ireland
E-mail: lenka.bajcetic@oeaw.ac.at, declerck@dfki.de, john@mccr.ae

In this paper we present ongoing work which aims to semi-automatically connect pronunciation information to lexical semantic resources which currently lack such information, with a focus on WordNet. This is particularly relevant for the cases of heteronyms — homographs that have different meanings associated with different pronunciations — as this is a factor that implies a re-design and adaptation of the formal representation of the targeted lexical semantic resources: in the case of heteronyms it is not enough to just add a slot for pronunciation information to each WordNet entry. Also, there are numerous tools and resources which rely on WordNet, so we hope that enriching WordNet with valuable pronunciation information can prove beneficial for many applications in the future. Our work consists of compiling a small gold standard dataset of heteronymous words, which contains short documents created for each WordNet sense, in total 136 senses matched with their pronunciation from Wiktionary. For the task of matching WordNet senses with their corresponding Wiktionary entries, we train several supervised classifiers which rely on various similarity metrics, and we explore whether these metrics can serve as useful features as well as the quality of the different classifiers tested on our dataset. Finally, we explain in what way these results could be stored in OntoLex-Lemon and integrated to the Open English WordNet.

Keywords: sense linking; heteronyms; Wordnets; Wiktionary

Dictionaries as collections of lexical data stories: an alternative post-editing model for historical corpus lexicography

Ligeia Lugli

SOAS, room 339, 10 Thornhaugh St, London WC1H 0XG, UK

E-mail: ll34@soas.ac.uk

This paper proposes a model of dictionary post-editing inspired by data-journalism. It starts by problematising the parallel, drawn in the description of this year's eLex conference theme, between lexicographic and machine-translation post-editing. It then proceeds to outline data- journalism workflows and to illustrate how these may offer a suitable blueprint for automating and post-editing corpus-driven historical dictionaries of low-resource languages. In particular, the paper highlights the usefulness of adopting an iterative development model, whereby minimal automated entries are incrementally augmented with curated information, and of switching to data-visualisations as the main medium of communication.

Data-journalists concentrate much of their post-editing efforts in plotting the data into highly customised visualisations capable of narrating their interpretation of a story while also allowing multiple lines of inquiry. This paper suggests that historical lexicographers would benefit from similarly directing their post-editing efforts into weaving data into customised, lemma-specific, visualisations capable of guiding users towards further exploration.

The paper concludes with practical examples drawn from two ongoing historical dictionary projects, *A Visual Dictionary and Thesaurus of Buddhist Sanskrit* and *A Visual Dictionary of Tibetan Verb Valency*, which are adopting data-journalism workflows to post-edit automatically generated entries and data-visualisations into 'lexical data stories'.

Keywords: historical lexicography; data-journalism; post-editing; Sanskrit; Tibetan

The structure of a dictionary entry and grammatical properties of multi-word units

Monika Czerepowicka

University of Warmia and Mazury in Olsztyn (Poland)
E-mail: monika.czerepowicka@uwm.edu.pl

Users of advanced inflectional languages expect dictionaries to provide clear inflectional information so that the creation or use of a given form does not generate additional problems. The development of technologies and tools for machine language processing has naturally made contemporary inflectional dictionaries advanced electronic works that contain tools for the individualisation of their content in line with users' needs. The main concern of this article is the influence of the grammatical properties of language units on lexicographic description, in particular the structure of a dictionary entry. This issue will be discussed with reference to *Verbel. The Inflectional Dictionary of Polish Verbal Phrases*, which is an electronic dictionary listing over 5,000 multi-word units, giving all their paradigmatic forms directly. Although it is a specialist study providing a formal description of units, thanks to the proper structure of entries it is possible to be used also by non-specialists. The opportunity of choosing the scope of lexicographic information in *The Verbel Dictionary* is guaranteed by a two-stage scheme of the entry which consists of a general and detailed description of units.

Keywords: multi-word units; inflection; dictionary; e-lexicography

Encoding semantic phenomena in verb-argument combinations

Elisabetta Jezek[1], Costanza Marini[1,2], Emma Romani[1]

[1] University of Pavia, Dipartimento di Studi Umanistici, Strada Nuova 65, 27100 Pavia

[2] University of Bergamo, Dipartimento di Lingue, Letterature e Culture Moderne, via Salvecchio 19, 24129 Bergamo

E-mail: e.jezek@unipv.it, costanza.marini01@universitadipavia.it, emma.romani01@universitadipavia.it

In this paper, we report the classification we adopted in two electronic resources of corpus-derived verbal patterns for Italian and Croatian (T-PAS and CROATPAS) to account for three different semantic phenomena that we observed occurring between nouns and verbs in valency structure contexts: Semantic Type alternation, Semantic Type shift (metonymy), and Complex Type exploitation. After presenting the two resources in the context of similar projects (Section 2), in Sections 3, 4, and 5 we examine the three phenomena in detail and show how we registered them in the editor we developed for this purpose, called Skema. The encoding of these phenomena in the editor is of paramount importance for being able to query them in the interface of the two resources, which will soon be publicly available online. In Section 5, we draw our conclusions and suggest possible ways to use the annotated data.

Keywords: pattern resource; verb argument structure; semantic type; corpus analysis; word sense

A cognitive perspective on the representation of MWEs in electronic learner's dictionaries

Thomai Dalpanagioti

School of English Language and Literature, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

E-mail: thomdalp@enl.auth.gr

One of the main pending methodological issues in lexicography is the representation of multiword expressions (MWEs). Their heterogeneous and fuzzy nature has given rise to diverse typologies in linguistic theory and to a variable and inconsistent treatment in lexicographic practice. Addressing this issue in the context of pedagogical lexicography is of vital importance because, due to a complex interplay of features of form, meaning and use, MWEs present major difficulties for learners as regards reception, production and retention. This paper thus examines the representation of different types of MWEs in online versions of English monolingual learner's dictionaries and points out the need for a more rational, motivated and systematic lexicographic treatment. We argue for a cognitively oriented approach to MWEs that draws on Frame Semantics and the Conceptual Metaphor and Metonymy Theory. The proposal is illustrated through two case studies, which demonstrate how MWEs are integrated in a motivated semantic network of the motion verbs *crawl* and *dash*. The flexibility of the electronic medium can make it feasible to design cognitively informed features of the dictionary microstructure to improve the representation of MWEs.

Keywords: multiword expressions; monolingual learner's dictionaries; Frame Semantics; Conceptual Metaphor and Metonymy Theory; motion verbs

Frame-based terminography: a multi-modal knowledge base for karstology

Špela Vintar[1], Vid Podpečan[2], Vid Ribič[3]

[1] University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana

[2] Jožef Stefan Institute, Jamova 39, SI – 1000 Ljubljana

[3] Kofein dizajn, Beethovnova 9, SI – 1000 Ljubljana

E-mail: spela.vintar@ff.uni-lj.si, vid.podpecan@ijs.si, vid@kofein.si

We present an innovative approach to the representation of domain-specific knowledge which combines traditional concept-oriented terminography with knowledge frames and augments linguistic data with images, videos, interactive graphs and maps. The interface is simple and intuitive, prompting the user to enter a query term in any of the three languages (English, Croatian and Slovene). If the term is found it is described through textual definitions from various sources, its frame derived from annotated data, a graph depicting the neighbourhood of the concept and – if feasible – a map of geolocations for the queried term. The frame represents aggregated and structured knowledge as it describes the concept through a set of semantic relations. Graphs enable the user to browse through related concepts and explore the domain in a visually represented network. The underlying knowledge base of karstology was created within the TermFrame project and is based on an implementation and extension of the frame-based approach to terminology.

Keywords: frame-based terminography; karstology; knowledge base; visualisation

Creating an Electronic Lexicon for the Under-resourced Southern Varieties of the Kurdish Language

Zahra Azin[1], Sina Ahmadi[2]

[1] Geomatics and Cartographic Research Center, Carleton University

[2] National University of Ireland Galway, Ireland

E-mail: zahraazin@cmail.carleton.ca, ahmadi.sina@outlook.com

Thanks to the advances in information technology and communication, many endangered, vulnerable and under-represented language communities have a chance to revitalise and document their languages. In comparison to other Kurdish variants such as central Kurdish (also known as Sorani) and northern Kurdish (also known as Kurmanji), southern Kurdish has received little attention, making it an under-documented and under-resourced language that is spoken primarily in the Kurdish regions of Iran, particularly Kermanshah and Ilam provinces. As the case of our study, we focus on creating an electronic monolingual lexicon of significant size for the southern variants of Kurdish in the OntoLex-Lemon ontology by converting a bilingual and monolingual dictionary. In addition, we report our efforts in using a semi-automatic pivot-based translation inference approach to align the current resource with other resources in Kurdish and Gorani. We believe that this resource increases inter-operability across various natural language processing systems and facilitates many tasks in computational linguistics for Kurdish. Our resource is publicly available under a Creative Commons Attribution-ShareAlike 4.0 International License (https://github.com/sinaahmadi/SKurdishLexicon).

Keywords: southern Kurdish; electronic lexicography; less-resourced languages; machine-readable dictionary

Lemmatisation, etymology and information overload on English and Swedish editions of Wiktionary

Allahverdi Verdizade
Uppsala University, P.O. Box 256, SE-751 05 Uppsala
E-mail: allahverdi.verdizade@lingfil.uu.se

Wiktionary is a user-generated wiki-project with the goal of building a universal dictionary covering all words in all languages. Various language editions of Wiktionary have community-specific policies regulating concrete lexicographic questions. The distinct entry structures of English and Swedish Wiktionaries are examined in the context of the relation between headword and etymological information, under special consideration of the user-friendliness of the respective approach. The English Wiktionary applies the etymological approach in setting the headword, which splits identical forms into parts of speech, but also into headwords based on word origin. Additionally, the semantic information is separated from non-semantic more rigorously than is done in the Swedish Wiktionary, placing lists of related and derived terms below the headword rather than under each definition. The Swedish Wiktionary applies the formal-grammatical approach, where division into headwords is made strictly based on identical form and part of speech. In this approach, homonymy is disregarded. The etymological information is nested under each definition rather than having a separate section above the headword. The analysis of the two language editions suggests that the different approaches lead to different amounts of information overload in users, depending on the extent of non-semantic information. Equally extensive entries are handled better within the layout structure of the English Wiktionary.

Keywords: Wiktionary; information overload; etymology

Multiword-term bracketing and representation in terminological knowledge bases

Pilar León-Araúz, Melania Cabezas-García, Pamela Faber

University of Granada, Granada, Spain

E-mail: pleon@ugr.es, melaniacabezas@ugr.es, pfaber@ugr.es

Multiword terms (MWTs) are frequently consulted in terminological resources due to their structural, cognitive, and conceptual complexity. However, in most terminological resources they are not always well described, since they are often included as independent term entries with no information on how their constituents are related. An accurate management of MWTs of three or more constituents requires, as a first step, their structural disambiguation, also called bracketing. This paper examines MWT bracketing in order to enhance MWT representation by describing their structural dependencies. Based on NLP advances in bracketing, a protocol has been designed through corpus queries and evaluated according to the reliability of corpora and rules as well as the causes underlying failure. Automatising bracketing can help enhance the representation of MWTs in terminological knowledge bases, assisting both the terminologist and the final user, since making their relational structure explicit can favour knowledge acquisition.

Keywords: multiword term; bracketing; terminological knowledge base; terminology

New developments in Lexonomy

Adam Rambousek[1,2,4], Miloš Jakubícek[1,2], Iztok Kosem[3,4]

[1] Faculty of Informatics, Masaryk University, Brno, Czech Republic

[2] Lexical Computing, Brno, Czech Republic

[3] Centre for Language Resources and Technologies, University of Ljubljana, Slovenia

[4] Jožef Stefan Institute, Ljubljana, Slovenia

E-mail: rambousek@fi.muni.cz, milos.jakubicek@sketchengine.eu, iztok.kosem@cjvt.si

This article describes new developments and enhanced features in the open-source web application for dictionary writing, Lexonomy. Since its introduction in 2017, a growing number of users and organisations have chosen Lexonomy to edit their dictionaries. We describe the motivation and process of the source code refactoring to Python programming language. Next, we provide details on integration with the Sketch Engine corpus manager. We also cover the completely new feature of dictionary linking, both as a graphical interface for users, and API to include Lexonomy in the process of automatic dictionary linking. Finally, the article describes the new functionality needed for Lexonomy integration within the ELEXIS project processes. Furthermore, we provide usage statistics on users and dictionaries they create.

Keywords: Dictionary editing; Dictionary writing system; Lexicographic tools; XML; Corpora connection

The Distribution Index Calculator for Estonian

Ene Vainik[1], Ahti Lohk[1], Geda Paulsen[1,2]

[1] Institute of the Estonian Language, Roosikrantsi 6, Tallinn 10119, Estonia
[2] Uppsala University, Thunbergsvägen 3 L, Uppsala 75126, Sweden
E-mail: Ene.Vainik@eki.ee, Ahti.Lohk@eki.ee, Geda.Paulsen@eki.ee

Lexicographers working with such morphologically rich languages as Estonian face the task of detecting the lexicographic status of some word forms that look like case forms of nouns but can behave as function words to a certain degree. Hence, a measurable criterion for making a word form an autonomous headword is needed. The present paper describes the idea and development of a tool called the Distribution Index Calculator (DIC) for Estonian. It is a web-based application which finds the frequency data of word forms and lemmas from an annotated corpus and retrieves a statistic called the Distribution Index (DI). The DI indicates the relative prominence of a word form as compared to its expected normative level of salience. The application is described in detail and some illustrations of its performance are provided. The evaluation of its quality is as follows: a higher than critical level of DI can be trusted as an indicator of the relative autonomy of a word form, while a lower than critical level of DI does not preclude such autonomy. The DIC thus gives relative heuristics rather than absolute ratings or true-value decisions.

Keywords: language technology; lexicography; morphology; distribution of case forms; the Estonian language

Compiling an Estonian-Slovak Dictionary with English as a Binder

Michaela Denisová[1]

[1] Masaryk University, Žerotínovo nám. 617/9, 601 77 Brno
E-mail: michaeladenisova@gmail.com

For such a rare language combination as Estonian-Slovak, it is complicated to find study materials designated for Slovaks learning Estonian, especially a bilingual dictionary, an essential language study resource. However, building a bilingual dictionary from scratch requires a lot of work and effort. The half-automatic computational methods and available open-source language resources offer a possible solution for this complicated task. One approach is to merge two already existing dictionaries that share a common language to derive a new language pair dictionary. However, as words are polysemous, many mistakes could occur while attempting so. Therefore, it is required to edit the aligned translations afterwards.

This article describes the process of compiling the Estonian-Slovak dictionary created from English-Estonian and English-Slovak dictionaries. English was chosen as an intermediate language, as it is a well-resourced language, and all materials are easy to find. Various automatic techniques were applied in the editing step to decrease the number of incorrectly aligned translations. Finally, the techniques used and quality of the dictionary were manually evaluated on a random sample of 1,000 translations.

The final version of the dictionary consists of 138,779 translations, and the Estonian headword list covers about 85% of basic Estonian vocabulary, which contains around 5,000 lemmas. The correct translations form approximately 40% of the dictionary. Additionally, a web application is being developed for this dictionary.[7]

Keywords: bilingual dictionaries; (semi)automatic compilation; intermediate language; Estonian; Slovak

[7] https://estonian-slovak-dictionary.herokuapp.com (23 March 2021).

Using Open-Source Tools to Digitise Lexical Resources for Low-Resource Languages

Ben Bongalon[1], Joel Ilao[2], Ethel Ong[3], Rochelle Irene Lucas[4], Melvin Jabar[5]

[1] Independent Researcher, California, USA

[2,3] College of Computer Studies, De La Salle University, Manila, Philippines

[4] English and Applied Linguistics, De La Salle University

[5] Behavioral Sciences, De La Salle University

E-mail: ben@isawika.org, {joel.ilao, ethel.ong, rochelle.lucas, melvin.jabar}@dlsu.edu.ph

Advances in open-source lexicography tools have made it more practical to digitise historical dictionaries and lexical resources. However, most retro-digitisation efforts have catered to dominant languages while ethnic minority and indigenous languages tend to be neglected. In countries with a large number of regional and local languages, such as the Philippines, retro-digitisation is a daunting challenge. Of its 186 languages and 500+ dialects, only a few are known to have e-dictionaries produced. The traditional "top-down" approach simply does not scale, since the community need for language documentation far outstrips the number of motivated linguists, lexicographers and funding entities available. This paper describes a complete tool chain and workflow that we used to digitise a Hanunoo-English dictionary originally published in the 1950s (Conklin, 1953). A trainable OCR engine, Tesseract (Smith, 2007), is used to handle the novel glyphs found in the dictionary. Post-edits were performed to fix OCR errors, extract lexical elements from the transcribed pages, and produce an XML-formatted electronic dictionary containing 5,779 entries. The Lexonomy dictionary editor (Měchura, 2017) was used to edit the entries and host the access-controlled electronic dictionary online.

Keywords: indigenous language; retro-digitisation; electronic lexicography; OCR; LSTM

A Word Embedding Approach to Onomasiological Search in Multilingual Loanword Lexicography

Peter Meyer[1], Ngoc Duyen Tanja Tu[1]

[1] Leibniz-Institut für Deutsche Sprache, R5, 6-13 68161 Mannheim Germany
E-mail: meyer@ids-mannheim.de, tu@ids-mannheim.de

Abstract

In this paper we present an experimental semantic search function, based on word embeddings, for an integrated online information system on German lexical borrowings into other languages, the *Lehnwortportal Deutsch* (LWPD). The LWPD synthesizes an increasing number of lexicographical resources and provides basic cross-resource search options. Onomasiological access to the lexical units of the portal is a highly desirable feature for many research questions, such as the likelihood of borrowing lexical units with a given meaning (Haspelmath & Tadmor, 2009; Zeller, 2015). The search technology is based on multilingual pre-trained word embeddings, and individual word senses in the portal are associated with word vectors. Users may select one or more among a very large number of search terms, and the database returns lexical items with word sense vectors similar to these terms. We give a preliminary assessment of the feasibility, usability and efficacy of our approach, in particular in comparison to search options based on semantic domains or fields.

Keywords: onomasiological search; word embeddings; multilingual lexicography; lexical borrowings

Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources

Carole Tiberius[1], Simon Krek[2], Katrien Depuydt[1], Polona Gantar[3], Jelena Kallas[4], Iztok Kosem[2], Michael Rundell[5]

[1] Instituut voor de Nederlandse Taal, Leiden, The Netherlands
[2] Jožef Stefan Institute, Ljubljana, Slovenia
[3]Faculty of Arts, University of Ljubljana, Slovenia
[4]Institute of the Estonian Language, Tallinn, Estonia
[5]Lexical Computing, Brno, Czech Republic
E-mail: {carole.tiberius,katrien.depuydt}@ivdnt.org, {simon.krek,iztok.kosem}@ijs.si, apolonija.gantar@ff.uni-lj.si, jelena.kallas@eki.ee, michael.rundell@gmail.com

In this paper we describe ongoing work on the identification and definition of core lexicographic elements to be used in the ELEXIS data model. ELEXIS is a European infrastructure project fostering cooperation and information exchange among lexicographical research communities. One of the main goals of ELEXIS is to make existing lexicographic resources available on a significantly higher level than is currently the case. Therefore, a common data model is being developed which aims to: a) streamline the integration of lexicographic data into the infrastructure (using the ELEXIFIER tool), b) enable reliable linking of the data in the ELEXIS Dictionary Matrix, and c) provide a basic template for the creation of new lexicographic resources, such that they can automatically benefit from the tools and services provided by the ELEXIS infrastructure. Here we focus on the development of a common vocabulary and report on the results of an initial survey that was conducted to collect feedback from experts in lexicography.

Keywords: data model; common vocabulary; lexicographic resource; interoperability

Visualising Lexical Data for a Corpus-Driven Encyclopaedia

Santiago Chambó[1], Pilar León-Araúz[2]

[1, 2] Department of Translation and Interpreting,
University of Granada, Buensuceso, 11, 18002 Granada (Spain)
E-mail: santiagochambo@ugr.es, pleon@ugr.es

The Humanitarian Encyclopedia (HE) is an ongoing corpus-driven project that aims at defining and documenting the dynamics of 129 concepts that are particularly controversial, fuzzy or ill-defined within the humanitarian action domain, thus enhancing communication in a sensitive area. In the HE, each entry is created according to an approach that combines corpus-driven knowledge with expert knowledge. Concept entries are authored by field experts who are provided with a Linguistic Analysis Report (LAR) created by a team of linguists. In LARs, HE linguists support their claims by i) presenting, quantifying and categorising textual data and by ii) making comparisons among subcorpora, which are created based on the corpus metadata (i.e. document type, region, organisation type, publication year). This article presents the visualisations created by HE linguists to represent both semantic information (i.e., conceptual combinations and non-hierarchically related concepts) and quantifiable concordance and collocational data. This includes approaches to disaggregating measures according to different kinds of subcorpus types and strategies to represent collocational intersections among subcorpora (i.e., collocates occurring in multiple subcorpora) as well as collocates unique to each subcorpus. Other concept-specific visualisations were also designed and are examined in this article.

Keywords: lexical data; visualisation; concept

Language Monitor: tracking the use of words in contemporary Slovene

Iztok Kosem[1,2], Simon Krek[1,2], Polona Gantar[1], Špela Arhar Holdt[1], Jaka Čibej[1]

[1] Centre for Language Resources and Technologies (CJVT), University of Ljubljana
[2] Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@cjvt.si, simon.krek@ijs.si, apolonija.gantar@guest.arnes.si, spela.arhar@cjvt.si, jaka.cibej@ff.uni-lj.si

In this paper, we present Language Monitor 1.0, a new online resource for monitoring language changes in Slovene, developed at the Centre for Language Resources and Technologies at the University of Ljubljana. The resource is another part of the newly developed infrastructure for researching and describing contemporary Slovene. Language Monitor 1.0 offers four sections to observe word usage: (1) a single-word list, (2) word groups, (2) a neologism section, and (4) word comparisons. The words for a single-word list are manually validated from a list of salient word candidates, which are identified using the Simple Maths method. The paper also describes future plans, including the setup of a relational database linked with a data warehouse solution for analysis purposes, which will include various statistical information on different language phenomena relevant for researchers, lexicographers, and other users, and will provide possibilities for adding several new features to the Language Monitor.

Keywords: Language Monitor; trends; neologisms; language change; corpus

Automatic induction of a multilingual taxonomy of discourse markers

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
E-mail: rogelio.nazar@pucv.cl

This paper describes a proposed method for the identification and classification of discourse markers (e.g., *however, therefore, by the way*) by applying statistical analysis to large parallel corpora. The objective is to build a lexical resource consisting of a multilingual taxonomy, so far in English, Spanish, German and French. A method is proposed that first separates discourse markers from the rest of the lexical units in the corpus using a measure of entropy, and then classifies them in groups by function using a clustering procedure especially designed for massive data processing. From that point onwards, the system is used to recursively identify and classify more units. Experimental evaluation shows that, in terms of precision, the automated method is able to perform as well as a team of human annotators (undergraduate students of linguistics), and it outperforms them in terms of recall.

Keywords: automatic creation of dictionary content; connectives; discourse markers; taxonomy induction; natural language processing

Porting the Latin WordNet onto OntoLex-Lemon

Stefania Racioppa[1], Thierry Declerck[1,2]

[1] German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2, Saarbrücken, Germany
[2] Austrian Centre for Digital Humanities and Cultural Heritage
Sonnenfelsgasse 19, Wien 1010, Austria
E-mail: stefania.racioppa@dfki.de, declerck@dfki.de

In this paper we describe the porting of the Latin WordNet data available at the University of Exeter onto the OntoLex-Lemon model, focusing on the representation of both morphological and conceptual information. In the longer term, we aim at integrating the resulting data set in the Linguistic Linked Open Data (LLOD) infrastructure, linking (or even merging) it to the Latin data sets already published in the LOD framework by the ERC "Linking Latin" (LILA) project. We discuss some lessons learned, as it turned out that such a transformation and linking exercise can lead to an improved consistency and accuracy of the original data.

Keywords: Latin; WordNet; Morphology; OntoLex-Lemon

Identifying Metadata-Specific Collocations in Text Corpora

Ondrej Herman[1,2], Miloš Jakubícek[1,2], Vojtech Kovár[1,2]

[1] Natural Language Processing Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic

[2] Lexical Computing, Brno, Czech Republic

E-mail: ondrej.herman@sketchengine.eu, milos.jakubicek@sketchengine.eu, vojtech.kovar@sketchengine.eu

Statistical corpus analysis of collocations is one of the important steps in creating a dictionary entry: collocations may distinguish senses, describe typical phrasemes and idioms and outline the whole picture of a word's behaviour. However, some collocations are domain-specific, typical only in particular contexts, and thus far there has been no easy way to distinguish "general" collocations from those that are predominantly typical in particular domains. In this paper, we present a tool which allows lexicographers to see typical domains in which a particular collocation occurs. We introduce a statistical procedure based on corpus metadata to identify domain-specific collocations in an intuitive way, and we also present a user interface connected to the word sketch feature of the Sketch Engine corpus interface (Kilgarriff et al., 2014a). The new feature can be used in the manual inspection of collocation lists, as well as when using the API or in a semi-automatic post-editing scenario of building a dictionary.

Keywords: collocations; word sketch; meta-data; text types; corpus

Corpus-based Methodology for an Online Multilingual Collocations Dictionary: First Steps

Adriane Orenha-Ottaiano[1], Marcos Garcia [2], Maria Eugênia Olímpio de Oliveira Silva[3], Marie-Claude L'Homme[4], Margarita Alonso Ramos[5], Carlos Roberto Valêncio[6], William Tenório[7]

[1] São Paulo State University (UNESP), Brazil
2 Universidade de Santiago de Compostela, Galiza, Spain
3 University of Alcalá, Spain
4 OLST, Université de Montréal, Québec, Canada
[5] Universidade da Coruña, Spain
[6] São Paulo State University (UNESP), Brazil
[7] São Paulo State University (UNESP), Brazil
E-mail: adriane.ottaiano@unesp.br, marcos.garcia.gonzalez@usc.gal, eugenia.olimpio@uah.es, mc.lhomme@umontreal.ca.ca, margarita.alonso@udc.es, carlos.valencio@unesp.br, williamtenoriotenorio@gmail.com

This paper describes the first steps of a corpus-based methodology for the development of an online Platform for Multilingual Collocations Dictionaries (PLATCOL). The platform is aimed to be customized for different target audiences according to their needs. It covers various syntactic structures of collocations that fit into the following taxonomy: verbal, adjectival, nominal, and adverbial. Part of its design, layout and methodological procedures are based on the *Bilingual Online Collocations Dictionary Platform* (Orenha-Ottaiano, 2017). The methodology also relies on the combination of automatic methods to extract candidate collocations (Garcia et al., 2019a) with careful post-editing performed by lexicographers. The automatic approaches take advantage of NLP tools to annotate large corpora with lemmas, PoS-tags and dependency relations in five languages (English, French, Portuguese, Spanish and Chinese). Using these data, we apply statistical measures (Evert et al., 2017; Garcia et al., 2019b) and distributional semantics strategies to select the candidates (Garcia et al., 2019c) and retrieve corpus-based examples (Kilgarriff et al., 2008). We also rely on automatic definition extraction (Bond & Foster, 2013) so that collocations can be more effectively organized according to their specific senses.

Keywords: collocations; collocations dictionary; online platform; automatic extraction; lexicography

Word-embedding based bilingual terminology alignment

Andraž Repar[1], Matej Martinc[2], Matej Ulčar[3], Senja Pollak[2]

[1] International Postgraduate School, Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia

[2] Institut Jozef Stefan, Jamova 39, Ljubljana, Slovenia

[3] Faculty of Computer and Information Science, University of Ljubljana, Vecna pot 113, Ljubljana, Slovenia

E-mail: andraz.repar@ijs.si, matej.martinc@ijs.si, matej.ulcar@fri.uni-lj.si, senja.pollak@ijs.si

The ability to accurately align concepts between languages can provide significant benefits in many practical applications. In this paper, we extend a machine learning approach using dictionary and cognate-based features with novel cross-lingual embedding features using pretrained fastText embeddings. We use the tool VecMap to align the embeddings between Slovenian and English and then for every word calculate the top 3 closest word embeddings in the opposite language based on cosine distance. These alignments are then used as features for the machine learning algorithm. With one configuration of the input parameters, we managed to improve the overall F-score compared to previous work, while another configuration yielded improved precision (96%) at a cost of lower recall. Using embedding-based features as a replacement for dictionary-based features provides a significant benefit: while a large bilingual parallel corpus is required to generate the Giza++ word alignment lists, no such data is required for embedding-based features where the only required inputs are two unrelated monolingual corpora and a small bilingual dictionary from which the embedding alignments are calculated.

Keywords: terminology alignment; word embeddings; embeddings alignment; machine learning

Semi-automatic building of large-scale digital dictionaries

Marek Blahuš[1], Michal Cukr[1], Ondrej Herman[1,2], Miloš Jakubíček[1,2], Vojtech Kovár[1,2], Marek Medved[1,2]

[1] Lexical Computing, Brno, Czechia

[2] Natural Language Processing Centre
Faculty of Informatics, Masaryk University

E-mail: {firstname.lastname}@sketchengine.eu

This paper presents a novel way of creating dictionaries by using a particular post-editing workflow, all of which is carried out in the context of building a set of three bilingual dictionaries – Tagalog, Urdu and Lao dictionaries with translations into English and Korean. The dictionaries were created completely from scratch without reusing any existing content and in a completely automatic manner, amounting to 50,000 headwords, out of which 15,000 headwords were subject to subsequent manual post-editing. In the paper we discuss the post-editing methodology that we used and its impact on the overall lexicographic workflow. We describe the web corpora that were built specifically for the purpose of building these three dictionaries as well as their annotations (such as PoS tagging and lemmatisation) and tools that were used for the corpus annotation and for automating individual entry parts and the post-editing thereof. Most of the automatic drafting and post-editing relied on a backbone consisting of the Sketch Engine corpus management system and Lexonomy dictionary editor We also detail the overall amount of work involved in each post-editing step, the technical and managerial difficulties faced alongside in the project, and the major technological issues that still need improvement in the post-editing scenario.

Keywords: post-editing lexicography; dictionary drafting; Sketch Engine

Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages

Federico Martelli[1], Roberto Navigli[1], Simon Krek[2], Jelena Kallas[3], Polona Gantar[4], Svetla Koeva[5], Sanni Nimb[10], Bolette Sandford Pedersen[8], Sussi Olsen[8], Margit Langemets[3], Kristina Koppel[3], Tiiu Üksik[3], Kaja Dobrovoljc[2], Rafael-J. Ureña-Ruiz[9], José-Luis Sancho-Sánchez[9], Veronika Lipp[11], Tamás Váradi[12], András Gyorffy[11], Simon László[11], Valeria Quochi[14], Monica Monachini[14], Francesca Frontini[14], Carole Tiberius[13], Rob Tempelaars[13], Rute Costa[6], Ana Salgado[6,7], Jaka Čibej[2], Tina Munda[2]

[1] Sapienza NLP Group, Department of Computer Science, Sapienza University of Rome, Italy
[2] Artificial Intelligence Laboratory, Jožef Stefan Institute, Slovenia
[3] Institute of the Estonian Language, Estonia
[4] Faculty of Arts, University of Ljubljana, Slovenia
[5] Institute for Bulgarian Language, Bulgarian Academy of Sciences, Bulgaria
[6] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
[7] Academia das Ciências de Lisboa, Portugal
[8] University of Copenhagen, Denmark
[9] Centro de Estudios de la Real Academia Española, Spain
[10] Society for Danish Language and Literature, Copenhagen, Denmark
[11] Hungarian Research Centre for Linguistics, Institute for Lexicology, Hungary
[12] Hungarian Research Centre for Linguistics, Institute for Language Technologies and Applied Linguistics, Hungary
[13] Instituut voor de Nederlandse Taal, The Netherlands
[14] Istituto di Linguistica Computazionale "A. Zampolli", Centro Nazionale delle Ricerche, Italy
E-mail: federico.martelli@uniroma1.it, roberto.navigli@uniroma1.it, simon.krek@ijs.si, jelena.kallas@eki.ee, apolonija.gantar@ff.uni-lj.si, svetla@dcl.bas.bg, kaja.dobrovoljc@ijs.si, lipp.veronika@nytud.hu, varadi.tamas@nytud.hu, simon.laszlo@nytud.hu, gyorffy.andras@nytud.hu, valeria.quochi@ilc.cnr.it, monica.monachini@ilc.cnr.it, francesca.frontini@ilc.cnr.it, jaka.cibej@ijs.si, tina.munda@ijs.si, bspedersen@hum.ku.dk, saolsen@hum.ku.dk, margit.langemets@eki.ee, kristina.koppel@eki.ee, tiiu.yksik@eki.ee, rute.costa@fcsh.unl.pt, anasalgado@campus.fcsh.unl.pt, carole.tiberius@ivdnt.org, rob.tempelaars@ivdnt.org

Over the course of the last few years, lexicography has witnessed the burgeoning of increasingly reliable automatic approaches supporting the creation of lexicographic

resources such as dictionaries, lexical knowledge bases and annotated datasets. In fact, recent achievements in the field of Natural Language Processing and particularly in Word Sense Disambiguation have widely demonstrated their effectiveness not only for the creation of lexicographic resources, but also for enabling a deeper analysis of lexical-semantic data both within and across languages. Nevertheless, we argue that the potential derived from the connections between the two fields is far from exhausted. In this work, we address a serious limitation affecting both lexicography and Word Sense Disambiguation, i.e. the lack of high-quality sense-annotated data and describe our efforts aimed at constructing a novel entirely manually annotated parallel dataset in 10 European languages. For the purposes of the present paper, we concentrate on the annotation of morpho-syntactic features. Finally, unlike many of the currently available sense-annotated datasets, we will annotate semantically by using senses derived from high-quality lexicographic repositories.

Keywords: Digital lexicography; Natural Language Processing, Computational Linguistics, Corpus Linguistics; Word Sense Disambiguation.

A Use Case of Automatically Generated Lexicographic Datasets and Their Manual Curation

Dorielle Lonke[1], Raya Abu Ahmad[1], Volodymyr Dzhuranyuk[1], Maayan Orner[1], Ilan Kernerman[1]

[1] K Dictionaries, Tel Aviv

E-mail: dorielle@kdictionaries.com, raya@kdictionaries.com, vova@kdictionaries.com, maayan@kdictionaries.com, ilan@kdictionaries.com

This paper provides an overview of a multi-layer project combining machine and manual processes in linking multilingual lexicographic resources and leading to the generation of over 200 new language pairs and the update of over 50 existing ones. In the first phase, we create multilingual glossaries by reversing entries from the Password English multilingual dataset of K Dictionaries, reformulating the L1 translations into headwords, aligning them to the original English entries that become their translations, and adding the other language translations of those English entries. The reversal is supplemented by rule-based algorithms to reduce noise; merge, duplicate and separate entries; and check duplicate senses for similar or identical definitions and examples of usage. This is followed by manual detection and amendment of erroneous grammatical categories and faulty meanings, and editing the translation links. The next phase concerns cross-linking each semi-automatically generated multilingual glossary from the first phase with another full lexicographic resource of that L1 from the Global Multilingual Data Series, including its own bilingual versions whenever available. We present the main tasks involved in this project, featuring the automated operations combined with post-editing, the outcomes, our conclusions and further plans.

Keywords: auto-generated data; automatic post-editing; semi-automated processes; manual curation; resource cross-linking

Abstracts of keynote talks

A White Paper on the Future of Academic Lexicography

Kris Heylen, Vincent Vandeghinste

Dutch Language Institute (INT), Netherlands
E-mail: Kris.Heylen@ivdnt.org, vincent@ccl.kuleuven.be

Academic, evidence-based lexicography has a long tradition of analyzing large amounts of language data in a scientific way in order to compile concise, high-quality knowledge about words and their usage with an eye to serving the entire language community. However, lexicography increasingly faces challenges with respect to:

1. its role in society, science and the knowledge economy,
2. the scalability of both the analysis and production process, and
3. the customizability and accessibility of its content for a diverse audience and for integration in new IT applications.

The Lorentz workshop on the "Future of Academic Lexicography" (Leiden 4-9 November 2019) brought together lexicographers and experts from neighboring disciplines like Data Analytics, Artificial Intelligence, Citizen Science, Human-Computer Interaction and Sociology to explore how each of these challenges can be tackled in a multidisciplinary way to strengthen the position of academic lexicography as a locus for scientific research with direct relevance for, and impact on, society. The conclusions and recommendations of the workshop were summarized in a White Paper and will be presented at the start of the panel session and will form the point of departure for the discussion, which will be moderated by Kris Heylen and Vincent Vandeghinste.

Scalability of Maths for Lexicography

Pavel Rychlý

Lexical Computing / Masaryk University, Czechia
E-mail: pary@fi.muni.cz

Many mathematical methods and formulas are used in lexicography to find interesting or important information about words, contexts, and other parts of natural language. The presentation will show several examples of using maths in terms of scalability and highlight the importance of scalability in practical lexicography.

Designing and Populating Specialised knowledge resources: EcoLexicon and by-products

Pilar León Araúz

University of Granada, Spain
E-mail: pleon@ugr.es

The design and population of specialized knowledge resources need a dynamic framework for knowledge extraction and representation. Frame-based term analysis and concept modelling allow for the representation of specialized knowledge in a meaningful way for target users, who need to accommodate specialized notions into previously stored knowledge structures. In EcoLexicon, a terminological knowledge base on the environmental domain, this translates into extracting, organizing and describing specialized concepts and terms in a wide range of different formats. In this presentation, EcoLexicon will be presented together with the methods employed in its construction as well as the by-products that came along, namely the EcoLexicon English corpus, the EcoLexicon Semantic Sketch Grammar and EcoLexiCAT, a terminology-enhanced translation tool.