eLex 2019

Electronic lexicography in the 21st century (eLex 2019): Smart lexicography

Book of abstracts

edited by Iztok Kosem
Tanara Zingano Kuhn

Sintra, Portugal, 1–3 October 2019

elex.link/elex2019

Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography

Book of abstracts

edited by	Iztok Kosem Tanara Zingano Kuhn
published by	Lexical Computing CZ s.r.o., Brno, Czech Republic
proofreading by	Paul Steed
licence	Creative Commons Attribution ShareAlike 4.0 International License

Sintra, October 2019

eLex 2019

ORGANIZERS

CELGA-ILTEC
Centro de Estudos de Linguística Geral
e Aplicada da Universidade de Coimbra

Univerza *v Ljubljani*

cjvt

elexis
european lexicographic infrastructure

LEXICAL COMPUTING

ACKNOWLEDGEMENT

We would like to thank our sponsors and supporting institutions for supporting the conference.

SPONSORS

Hornby
A. S. Hornby Educational Trust

Dictionary.com

Oxford Dictionaries

Porto Editora

elex.link/elex2019

eLex 2019

CONFERENCE COMMITTEES

Organising Committee

Tanara Zingano Kuhn
Margarita Correia
José Pedro Ferreria
Maarten Janssen
Isabel Pereira
Jelena Kallas
Miloš Jakubíček
Iztok Kosem
Simon Krek
Carole Tiberius
Ondřej Matuška
Teja Goli

Scientific Committee

Andrea Abel
Špela Arhar Holdt
Vit Baisa
Gerhard Budin
Nicoletta Calzolari
Lut Colman
Paul Cook
Margarita Correia
Gilles-Maurice de Schryver
María José Dominguez Vazquez
Patrick Drouin
Jose Pedro Ferreira
Edward Finegan
Thierry Fontenelle
Polona Gantar
Yongwei Gao
Radovan Garabik
Alexander Geyken
Kris Heylen
Ales Horak
Miloš Jakubíček
Maarten Janssen
Jelena Kallas
Ilan Kernerman
Maria Khokhlova
Annette Klosa-Kückelhaus
Svetla Koeva
Iztok Kosem
Vojtěch Kovář
Simon Krek
Michal Kren
Tanara Zingano Kuhn
Margit Langemets
Lothar Lemnitzer
Robert Lew
Pilar León Araúz
Nikola Ljubešič
Henrik Lorentzen
Tinatin Margalitadze
Stella Markantonatou
John P. McCrae
Amalia Mendes
Michal Boleslav Měchura
Julie Miller
Victor Mojela
Monica Monachini
Orion Montoya
Sara Može
Christine Möhrs
Chris Mulhall
Carolin Müller-Spitzer
Lionel Nicolas
Sussi Olsen
Vincent Ooi
Isabel Pereira
Jordi Porta
Adam Rambousek
Laurent Romary
Klaas Ruppel
Roser Sauri
Tanneke Schoonheim
Hindrik Sijens
Emma Sköldberg
Nicolai Hartvig Sørensen
Egon Stemle
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamas Varadi
Carlos Valcárcel Riveiro
Serge Verlinde
Piotr Zmigrodzki

elex.link/elex2019

TABLE OF CONTENTS

EXTENDED ABSTRACTS

ABSTRACTS OF PAPERS

ABSTRACTS OF KEYNOTE TALKS

Extended abstracts

Augmented Writing:
New Opportunities for Lexicography?

Henrik Køhler Simonsen

Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg
E-mail: hks.msc@cbs.dk

We live in an age of disruption, and lexicography as a scientific discipline and practice has already been disrupted in several ways. A relatively new disruptive technology is augmented writing, which seems to challenge the role of dictionaries in a dramatic fashion. Augmented writing (AW) may even be described as Lexicography 3.0 and might even be called “smart lexicography”. The discussion offered in this presentation indicates that AW disrupts not only lexicography to a very high degree, but also writing and copyediting.

The presentation discusses the strengths and weaknesses of augmented writing in relation to lexicography, and how AW might challenge lexicography in a number of areas. The presentation also offers various theoretical reflections on AW. The discussion is based on both a structured literature review and on empirical data from a structured test and analysis of sixteen different AW technologies.

Based on the analysis and discussion this presentation offers a number of theoretical considerations on how lexicography is affected by AW. The presentation also offers theoretical reflections on how lexicography may make AW even better by providing what AI cannot; transfer of knowledge into practice, knowledge of the world, emotional knowledge and relational intelligence. Based on the analysis and discussion it was found that the entire lexicographic process chain might very well be severely impacted and to some extent completely changed by AW technologies. The analysis also seems to indicate that AW technologies threaten to leave behind certain lexicographic processes, disciplines and data categories, and that such technologies will have severe implications for the core competences of lexicographers and for the way the user uses AW-supported lexicographic services. The analysis also seem to indicate that the business potential of AW services seems to be significant. Finally, the presentation includes a number of theoretical reflections based on a symbiotic division of labour, which enables both AI and humans to do what they do best.

The analysis of the sixteen AW services and the related findings are thus estimated to be highly relevant for everybody interested in lexicographic business development, lexicographic business models and innovative technologies.

Keywords: augmented writing; artificial intelligence; lexicographic business model

Writing Assistants: From Word Lists to NLP and Artificial Intelligence

Serge Verlinde, Lieve De Wachter, An Laffut, Kristin Blanpain, Geert Peeters, Ken Sevenants, Margot D'Hertefelt

KU Leuven, Leuven Language Institute
E-mail: {serge.verlinde, lieve.dewachter, an.laffut, kristin.blanpain, geert.peeters, ken.sevenants, margot.dhertefelt}@kuleuven.be

As noted in Strobl e.a. (2019), increased attention to writing support along with technological progress have driven the development of a new generation of writing assistants.

Various (meta-)lexicographers (such as Frankenberg-Garcia et al., 2019; Granger and Paquot, 2015; Tarp et al., 2017; Wanner et al., 2013) consider these writing assistants as a tool to (re)use lexicographical data, offering an alternative to the use of (electronic) dictionaries. According to Frankenberg-Garcia et al. (2019: 24), the lack of dictionary consultation skills exhibited by many users is addressed by the development of tools requiring little or no training or instruction.

The writing assistants that we have developed for (academic) English, Dutch (Dutch as a foreign language, general Dutch and academic Dutch) and Afrikaans exhibit the following characteristics, they:

- are web-based applications (similar to, e.g., WriteAway), while other writing assistants are available as an app (e.g. Writefull), a software package (e.g. SWAN) or as an add-in for a word processor (e.g. ProWritingAid);
- offer immediate feedback, which allows users to improve and/or enrich their text;
- support the writing process.

Our writing assistants thus serve the following three functions (Tarp et al., 2017; Ziyuan, 2012):

- correction: by flagging mistakes (as do spell and grammar checkers, such as Grammarly, as well as applications for word combinations, such as HARenES)
- detection: by highlighting words or phrases that may need to be improved (e.g. overly long sentences, recurring patterns)
- prediction: by providing suggestions to enrich the text, as does, e.g., ColloCaid (Lew et al., 2018)

In terms of the typology proposed by Allen et al. (2016), the ILT writing assistants belong to the category of Automated Writing Evaluation tools, with a number of additional features of Intelligent Tutoring Systems, such as more individualised feedback.

While presenting our writing assistants, we aim to show that in order to develop a high-quality application, an interdisciplinary approach is required. Our writing assistants are mainly based on lexicographical data and data from corpus analyses, but we have also integrated NLP techniques, with further opportunities offered by word embeddings and AI-based language models. In the presentation, we will illustrate how we have used each of these building blocks on the basis of a number of specific examples.

Keywords: writing assistants; web-based; correction; detection; prediction

References

Allen, L., Jacovina, M. & McNamara, D. (2016). Computer-based writing instruction. In C.A. MacArthur, S. Graham & J. Fitzgerald (eds.) *Handbook of writing research.* New York, NY: Guilford, pp. 316-329.

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1), pp. 23-39.

Granger, S. & Paquot, M. (2015). Electronic lexicography goes local: Designs and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1), pp. 118–141.

Lew, R., Frankenberg-Garcia, A., Rees, G., Roberts, J. & Sharma, N. (2018). ColloCaid: A real-time tool to help academic writers with English collocations. Kerk, S., Cibej, J., Gorjanc, V., Kosem, I. (eds.) *Proceedings of the XVIII EURALEX International Congress.* Ljubliana University Press, Faculty of Arts, pp. 247-254.

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

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

Wanner, L., Verlinde, S. & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M. & Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*, pp. 472-487.

Ziyuan, Y. (2012). Breaking the language barrier : a game-changing approach. (https://sites.google.com/site/yaoziyuan/publications/books/breaking-the-language-barrier-a-game-changing-approach)

Websites:

ColloCaid. Accessed at: http://www.collocaid.uk/ (17 August 2019)

Dutch writing assistant. Accessed at: https://schrijfassistent.be (17 August 2019). The other writing assistants are licenced products.

Grammarly. Accessed at: https://www.grammarly.com/ (17 August 2019)

HARenEs. Accessed at: http://harenes.taln.upf.edu/CakeHARenEs/check (17 August 2019)

ProWritingAid. Accessed at: https://prowritingaid.com/art/372/Getting-Started-with-ProWritingAid-s-MS-Word-Add-in.aspx (17 August 2019)

spaCy. Accessed at: https://spacy.io/ (17 August 2019)

SWAN. Accessed at: http://cs.joensuu.fi/swan/ (17 August 2019)

WriteAway. Accessed at: http://writeaway.nlpweb.org/ (17 August 2019)

Writefull. Accessed at: https://writefullapp.com/ (17 August 2019)

Analysing the Meanings of *Rumour* in the *Oxford English Dictionary*: A Corpus-based Approach

Yoko Iyeiri

Kyoto University, Kyoto, Japan

Rumour in English is a loan word attested from the 14th century onwards. The *Oxford English Dictionary* provides six different meanings under the entry of this word, half of which are obsolete in contemporary English (*OED*, s.v. *rumour*): "1. General talk or hearsay, not based upon definite knowledge; General talk or hearsay personified; 2. A widespread report of a favourable or laudatory nature (obsolete); Talk or report of a person or thing in some way noted in some respect (obsolete); The fact of being generally talked about; reputation, renown (obsolete); 3. An unverified or unconfirmed statement or report circulating in a community; 4. Loud expression or manifestation of disapproval or protest; an instance of this (obsolete); 5. Clamour, outcry; noise, din; Also an instance of this (chiefly archaic); and 6. Uproar, tumult, disturbance; an instance of this (obsolete)".

This clarifies that the positive meaning attached to *rumour* in the past fell into disuse during the Modern English period. Also, the "loudness" inherent in the meaning of this word in the past disappeared in the course of the history of English. I will examine these historical shifts by exploring examples of *rumour* found in the *Early English Books Online Corpus* and contrasting the results with the usage of *rumour* in contemporary English (Iyeiri 2019, forthcoming). I will demonstrate how quantitative collocational analyses within the framework of corpus linguistics can contribute to the writing of definitions in dictionaries, using *rumour* as an illustrative case. The discussion will also include some idiomatic expressions with the word *rumour.*

While the investigation of the noun *rumour* will be the central concern of my paper, I will also discuss the verb *rumour*, showing how its use expanded in the course of the Modern English period. The result of my exploration shows that the verbal use of *rumour* was very limited in the 16th century, whereas its use gradually gained ground up to the present day. Although it is one of those verbs whose use in today's English is usually confined to the passive (cf. Levin 1993: 107), its use in the active voice was not uncommon in the past. Detailed analyses of lexical items likes this will contribute to the compilation of dictionaries, especially those on historical principles.

Keywords: rumour; Oxford English Dictionary; corpus linguistics; Early English Books Online Corpus; collocation

References

Corpus of Contemporary American English. <https://corpus.byu.edu/coca/> (1 February 2019).

Early English Books Online. <https://corpus.byu.edu/eebo/> (1 February 2019).

Iyeiri, Yoko. 2019 (forthcoming). "Gendai America Eigo no *rumor*: *Corpus of Contemporary American English* no Bunseki kara" (*Rumor* in *Contemporary American English: An Analysis of the Corpus of Contemporary American English*). (In Japanese).

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Oxford English Dictionary <http://www.oed.com/> (1 February 2019).

Enhancing Mobile Learning by Linking Japanese Dictionary Apps

Jack Halpern

The CJK Dictionary Institute, 34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 Japan
E-mail: jack@cjki.org

The emergence of smart platforms has prompted a surge in mobile applications which significantly enhance the user experience of dictionary users and language learners. This paper describes how four mobile apps exploit the unique features of the mobile platform to help learners study Japanese effectively in previously unavailable ways (Halpern, 2015).

The first is a kanji-English dictionary (The CJK Dictionary Institute, 2019), which provides learners with an in-depth understanding of the meanings and functions of kanji in contemporary Japanese. The second is a kanji homophones dictionary (The CJK Dictionary Institute, 2018), which enables more advanced learners to understand the differences between *kun* homophones such as 変える 'change', 代える 'substitute' and 換える 'exchange'. The third is the first *kanji thesaurus* (The CJK Dictionary Institute, 2018), a new type of reference tool, which enables advanced learners to deepen their understanding of the difference between kanji synonyms like 破 ‘break', 壊 'break down' and 崩 'crumble'.

The KKLD app enables the learner to gain a full understanding of kanji through (1) the core meaning, a concise keyword representing the fundamental concept for each kanji, (2) character meanings grouped around the core meaning that show their semantic interrelatedness, (3) numerous illustrative compounds and examples, and (4) an efficient lookup method (SKIP) which enables beginners to look up characters quickly and accurately (in addition to other search methods such as by radical, by reading and by English) (Halpern, 2016). A useful feature of these apps is the use of state-of-the-art *app-to-app linking* technology that enables the user to freely switch between the dictionary apps by tapping on a link. For example, in KKLD tapping the entry for 換える 'exchange' launches the KKUG app to display the homophone group for *kaeru.*

Another application is a ground-breaking *parallel text reader* Libera (The CJK Dictionary Institute, 2017). Libera combines the strengths of traditional bilingual parallel texts with the untapped potential of mobile app technology. This multi-panel platform introduces a new form of hypertext called *Interactive Parallel Text* (Halpern, 2015), which displays texts side by side in separate panels for easy viewing. Tapping on a word/phrase in the kanji panel simultaneously highlights the corresponding segment in the English, hiragana/romaji and dictionary panels, and provides context-sensitive examples, grammar notes, and high-quality (human recorded) audio. This enables learners to enjoy uninterrupted reading combined with efficient vocabulary acquisition. A unique feature is the customized *context-sensitive dictionary* (CSD), which highlights the particular sense of the selected word in context.

The paper also describes how pedagogical lexicography was used to create dictionary apps that enable learners to gain an in-depth understanding of how kanji are used in Japanese. The method of compilation was based on the descriptive approach, whose aim is to record usage based on actual occurrences. Each meaning was written anew, the result of exhaustive semantic analysis, using such techniques as componential analysis and the study of near-synonyms.

In summary, by providing learners with a user-friendly and aesthetically pleasing environment, this new platform makes the learning experience enjoyable and represents an innovative new direction in mobile-assisted language learning.

Keywords: parallel text; dictionary; kanji dictionary; Japanese; linking

References

The CJK Dictionary Institute, Inc. (2019). The Kodansha Kanji Learner's Dictionary. [Mobile application software]. Retrieved from https://apps.apple.com/us/app/kanji-learners-dictionary/id716481897?ls=1

The CJK Dictionary Institute, Inc. (2018). The Kodansha Kanji Usage Guide. [Mobile application software]. Retrieved from https://apps.apple.com/us/app/the-kodansha-kanji-usage-guide/id1069262969?ls=1

The CJK Dictionary Institute, Inc. (2018). Kodansha Kanji Synonyms Guide. [Mobile application software]. Retrieved from https://apps.apple.com/us/app/kodansha-kanji-synonyms-guide/id1406222615?ls=1

The CJK Dictionary Institute, Inc. (2017). Libera: Parallel Text Reader. [Mobile application software]. Retrieved from https://apps.apple.com/us/app/libera-parallel-text-reader/id943767191?ls=1

Halpern, Jack. (2015). Dictionaries and Mobile Tools for Effective Language Learning. [PowerPoint slides]. Retrieved from http://www.cjk.org/cjk/reference/workshop_pp.pdf

Halpern, Jack. (2015). Interactive Parallel Text: A New Paradigm for Language Learning. [PowerPoint slides]. Retrieved from http://www.cjk.org/cjk/reference/elec_poster_pp.pdf

Halpern, Jack. (2016). Compilation Techniques for Pedagogically Effective Bilingual Learners’ Dictionaries. *International Journal of Lexicography*, 29(3), pp. 323–338.

Going Native: Introducing Sanakirja.fi, a Digital-First Smart Dictionary Service

Elina Söderblom, Piia Saresoja

Kielikone Ltd., Helsinki, Finland
E-mail: elina.soderblom@kielikone.fi, piia.saresoja@kielikone.fi

In this demo, we introduce Sanakirja.fi (www.sanakirja.fi), our next-generation bilingual dictionary service with both free and premium features. Our goal has been to imagine a digitally native dictionary – an easy-to-use dictionary as well as a one-stop service for all your language needs. Our efforts have focused on multiple different areas, from the data layer to the user interface and new smart features.

Firstly, our data format is machine-readable JSON instead of human-readable XML. This is one of the core ideologies behind the service: the data has to be easy to parse and easy to display. In addition, we offer our dictionary data to others through an API, so the data model and format is very important.

Next, we have focused on the mobile user experience. Our website is built on responsive technology, aided by the JSON data format which allows us to reorganize the data on the screen without fear of breaking the dictionary entry's readability.

We have also improved the user experience in general – for example, by indexing more terms, reducing the amount of obsolete links, and linking translations back to headwords in the opposite language pair. What is more, we have introduced other smart features such as integrated lemmatization which suggests the base forms of searched words.

Specifically for L2 users, we have useful features such as inflection and conjugation tables for headwords, and speech-to-text to help with pronunciation. Users can also create word lists and practice these words by playing games. For users who want to translate phrases or sentences not found in our dictionary, we have integrated a machine translation service.

Lastly, our service enables users to submit translation requests for words that are missing from the dictionary, and to leave feedback on any dictionary entry. The data can then be updated by lexicographers using our real-time dictionary editor, and the edited entries will be instantly visible to all users.

Keywords: e-dictionary; JSON data; usability; user experience

LexBib: Towards a Reference Platform for the Domain of Lexicography

David Lindemann, Ulrich Heid

Universität Hildesheim, Germany
E-mail: david.lindemann@uni-hildesheim.de

In this demo, we present aspects of workflow, methods and tools employed in the creation of LexBib[1], a corpus and bibliography of metalexicography (Lindemann et al., 2017). So far, we have merged publication metadata extracted from several existing bibliographies and tables of contents of journals and conference proceedings central to our discipline. We also collect full texts, focusing at this stage on publications in English, published between 2000 and 2018; but in the near future we intend to also work on older items and publications in other languages, in order to obtain additional metadata, such as term candidates, topic models, and citation relations, which will be gathered from the full texts by means of computational methods. Term candidates are mapped to a multilingual domain ontology, the development of which is another central pillar in the LexBib project.

It is our goal to build a digital reference platform for the domain of lexicography, with good coverage and high quality data. How to achieve a satisfactory balance between automation and manual work is a constant concern present in both Lexicography and Library Science. By discussing the LexBib workflow we offer preliminary results regarding this central question: How much manual work will be necessary to obtain a comparatively complete and noise-free resource?

As a first milestone, we present the manually validated publication metadata collection so far completed and made available via an open Zotero group. We explain how to make use of the data, and how to collaborate with the project, e.g. by contributing bibliographical data and/or full text contents, by validating metadata, or by contributing localized lexicalizations for integration into the domain ontology.

It will also be possible for authors to enrich their personal profile page on the LexBib online platform with additional information like ORCID IDs or other identifiers and links to home pages, etc. Information related to persons will be curated manually; this includes author name disambiguation, a working step that already has been performed on the present dataset. At Sintra, we ask all authors to check their personal bibliographies at LexBib for completeness, and whether we are using the preferred

[1] See https://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/

name form: for some authors, we have found more than one version in the sources.

Keywords: metalexicography; bibliography; domain ontology; community building

Smart e-Dictionaries on the Continuum of Information Tools in the R-environment

Theo JD Bothma[1], Rufus H Gouws[2]

[1] Department of Information Science, University of Pretoria, South Africa
[2] Department of Afrikaans and Dutch, Stellenbosch University, South Africa
E-mail: theo.bothma@up.ac.za, rhg@sun.ac.za

Traditional dictionaries offer curated data to end-users. End-users should therefore be able to find the correct data to meet their information needs. Lexicographers add a caveat – the information need should be a lexicographic information need. However, end-users do not necessarily know the exact scope of lexicographic information. A further complication arises where words are not included in the dictionary. If the lemma is found, the dictionary does not necessarily provide the end-user with only the correct data for the context or does not necessarily contain the information required to satisfy specific needs. Dictionary articles can still demand considerable interpretation by the end-user to select the appropriate meaning or equivalent.

Users often have to consult multiple sources in a paper-based environment, which could be time consuming, and could lead to end-users' information needs not adequately satisfied.

The situation is different in the e-environment, where the end-user can navigate between sources – both e-dictionaries and other e-sources. This is especially evident on various e-book platforms. One can link multiple dictionaries to a text. By clicking on a word, the end-user has access to them; or even to specified articles elsewhere and directly to Google, to search the internet. Such content is obviously not curated, and providing access to such data is therefore anathema to the traditional lexicographer, because they have no control over what is presented to the end-user. A traditional dictionary is the result of an application of data pushing procedures. The online environment enables the use of data pulling procedures that give users access to both curated and non-curated data. The end-user is educated to distinguish between these data types.

These issues will be illustrated using a number of examples:

- Complex dictionary articles, where the context is not necessarily evident, and end-users are required to make informed choices;
- Cases where a word or phrase does not occur in the dictionary, and end-users are required to consult other sources;

- Cases where additional information on demand is required, e.g. multimedia.

These examples will show that, in an online environment, end-users can easily move between a large number of different and disparate information sources to satisfy any specific information need, and that a dictionary is one of a plethora of information sources. End-users have to make informed choices, based on context, prior and general world knowledge. It will be argued that everywhere an information need exists, they have to make such choices. They have the option to ignore the information need, or to explore the concept in more detail, either by consulting a dictionary entry, or by delving deeper into other information sources with the dictionary as point of departure or with a dictionary-external point of departure. The information is therefore available on demand, without risking information overload.

The paper will further explore to what extent the preceding has an influence on a theory/praxis of dictionary use, dictionary construction for an online environment, and ultimately on a general theory of lexicography.

Keywords: smart e-dictionaries; information tools; end-users; general theory of lexicography

Catching New Words from the Crowd: Citizen Science as a Lexicographical Tool

Nava Maroto[1], Miguel Sánchez Ibáñez[2,] Pablo Gómez Sanz[3]

[1] Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología (Universidad Politécnica de Madrid) and the research group NeoUSal (Universidad de Salamanca), ETSI de Telecomunicación, Avda. Complutense, 30, 28040, Madrid (Spain)
[2] Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología (Universidad Politécnica de Madrid) and the research group NeoUSal (Universidad de Salamanca), ETSI de Telecomunicación, Avda. Complutense, 30, 28040, Madrid (Spain)
[3] Escuela Técnica Superior de Ingenieros Informáticos, (Universidad Politécnica de Madrid), Calle de los Ciruelos, s/n, 28660, Boadilla del Monte (Madrid, Spain)
E-mail: mariadelanava.maroto@upm.es, miguel.sanchezi@upm.es, pablo.gomez.sanz@alumnos.upm.es

Since 2009, the research group ATeNeo-NeoUsal, which runs the Observatory of Neology of the Spanish region of Castile and Leon, has detected, extracted and classified over 5,500 neologisms appearing in general newspapers published in the Spanish region of Castile and Leon.

The identification and collection of neologisms has proved to be a difficult and time-consuming task. Moreover, the lexicographic approach (Freixa & Sole, 2006) alone is an insufficient filter in the automatic detection of neologisms from newspapers (Sánchez Ibáñez, 2018). Therefore, in recent years our group has started to consider the possibilities of managing and organizing such a vast amount of valuable data. First, we designed a scoring scale, which positively assesses the neologisms that conform to their main characteristics (formal instability, motivation, diachronic novelty) and penalizes those that do not do so. We implemented and applied it to our corpus, and presented the preliminary (and quite satisfactory) results at eLex 2017. However, some flaws found in the process, like the selection of non-neological units as neologisms even after the application of the scoring scale, made us go further in the management of our neologisms. That is why we introduced a citizen science (European Commission, 2013) approach in the development of our task. We believe that ordinary citizens should have a say in the process of identification and compilation of new words, and that technology can facilitate interactions with the "crowd" in this task. Also, their linguistic intuition as speakers is a key validation tool that should be included in our methodology. That is why we have designed a mobile application that allows registered users to identify and assess neologisms.

The aim of the app is to allow us to interact with lay people in order to identify, contrast and assess neologisms. This process obviously needs to be monitored and mediated by linguists, who will guide and motivate collaborators through the use of games and small competitions that will boost collaboration. Participants' motivation and training is a key issue that we are addressing from a citizen science approach, while considering the role that public libraries may play in dissemination and training (Maroto, 2019).

The app (Gómez Sanz, 2019), which is now ready to enter the production stage, will serve as a platform for the interaction between lexicographers and the general public. On the one hand, citizens are expected to collaborate in the identification, assessment and broadcasting of new words through gamification. Linguists, on the other hand, will benefit from their suggestions to help in the process of validation of new words for their inclusion in a dictionary of neologisms. Control mechanisms are considered in order to ensure that the data introduced by the contributors meet the linguistic requirements.

In this presentation we will describe the app and the envisaged protocol to motivate and train contributors, as well as the preliminary results and conclusions drawn from our first uses.

Keywords: crowdsourcing; neology; app development; citizen science

References

European Commission. 2013. *Green Paper on Citizen Science. Towards a Society of Empowered Citizens and Enhanced Research.* Accessed at https://ec.europa.eu/digital-single-market/en/news/green-paper-citizen-science-europe-towards-society-empowered-citizens-and-enhanced-research (20 August 2019)

Freixa, J. & Solé, E. (2006). Análisis lingüístico de la detección automática de neologismos léxicos, *Sendebar,* 17, pp. 135-147.

Gómez Sanz, P. (2019) *Desarrollo de una aplicación móvil para la captura y evaluación de palabras nuevas en español.* Final Year Project, Universidad Politécnica de Madrid.

Maroto, N. (2019). *Las bibliotecas públicas como punto de encuentro para la ciencia ciudadana. Experimento en el ámbito de la lingüística aplicada.* Master thesis, Universidad Carlos III de Madrid.

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

Smart Advertising and Online Dictionary Usefulness

Anna Dziemianko

Adam Mickiewicz University, Poznań, Poland
E-mail: danna@wa.amu.edu.pl

Online advertisements are smart due to targeting. Personalized, targeted advertisements often trail Internet users to make them purchase the goods they have been looking for. One obvious question is whether smart advertising influences online dictionary use.

The aim of the paper is to investigate the effect of targeted advertisements in online dictionaries on language reception, production and learning. Four questions are posed:

- Does advertisement targeting affect dictionary-based language reception, production and retention?
- Are advertisement-free online dictionaries as useful for language reception, production and learning as those with advertisements (either targeted or non-targeted ones)?
- Does the time of online dictionary consultation depend on advertisement targeting?
- Is the effect of advertisements (targeted or otherwise) on language reception, production and retention dependent on sense position in the entry?

In addition to those primary aims, the study investigates dictionary users' attitudes to targeted and non-targeted advertisements.

To achieve the major aims of the study, a pre-test, a main test and a retention test were conducted online. The main test involved 14 infrequent words which had two-sense entries in e-OALD9. For each test item there were two tasks: decoding (the meaning of the word was to be explained) and encoding (the word had to be used in a sentence). Access to three OALD9-based dictionary versions was provided: with targeted advertisements, with non-targeted advertisements and without any advertisements. In the other two tests, the participants explained the words without any dictionary. The pre-test identified the cases where the words were known. Meaning retention was checked in the post-test. 162 learners of English (C1 in CEFR) participated in the experiment. 56 subjects used dictionaries with advertisements targeted at them, 52 those with non-targeted advertisements, and 54 those with no advertisements. Apart from the tests, a short survey was carried out to explore the participants' opinions on advertisements in online dictionaries.

The results reveal that either targeted or non-targeted advertisements in online dictionaries do not significantly affect language reception, production and learning, irrespective of sense position in entries. However, both targeted and non-targeted advertisements prolong dictionary consultation. Non-targeted advertisements were found to be a little more disruptive than targeted ones, but the assessment of advertisements was not dependent on the experimental condition. The implications and limitations of the findings are explored in the full paper.

Keywords: online dictionary; targeted advertisements; dictionary use; reception; production; learning time

How Much "Tourism" Is There in Dictionary Apps? An Empirical Study of Lexicographical Resources on Mobile Devices (German, Italian, Spanish)

Carolina Flinz[1], Maria Egido Vicente[2]

[1] University of Milan, Italy
[2] University of Salamanca, Salamanca, Spain
E-mail: carolina.flinz@unimi.it, mariaegido@usal.es

In the general teaching-learning process of foreign languages, the use of electronic dictionaries accessed via hand-held devices has grown considerably over the past two decades, especially since the arrival of the first generations of digital natives into the classroom. This new type of user, and the development of educational processes within the framework of an increasingly digitalized society, has led to the emergence of a large variety of electronic dictionaries and dictionary apps (Nesi, 2000). Lexicographers, dictionary designers and publishers soon became aware of the advantages of the new medium and "jumped on the appification bandwagon" (Gao, 2013: 213). Currently, the app format presents a wide range of new possibilities and challenges compared to print and web dictionaries (Holmer et al., 2015: 356).

Numerous studies have focused not only on the specific case of the use of bilingual electronic dictionaries, but also on the pairs of languages under study in the present contribution, German-Italian and German-Spanish (Domínguez et al., 2013; Flinz, 2014; Nied Curcio, 2014; Meliss, 2015; Fernández et al., 2016;). Fewer studies have addressed the topic of the production and use of dictionary apps (Gao, 2013; Marello, 2014; Simonsen, 2014; Holmer et al., 2015) while app studies dealing with the pairs of languages mentioned above are rare (Nied Curcio, 2014).

This work approaches the study of bilingual dictionary apps from the content perspective with the aim of describing the information provided to Spanish and Italian learners of German for Specific Purpose in Tourism Faculties. The framing of our study in the context of the language of tourism derives from the specific communicative need of an increasingly large group of students, when considering the economic and social impact that tourism has in southern European countries such as Spain and Italy (Cortés-Jiménez & Pulina, 2006). However, the results obtained can be extended to a more general context.

Under this premise we will carry out an analysis of two of the most popular bilingual dictionary apps for general language among Italian- and Spanish-speaking students of German as a Foreign Language (Leo and Pons) from the search of 50 common terms

in the field of tourism. Our aim is twofold: (1) to evaluate if the analysed dictionary apps really use the advantages of the new medium in comparison with web dictionaries (Leo and Pons), and (2) to reflect on the relevance of these tools in the teaching-learning context of German for Specific Purposes in the field of Tourism. For this purpose, we will first check the presence of the selected terms in the dictionary apps and web dictionaries. Then we will focus on their macro- and microstructural features, concentrating on different aspects, among others, outside matter, access structure to the entries, integration of multimedia features, organization of entries, cross-reference structure (Engelberg et al., 2016: 157-160), and reference to cultural contents. Finally, reflections on the influence of the teaching-learning of German for LSP will also be drawn.

Keywords: mobile devices; dictionaries; tourism; German, Italian, Spanish

References

Cortés-Jiménez, I. & Pulina, M. (2006). Tourism and Growth: Evidence for Spain and Italy. In *Proceedings of the 46th Congress of the European Regional Science Association.*

Domínguez Vázquez, M. et al. (2013). Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In M. J. Domínguez Vázquez (ed.) *Trends in der deutsch-spanischen Lexikographie.* Frankfurt am Main: Peter Lang, pp. 135-172.

Engelberg, S., Müller-Spitzer, C. & Schmidt, T. (2016): Vernetzungs- und Zugriffstrukturen. In A. Klosa & C. Müller-Spitzer (eds.) *Internetlexikographie. Ein Kompendium.* Berlin/Boston: de Gruyter, pp. 153-195.

Fernández Méndez, M. et al (2016). Uso de diccionarios de aprendizaje: Análisis de una encuesta desde una perspectiva contrastiva sobre el uso de los diccionarios bilingües alemán-español entre aprendices de ELE y DaF. In F. Robles et al. (eds.) *Sprachdidaktik Spanisch - Deutsch. Forschungen an der Schnittstelle von Linguistik und Fremdsprachendidaktik.* Tübingen: Narr Francke Attempto, pp. 73-92.

Flinz, C. (2014). Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernern. In A. Abel et al. (eds.) *Proceedings of the XVI Euralex international Congress: The User in Focus*, 213-224. Retrieved from: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

Gao, Y. (2013). The Appification of Dictionaries: From a Chinese Perspective. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference.* Tallinn/Ljubljana: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 213–224.

Holmer, L., Hult, A.K. & Sköldberg, E. (2015). Spell-checking on the fly? On the use of a Swedish dictionary app. In I. Kosem et al. (eds.) *Proceedings of the eLex 2015 conference, 11-13 Aug. 2015*, pp. 356-371.

Holmer, L., von Martens, M. & Sköldberg, E. (2015). Making a dictionary app from a lexical database: the case of the Contemporary Dictionary of the Swedish Academy. In I. Kosem, M. Jakubiček, J. Kallas & S. Krek (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 32-50.

Marello, C. (2014). Using Mobile Bilingual Dictionaries in an EFL Class. In A. Abel, et al. (eds*.) Proceedings of the XVI Euralex international Congress: The User in Focus.* Bolzano/Bozen, pp. 63–83.

Meliss, M. (2015). Was suchen und finden Lerner des Deutschen als Fremdsprache in aktuellen Wörterbüchern? Auswertung einer Umfrage und Anforderungen an eine aktuelle DaF-Lernerlexikographie. In T. Roelcke (ed.) *Wörterbücher für Deutsch als Fremdsprache - Probleme und Perspektiven. Themenreihe. InfoDaF 4/42*, pp. 401-432. Retrieved from: https://www.degruyter.com/downloadpdf/j/infodaf.2015.42.issue-4/infodaf-2015-0407/infod af-2015- 0407.pdf.

Nesi, H. (2010). Electronic dictionaries in second language vocabulary comprehension and acquisition. The state of the art. In *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000.* Retrieved from: http://www.euralex.org/elx_proceedings/Euralex2000/099_Hilary%20NESI_Electronic%20Dictionaries%20in%20Second%20Language%20Vocabulary%20Comprehension%20and%20Acquisition_the%20State%20of%20the%20Art.pdf.

Nied Curcio, M. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Abel et al. (eds.) *Proceedings of the XVI Euralex international Congress: The User in Focus*, pp. 263-280. Retrieved from: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX%202014_gesamt.pdf.

Simonsen, H. K. (2014). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel et al. (eds.) *Proceedings of the XVI Euralex international Congress: The User in Focus.* Bolzano/Bozen, pp. 249–261.

Online dictionaries and dictionary apps:

LEO = Leo Wörterbücher Deutsch, Italienisch; https://dict.leo.org/italienisch-deutsch/ (12.08.19)

Leo Wörterbücher Deutsch, Spanisch; https://dict.leo.org/spanisch-deutsch/ (12.08.19).

Leo Wörterbuch App for Android-Devices; https://play.google.com/store/apps/details?id=org.leo.android.dict&hl=de (12.08.19)

LEO Wörterbuch App for iOS-Devices; https://apps.apple.com/de/app/leo-wörterbuch/id396838427 (12.08.19)

PONS = Online-Wörterbuch; https://de.pons.com (12.08.19)

Wörterbuch App for Android-Devices; https://play.google.com/store/apps/details?id=com.pons.onlinedictionary&hl=de (12.08.19)

Wörterbuch App for iOs-Devices; https://apps.apple.com/de/app/pons-übersetzer/id577741918?mt=8&ign-mpt=uo%3D4 (12.08.19)

Slipping Through the Cracks of e-Lexicography

Geraint Rees[1], Ana Frankenberg-Garcia[1], Robert Lew[2], Jonathan C. Roberts[3], Nirwan Sharma[3], Peter Butcher[3]

[1] University of Surrey, Guildford, UK
[2] Adam Mickiewicz University, Poznań, Poland
[3] Bangor University, Bangor, UK
E-mail: g.rees@surrey.ac.uk, a.frankenberg-garcia@surrey.ac.uk, rlew@amu.edu.pl, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk, p.butcher@bangor.ac.uk

Thanks to the corpus revolution, which underpins e-lexicography, headword lists and defining vocabularies can now be adjusted to better reflect current language use. Definitions can be enhanced with information that goes beyond introspection alone. Syntactic patterning, lexical collocations and other phraseology can be given more comprehensive coverage. Good dictionary examples are easier to find. Yet despite these unparalleled and undeniable advantages, there are elements of word usage that appear to be slipping through the cracks of corpus analyses and corpus-based lexicographic resources. Issues encountered during the development of the lexicographic database behind the ColloCaid tool provide examples of such slips and a valuable opportunity to reflect on why they come about and how they might be addressed in future e-lexicography projects.

The ColloCaid text editor was designed to help writers enhance their use of general academic English collocations in a way that does not distract them from the writing task at hand (see eLex 2019 demo *ColloCaid: assisting writers with academic English collocations*). As detailed in Frankenberg-Garcia et al. (2019), the lexicographic database behind ColloCaid draws initially on a combination of high-frequency noun, verb and adjective lemmas from three respected corpus-based general academic English word lists: the Academic Keyword List (Paquot, 2010), the Academic Collocation List (Ackermann & Chen, 2013), and the Durrant (2016) subset of the Gardner and Davies (2014) Academic Vocabulary List. The database was then populated using word sketches (Kilgarriff et al., 2004), to automatically extract collocations sorted according to grammar relation from corpora of expert academic English writing, and GDEX (Kilgarriff et al., 2008) to select good corpus examples for language production.

While there is no doubt about the usefulness of the above e-lexicography tools and resources, manual curation of the data became necessary at several points. In this paper we discuss specific cases of high-frequency academic lemmas that slipped through initial selection thresholds, situations where it was necessary to distinguish between interdisciplinary and multidisciplinary lemmas, circumstances where part-of-speech tags had to be overruled, and instances where concordances that ranked highly in

GDEX did not reflect conventional colligation patterns. These examples not only serve as a reminder of the dangers of an uncritical use of lexicographical tools and resources, but also highlight the limits of certain e-lexicographical methods and suggestions for their improvement.

Keywords: writing assistants; collocation; corpora; EAP; word lists

Acknowledgements

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

References

ColloCaid (online) http://www.collocaid.uk [05/09/2019]

Frankenberg-Garcia, A. Lew, R., Roberts, J., Rees, G. & Sharma, N. (2019) Developing a writing assistant to help EAP writers with collocations in real time, *ReCALL*, 32(2), pp. 23-39.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008) GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX International Congress 2008*, Universitat Pompeu Fabra, Barcelona, pp. 425-433.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell., D. (2004) The Sketch Engine. *Proceedings from EURALEX 2004*, Lorient, France, pp. 105-116.

ColloCaid: Assisting Writers with Academic English Collocations

Ana Frankenberg-Garcia[1], Robert Lew[2], Geraint Rees[1], Jonathan C. Roberts[3], Nirwan Sharma[3], Peter Butcher[3]

[1] University of Surrey, Guildford, UK
[2] Adam Mickiewicz University, Poznań, Poland
[3] Bangor University, Bangor, UK
E-mail: a.frankenberg-garcia@surrey.ac.uk, rlew@amu.edu.pl, g.rees@surrey.ac.uk, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk, p.butcher@bangor.ac.uk

Despite the existence of excellent resources that writers can use to look up collocations – for example, general language dictionaries, collocation dictionaries, corpora and tools like SkELL, Just the Word or FlaxLC – writers may not be aware of them, or may simply not realize that their use of collocations could be enhanced. Another problem is that looking up collocations whilst writing can be distracting and disruptive. This is particularly true when coping with cognitively demanding tasks, such as academic writing.

This demo introduces you to ColloCaid, a text editor that provides academic English writers with collocation suggestions as they write.

At the time of writing this abstract, the lexicographic database underlying ColloCaid (version 0.3) covers: 551 lemmas derived from well-known general academic English word lists; 9,257 core interdisciplinary collocations extracted from corpora of expert academic writing; 27,771 curated corpus examples of core collocations in context; and 31,401 additional interdisciplinary collocations. For further details about our lexicographic coverage, see Frankenberg-Garcia et al. (2019) and "Slipping Through the Cracks of e-Lexicography: Lessons from ColloCaid" (paper presented at eLex 2019).

The ColloCaid database is integrated with TinyMCE, a widely used open-source text editor that can be accessed from a normal browser, compatible with multiple devices and operating systems, without the need to download additional software. When typed into ColloCaid, the lemmas covered in the lexicographic database are underlined with a green dotted line to signal that collocation suggestions are available. If wanted, users can then click on the highlighted word to trigger interactive menus that display collocations sorted according to grammar relation and strength of association, and examples to support language production (Frankenberg-Garcia, 2015). For further information, see "Collocations in e-Lexicography: Lessons from Human Computer Interaction Research" (paper presented at the eLex 2019 Workshop on Collocations).

In this demo, we invite you to: (1) do a very short collocation awareness exercise; (2) watch a 2-minute video about ColloCaid; (3) try out our prototype on your own device; (4) fill in a short feedback questionnaire; and, last but not least, (5) ask us questions and tell us what you think.

Keywords: collocation; academic writing; writing assistant

Acknowledgements

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

References

ColloCaid (online) http://www.collocaid.uk [05/09/2019]

Flax LC (online) http://flax.nzdl.org/greenstone3/flax [05/09/2019]

Frankenberg-Garcia, A. (2015) Dictionaries and Encoding Examples to Support Language Production. *International Journal of Lexicography*, 28(4), pp. 490-512.

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.

Just the Word (online) http://www.just-the-word.com/ [05/09/2019]

SkELL (online) https://skell.sketchengine.co.uk/run.cgi/skell [05/09/2019]

TinyMCE (online) https://www.tiny.cloud [05/09/2019]

Nénufar: Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource

Hervé Bohbot[1], Francesca Frontini[1], Fahad Khan[2], Mohamed Khemakhem[3,4,5], Laurent Romary[3,4,6]

[1] Praxiling UMR 5267 CNRS, Université Paul-Valéry Montpellier 3
[2] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa
[3] ALMAnaCH - INRIA
[4] CMB - Centre Marc Bloch, Berlin
[5] UPD7 - Université Paris Diderot, Paris 7
[6] BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
E-mail: herve.bohbot@cnrs.fr, francesca.frontini@univ-montp3.fr, fahad.khan@ilc.cnr.it, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

The *Petit Larousse Illustré* (PLI) is a monolingual French dictionary which has been published every year since the 1906 edition[2], and which is therefore a fundamental record of the evolution of the French language. As a consequence of the pre-1948 editions of the PLI entering the public domain in 2018 the *Nénufar* (Nouvelle édition numérique de fac-similés de référence) project was launched at the Praxiling laboratory in Montpellier with the aim of digitizing and making these editions available electronically. The project is still ongoing; various selected editions from each decade are going to be fully digitized (so far the 1906, 1924 and 1925 editions have been completed), and changes backtracked and dated to the specific year.

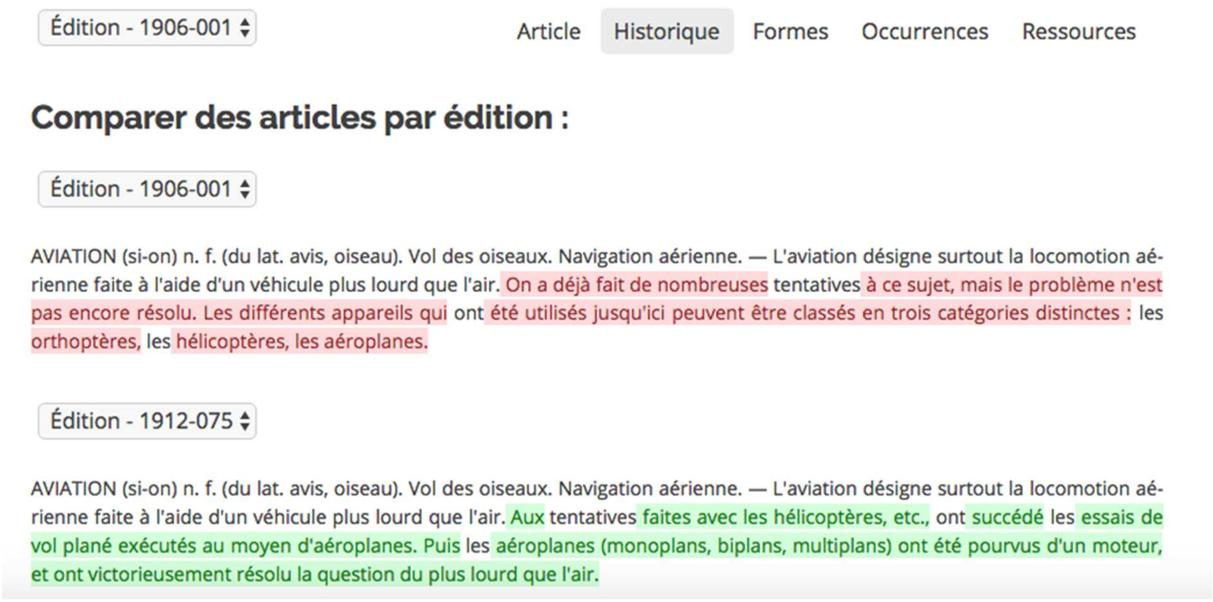
Nénufar's primary aim is to make the editions available and searchable via an advanced search interface which will not only enable the selective querying of text by lemma and type of content (definitions, examples, ...), but crucially also detect and study changes by comparing different editions. In order to do so, a specific web interface has been put in place (see Figure 1 and the project's website[3]). Alongside the digitized text, the Nénufar website[4] contains high quality scans for each page. In compliance with current open data best practices (Wilkinson et al., 2016), the project also aims to make the source data available separately from the querying interface both for research and for

[2] Published in 1905 but dated 1906.

[3] http://nenufar.huma-num.fr/?article=3807

[4] A similar project which presents data and scans from subsequent editions of the same legacy dictionary has been carried out by the team behind the Swedish Academy's Wordlist (see Holmer, Malmgren, and Martens (2016) and http://spraakdata.gu.se/saolhist/).

long-term preservation. The primary encoding format is TEI-XML; however in our case the TEI encoding is closely inspired by the latest version of the TEI-Lex0 (Bański et al., 2017, Romary & Tasovac, 2018) guidelines for encoding lexicographic resources[5], which are based upon TEI. The choice of a TEI[6] based approach allows the Nénufar project to align itself to other pre-existing initiatives and tools. By aligning ourselves to TEI-Lex0 we will be able to make use of digitisation tools such as Grobid (Khemakhem et al., 2017) which have TEI-Lex0 as their native format and which have already been tested and used within the Nénufar project to speed up the digitization of new editions. In addition we will be able to make use of ongoing initiatives to convert TEI-Lex0 datasets to RDF using the W3C recommendation for publishing lexicons as Linked Data, namely OntoLex-Lemon (McCrae et al., 2017; Bosque-Gil et al., 2016) which will allow for the publication of the Nénufar dataset as an LOD graph. The LOD version of the Nénufar dataset, now currently being developed, will be queryable from the available SPARQL endpoint and contain all available editions as one single graph, allowing for expert users to perform complex queries that could detect systematic changes in the dataset. The LOD version is particularly adapted to be linked to other datasets; more recent editions, once added, could also be of interest for NLP applications.



Comparer des articles par édition :

AVIATION (si-on) n. f. (du lat. avis, oiseau). Vol des oiseaux. Navigation aérienne. — L'aviation désigne surtout la locomotion aérienne faite à l'aide d'un véhicule plus lourd que l'air. On a déjà fait de nombreuses tentatives à ce sujet, mais le problème n'est pas encore résolu. Les différents appareils qui ont été utilisés jusqu'ici peuvent être classés en trois catégories distinctes : les orthoptères, les hélicoptères, les aéroplanes.

AVIATION (si-on) n. f. (du lat. avis, oiseau). Vol des oiseaux. Navigation aérienne. — L'aviation désigne surtout la locomotion aérienne faite à l'aide d'un véhicule plus lourd que l'air. Aux tentatives faites avec les hélicoptères, etc., ont succédé les essais de vol plané exécutés au moyen d'aéroplanes. Puis les aéroplanes (monoplans, biplans, multiplans) ont été pourvus d'un moteur, et ont victorieusement résolu la question du plus lourd que l'air.

Figure 1: Comparing the differences in the definition of *aviation* (aviation) between the 1906 and 1912 editions. In the first one air navigation is still a possibility, whereas in the second one it is a reality.

The presentation will illustrate the state of the art of the project, showcase the web

[5] We chose not to encode the dataset directly into TEI-Lex0 since the guidelines are still incomplete. But our dataset is as closely aligned as possible to the current version of the TEI-Lex0 guidelines in order to make any future conversion as straightforward as possible.

[6] On the Nénufar website select an entry and then *Ressources* to inspect the xml encoding.

site, outline the principles guiding the TEI encoding with examples, and discuss the issues concerning the conversion from TEI-Lex0 to OntoLex-Lemon.

Keywords: French dictionary; diachronic lexical resource; Petit Larousse Illustré; TEI; OntoLex-Lemon

References

Bański, P., Bowers, J. & Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In I. Kosem et al. (eds.) *Proceedings of eLex2017, Leiden, Netherlands.* Brno: Lexical Computing Ltd.

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M. & Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. In *GLOBALEX 2018 - Globalex Workshop at LREC2018.* Miyazaki, Japan.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. & Aguado-de-Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. In *GLOBALEX 2018 - Globalex Workshop at LREC2018.* Miyazaki, Japan.

Holmer, L., Malmgren, S.-G. & von Martens, M. (2016). "SAOLhist.se – för allmänt och vetenskapligt bruk." *Nordiske Studier i Leksikografi*, 13, pp. 349–58.

Khemakhem, M., Foppiano, L. & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In I. Kosem et al. (eds.) *Proceedings of eLex2017, Leiden, Netherlands.* Brno: Lexical Computing Ltd.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem et al. (eds.) *Proceedings of eLex2017, Leiden, Netherlands.* Brno: Lexical Computing Ltd.

Romary, L. & Tasovac, T. (2018). 'TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources' (presented at the TEI Conference, Tokyo, 2018) https://hal.inria.fr/hal-02265312.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018 doi:10.1038/sdata.2016.18.

EFL Learners' Use of Smartphones as a Substitute to Electronic Dictionaries

Makimi Kano, Gobel Peter

Kyoto Sangyo University, Kamigamo Motoyama Kita-ku, Kyoto, Japan
E-mail: makimik@cc.kyoto-su.ac.jp, pgobel@cc.kyoto-su.ac.jp

Since the spread of smartphones, dictionary use in classrooms has changed dramatically. Many studies have been done on the effectiveness of the use of electronic dictionaries in EFL classrooms, reporting positive results especially in students' vocabulary mastery (Yanti, 2016; Rezaei & Davoudi, 2016). However, with many dictionary sites and translation sites available free of charge, students choose to use their smartphones rather than their electronic dictionaries when they read and write in English. There have been several studies about positive influences of the use of online translation sites on EFL students (Benda, 2013; Groves & Mundt, 2015), but without proper understanding and instructions by the teacher, students cannot benefit from these convenient tools.

This study aims to reveal students' attitudes toward dictionary use in writing in English in an EFL classroom in Japan. The study is carried out in low-intermediate writing classes for first/second-year students in a private university in Japan. In these classes, students learn how to write various types of short essays. In this study, four questionnaires were created based on previous observation and administered over a semester to uncover students' search behaviour. We try to find out what kind of devices, websites, and/or apps they use to complete the assignments, if they use them differently depending on the tool, and if they use them differently depending on the task (brainstorming, writing, editing). Preliminary results suggest that students are highly dependent on smartphones during the assignments, even though all of them have their electronic dictionaries with them. Rather than looking up words in a dictionary, a lot of students used online translation sites such as Google Translate, where you get only one answer to an inquiry. Some prefer to use chat-style online translation, such as LINE Translation, with which you type in a word or phrases just as you chat with your friends, and the answer comes back right away in a conversation bubble. These practices, of course, lead to poor or erroneous word choices in their writing, but the fact that only a single search result is displayed (whether appropriate for the context or not) seems to be the convenience they like. However, over the semester most students learned to use both dictionaries and online apps, increasing their use of multiple tools. No students used dedicated dictionary apps on their smartphones. The questionnaire results show that there is little difference in use, depending on the tool, and only slight differences in use depending on the task, with more spelling checks in brainstorming and more usage checks in editing. The survey results will then be discussed in light of creating instructional materials for a more efficient use of reference tools in the classroom.

Keywords: Japanese EFL students; electronic dictionaries; Internet dictionaries/applications; dictionary consultation behaviour

New Digital Didactic Dictionary Model to Facilitate the "Entry" into New Languages. An Application to Latin and German

Manuel Márquez, Ana Fernández-Pampillón, Paloma Sánchez

Universidad Complutense de Madrid, Facultad de Filología, C/ Ciudad Universitaria s/n,
28040 Madrid, Spain
E-mail: manmarq@ucm.es, apampi@ucm.es, palomash@ucm.es

This work presents a new cognitive dictionary model that facilitates the comprehension and use of a new language in the first phases. The problem addressed is students' lack of motivation and sufficient linguistic knowledge on which to rely, what hampers the learning of a new language, among other factors. This situation happens, mainly, in secondary and early tertiary education and, in many cases, leads to early abandonment of learning.

One way to address the problem is the proposal presented in this paper. The aim is to make students feel they are able to understand simple sentences in the new language only with their mother tongue knowledge and with the support of the new dictionary. The lexical-semantic information contained in the dictionary is based on valence theory and Fillmore case grammar (Ágel & Fischer, 2010; Fillmore, 1968; Babic, 2011). The novel idea is to help to understand the global functioning of the language using a "sentence-puzzle" metaphor. According to this metaphor, a sentence can be understood as a puzzle whose pieces are its words. The meaning of the sentence is constructed from the meaning of the main piece, the verb. The verb needs, to complete its meaning, other words (with the role of agent, object, beneficiary, ...) that must semantically fit with it. Since this idea is valid for both the mother tongue and the new language, the learner can easily apply it to the new language.

The dictionary is, therefore, conceived as a repository of pieces, which are the lemmas. The verb, which is the main piece, contains in its microstructure the predicative frames with basic morphological and semantic information that helps the learner how to look for its arguments. Nouns and adjectives contain, in addition to morphological information, their semantic features, which must fit with the semantic features expected in verb arguments. The microstructure of the dictionary is completed with images about the meaning of the lemmas, the translation into the learner's mother tongue and examples, and in the case of the German dictionary, its pronunciation.

The macrostructure of the dictionary is based on the Hierarchical Faceted Category model (Denton, 2003). There are eight faceted-taxonomies, one for each lexical category.

The taxonomy of verbs is the most complex, with a depth of up to six levels. It includes the monovalent, bivalent and trivalent categories taking into account the possible meanings of each verb. Nouns are organized according to their semantic features. In the current version, adjectives, adverbs and prepositions are barely developed as it has been found that they are not necessary in an initial learning phase.

The dictionary model has been implemented in two inflected languages, Latin and German for Spanish mother tongue students (German_DDDictionary; Latin_DDDictionary, 2016). The didactic effectiveness has been empirically tested through four case studies with high-school students (Márquez & Chaves, 2016). The results indicate an improvement in grades and motivation. Currently, we are working on the construction of a virtual course to introduce into Latin or German thanks to the dictionary.

Keywords: computational lexicography; valence grammar; Latin dictionary; German dictionary; semantic features

Acknowledgements

This work has been partially funded by Spanish Government research projects TIN2014-52010-R (RedR+Human) and TIN2017-88092 R (CetrO+Spec) and by Universidad Complutense de Madrid Educational Innovation projects PIE 164-2018, PIE 193-2017 and PIE 269-2016.

References

Ágel, V. & Fischer, K. (2010). Dependency Grammar and Valency Theory. In B. Heine & Narrog, H. (eds.) *The Oxford Handbook of Linguistic Analysis.* Oxford, pp. 225-257.

Babic, M. (2011). Tesnière's Dependency Grammar And Its Application In Teaching Latin: Aslovenian Experience. In R. Oniga et al. (eds) *Formal Linguistics and the Teaching of Latin: Theoretical and Applied Perspectives in Comparative Grammar.* Cambridge, pp. 401-412.

Denton, W. (2003). "How to make a faceted classification and put it on the web". In Miskatonic University Press. Available at: http://www.miskatonic.org/library/facet-web-howto.html (27 August 2019)

Fillmore, C. J. (1968). The Case for Case". In Bach & Harms (eds.) *Universals in Linguistic Theory.* New York: Holt, Rinehart, and Winston, pp. 1-88. Available at: http://linguistics.berkeley.edu/~syntax-circle/syntax-group/spr08/fillmore.pdf (27 August 2019)

German_DDDictionary: *Digital Didactic Dictionary of German* (2016). Universidad Complutense de Madrid. Accessed at:

http://repositorios.fdi.ucm.es/DiccionarioDidacticoAleman/ (27 August 2019)

Latin_DDDictionary: *Digital Didactic Dictionary of Latin* (2016). Universidad Complutense de Madrid. Accessed at: http://repositorios.fdi.ucm.es/DiccionarioDidacticoLatin/ (27 August 2019)

Márquez Cruz, M. & Chaves Yuste, B. (2016). A Latin Functionalist Dictionary as a Self-Learning Language Device: Previous Experiences to Digitalization. *Education Sciences* 6 (23). Available at: https://www.mdpi.com/2227-7102/6/3/23 (27 August 2019)

The Sintra Variations – Thinking Outside the Box in Designing Online Dictionaries

Frank Michaelis, Carolin Müller-Spitzer, Sascha Wolfer

Institut für Deutsche Sprache, D-68161 Mannheim, Germany
[2] Affiliation of Author2, Address
E-mail: [michaelis/wolfer/mueller-spitzer]@ids-mannheim.de

In music, variations and improvisation on a theme are a common creative technique: existing material is modified step by step in melody, rhythm or harmony. If this succeeds, something completely new and unexpected can arise. In a similar way, the decomposition and (re)composition of a product is a common problem-solving strategy in design to achieve new solutions for old problems.

In our talk we apply this technique in the field of dictionary interfaces. To this end, we will focus on one dictionary entry with one particular, but characteristic, problem, i.e. that the entry is too long to show all content on one screen. Typically, a dictionary entry is structured in a horizontal (different senses of one word) and in a vertical line (different kind of information on one sense). So, the dictionary designer is confronted with the issue of which information to present first at the expense of others.

In most of the current online dictionaries, an expandable text structure is the method of choice. Other common design solutions are tab views or more print oriented options with signposts or menus (cf. e.g. Koplenig & Müller-Spitzer, 2014; Dziemianko, 2015). It is less widespread to use more vertical screen space. However, this could also be an interesting solution for mobile devices (cf. e.g. Storjohann, 2018).

We will show how variations and rearrangements of the basic building blocks of a dictionary entry can be used to create new and completely different interfaces. With focus on usability and new hardware devices, we will compare different design options and try to figure out potential advantages these new interfaces may have over standard ones.

Our “variations” may not be suitable for productive use. But similar to concept studies in the automotive industry, they may help us to get a better understanding of our design problems and hint at directions we can pursue, when revising existing dictionary interface forms.

Keywords: user interface; adaptive design; user-centred design; online dictionaries

References

Dziemianko, A. (2015). An insight into the visual presentation of signposts in English learners' dictionaries online. *International Journal of Lexicography* ecv040. doi:10.1093/ijl/ecv040.

Koplenig, A. & Müller-Spitzer, C. (2014). Questions of design. In Carolin Müller-Spitzer (ed.), *Using Online Dictionaries.* Berlin, Boston: de Gruyter, pp. 189–204.

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

From Thousands of Graphics to One Conclusion. Visualization of the Vocabulary of Quotation Expressions

Ngoc Duyen Tanja Tu

Leibniz Institute for the German Language, R5,6-13 68161 Mannheim, Germany
E-mail: tu@ids-mannheim.de

Nowadays electronic corpora are larger and larger, and thus visualisation of the data is a very important task if you want to explore it in a practical way. For this reason visual linguistics is attracting a growing attention (Bubenhofer, 2018: 26). But as Bubenhofer (2018: 25-26) states, the field is not yet well-developed, and more reflection is needed on the use of visualization and its relevance for linguistic work.

In my study, I analyse the vocabulary of quotation expressions. Quotation expressions are expressions which refer a direct or indirect speech representation (e.g. the verb *says* in *She says: "I am happy."* or the phrase *pat on the back* in *He pats him on the back: "I am proud of you."*). Analyses conducted by Kurz (1966), Michel (1966) and Jäger (1968) already showed that the vocabulary of quotation expressions is very large and diverse. But none of these authors worked out the factors which account for the diversity of the vocabulary of quotation expressions. Finding this out is not an easy task, because corpora are too big to be investigated in a qualitative way. Fortunately, corpora are now available in electronic form. As a consequence, it is possible to implement algorithms which visualize different aspects of the data, i.e. to use exploratory graphics (Unwin et al., 2008: 4). Therefore, conclusions can be made with the help of visualization.

My poster presents the use of exploratory graphics to find out which factors determine the use of a certain quotation expressions. The main factor which will be investigated is the position of the quotation expression in the syntagm – i.e., does the quotation expression stand before, between or after a speech representation? Every graphic shows another combination of the main factor and an additional factor, e.g. one will show the distribution of quotation expressions among the three positions (= the main factor) with regard to their semantic class (= an additional factor), another will show the distribution of quotation expressions among the three positions with regard to their part of speech. Overall, I want to reflect on the use of different visualization methods and also point out why the visualization of linguistic data is so important for corpus linguistic work.

With regard to the data I extracted about 2,000 quotation expressions from text

samples of the “Redewiedergabe-Korpus”[7]. The corpus consists of 293 fictional (narratives) and 327 non-fictional (newspaper and magazine articles) German text samples, each about 500 words, and written between 1840-1920. The text samples are manually labelled for quotation expressions and the direct speech they refer to.

Keywords: corpus linguistics; visualisation; vocabulary of quotation expressions

References

Bubenhofer, N. (2018). Visual Linguistics: Plädoyer für ein neues Forschungsfeld. In N. Bubenhofer & M. Kupietz (eds.) *Visualisierung sprachlicher Daten.* Heidelberg: Heidelberg University Publishing, pp. 25-62.

Jäger, S. (1968). Die Einleitungen indirekter Reden in der Zeitungssprache und in anderen Texten der deutschen Gegenwartssprache. *Muttersprache* 78, pp. 236-249.

Kurz, J. (1966). *Die Redewiedergabe. Methoden und Möglichkeiten.* Leipzig: Karl-Marx-Universität Leipzig, Sektion Journalistik.

Michel, G. (1966). Sprachliche Bedingungen der Wortwahl.. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationswissenschaft,* 19, pp. 103-129, 213-240, 339-364, 515-532.

Unwin, A., Chen, C. & Härdle, W. (2008). Introduction. In C. Chen, C., W. Härdle, W. & A. Unwin (eds.) *Handbook of Data visualization. With 569 Figures and 50 Tables.* Berlin/Heidelberg: Springer-Verlag, pp. 3-14.

[7] The corpus was created by the DFG-funded project “Redewiedergabe”. It is available on GitHub: https://github.com/redewiedergabe/corpus.

Lexicographic System of Explanatory Dictionary in Digital Environment

Yevhen Kupriianov[1], Iryna Ostapova[2], Volodymyr Shyrokov[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@gmail.com, vshirokov48@gmail.com

The experience of elaborating the project “Dictionary of the Ukrainian language in 20 volumes” is described. The dictionary in question is a heavily revised version of the 11-volume Ukrainian Explanatory Dictionary edited during 1970-1980. Currently the ninth volume has been published. The on-line version of the dictionary is available at: http://services.ulif.org.ua/expl/, SUM20ua.com. The project “Dictionary of the Ukrainian language in 20 volumes” is implemented under the government-sponsored program of developing the National Dictionary Base of Ukraine.

To facilitate the process of compiling the dictionary a virtual lexicographic laboratory (VLL) was created. The VLL offers a networked environment to work with the structure of the explanatory dictionary.

For developing the VLL the method of a lexicographic system was used. The dictionary is considered as a lexicographic system as an information system of a special type. The system architecture corresponds to ANSI/X3/SPARK architecture, in which the conceptual, internal and external data levels can be marked out. A lexicographic data model, which determines the entry structural elements and sets the relations among them, was taken as a conceptual model.

The conceptual model was built up based on the analysis of the printed version of the 11-volume Dictionary. The entire text was converted into digital form (scanned and recognized). Then the digital dictionary text was parsed, and the database was automatically generated to serve as a basis for creating the VLL.

The existing VLL is primarily aimed at supporting the lexicographic process. Currently, the main functions of VLL include:

- user authorization and identification;
- access rights administration (entries viewing, editing, access to interfaces, etc.);
- adding and removing entries in the lexicographic database;

- entry editing (within the given structure);
- representing entries in the selected format (e.g. publishing system format);
- data analysis (lexicographic statistics, editing history of each entry, etc.);
- selecting samples from the database according to the specified parameters;
- online dialogue between lexicographic process participants.

But the issue of using the VLL for conducting research into explanatory dictionary digital text needs deep consideration. The approaches to creating linguistic tools, based on the structural profile of the entry, are discussed. While working with the VLL, the requirements for the computer tools of the explanatory dictionary to be accessible to users with high linguistic competence via the online interfaces were defined. For this purpose "Diccionario de la lengua Española" (ed. 23) (on-line version at https://dle.rae.es/) was chosen for the experiment.

Keywords: digital lexicography; theory of lexicographic systems; lexicographic database; virtual lexicographic laboratory

Integration of the Electronic Dictionary of the 17th—18th C. Polish and the Electronic Corpus of the 17th and 18th C. Polish Texts

Joanna Bilińska[1], Renata Bronikowska[2], Zbigniew Gawłowicz[3], Maciej Ogrodniczuk[3], Aleksandra Wieczorek[2], Mateusz Żółtak[4]

[1] Institute of Western and Southern Slavic Studies, University of Warsaw,
Krakowskie Przedmieście 26/28, 00-927 Warszawa, Poland
[2] Institute of Polish Language, Polish Academy of Sciences,
Ratuszowa 11, 03-450 Warszawa, Poland
[3] Institute of Computer Science, Polish Academy of Sciences,
Jana Kazimierza 5, 01-248 Warszawa, Poland
[4] Austrian Centre for Digital Humanities, Austrian Academy of Sciences,
Sonnenfelsgasse 19, 1010 Wien, Austria
E-mail: j.bilinska@uw.edu.pl, renata.bronikowska@ijp.pan.pl,
zbigniew.gawlowicz@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl,
aleksandra.e.w@gmail.com, mateusz.zoltak@oeaw.ac.at

Our aim is to present the integration of the Electronic Corpus of the 17th and 18th c. Polish Texts, also referred to as The Baroque Corpus (Pl. korpus barokowy – hence the acronym KorBa), with the Electronic Dictionary of the 17th—18th C. Polish (henceforth e-SXVII). KorBa is a richly annotated corpus of Polish texts from the 17th and 18th centuries containing around 13.5 million automatically annotated and around 0.5 million manually annotated segments. Texts in the corpus are available in two orthographic layers: transliteration and transcription. The morphosyntactic annotation layer for each segment includes the basic form (lemma), the grammatical class and the relevant grammar category values. The corpus uses the MTAS software and supports the CQL query language.

The e-SXVII, launched in 2005, was among the first dictionaries created and provided entirely as a web application. It is a medium-size historical dictionary (currently having around 38,000 entries) which uses a dedicated software written in PHP and stores data in a relational database.

Both the dictionary and the corpus benefit from the integration. The corpus users can now get in-depth information about the displayed segments (their meanings, etymology, etc.). The users of the dictionary have gained easy access to a much wider set of usage examples, including advanced queries like searching for a given inflection form. The dictionary editors have also profited. First, the integration has allowed them to identify missing dictionary entries (i.e. the ones existing in the corpus but missing in the dictionary). Second, finding usage examples for the existing entries has been simplified. Third, a cross-check on finding a complete set of inflection forms has become possible. Finally, searching for phrasal verbs, proverbs, etc. (all present in the dictionary entries) has been made easier.

The integration required several kinds of actions. Firstly, the source texts and inflection forms identifiers had to be synchronized. Secondly, a corresponding REST API had to be developed on the dictionary side and an already existing one had to be adapted on the corpus side. Thirdly, user interfaces had to be adjusted. As a side-effect of the integration, the e-SXVII gained a REST API, making it machine searchable and readable.

Large differences between the lemmatization used by the corpus and the base forms noted by the dictionary turned out to be the biggest difficulty. Two solutions were implemented to mitigate the problem. Whenever it was possible to derive a transformation rule valid for a broad class of dictionary entries, the rule has been implemented (e.g. gerunds lemmatized to verbs in the corpus and to nouns in the dictionary). In other cases, the dictionary entry is processed on-the-fly by the morphological analyser which has been used to annotate the corpus. This is a fallback solution because it generates over-complicated corpus queries which can be difficult to understand and to adjust by the user.

Keywords: text corpus; electronic dictionary; historical electronic lexicography; integration of electronic lexical resources

Identification of Cross-disciplinary Spanish Academic Collocations for a Lexical Tool

Margarita Alonso Ramos, Marcos García Salido,

Marcos Garcia, Eleonora Guzzi

Universidade da Coruña, CITIC, Grupo LyS,
Dpto. de Letras, Fac. de Filoloxía. 15071, A Coruña
E-mail: {margarita.alonso, marcos.garcias, marcos.garcia.gonzalez}@udc.ga,
eleonora.guzzi@udc.es

We aim to build a lexical tool that helps novice writers in their academic writing in Spanish (Alonso-Ramos et al., 2017). Although most academic texts at Spanish universities are written in Spanish and Spanish is the mother tongue of the majority of students, the latter does not guarantee a good writing performance in academic discourse. Academic writing has to be learnt, since there is no native speaker of this genre. In fact, the academic writing of university students often shows certain deficiencies, many of which come from a poor knowledge of collocations. The proposed lexical tool offers suggestions of Spanish cross-disciplinary collocations, in order to help university students by improving the quality of their academic lexicon (for further details, see also García-Salido et al., 2018).

We focus on the proper method to identify cross-disciplinary collocations in a Spanish academic corpus consisting of research articles. Even though there are important lexical differences in different domains (Hyland & Tse, 2007), our project follows the approach according to which specialized texts contain, besides general lexicon (Drouin, 2007; Jacques & Tutin 2018: 1) domain-specific lexicon (or terminology) and 2) cross-disciplinary lexicon (or academic lexicon), which is in line with several works on academic English (Coxhead, 2000, Ackermann & Chen 2013; Gardner & Davies, 2014, Frankenberg-Garcia et al., 2018). However, the distinction between both kinds of lexicon is not clear-cut, especially when we deal with collocations. It is not enough to verify that the two elements of collocations are sufficiently represented in different domains of the academic corpus separately, but also the collocation as a whole. For instance, the noun *actividad* ‘activity’ and the verb *presentar* ‘to present’ have been selected for their specificity in the academic corpus, but the collocation *presentar actividad* is only specific to the domain of Natural Sciences.

We will describe the process of extraction of collocations from our academic corpus and the process of manual filtering that we employed until now. Firstly, we extracted a list of academic word candidates based on their specificity and on dispersion across all the domains. Secondly, we parsed our academic corpus to build a list of word combinations

using syntactic dependencies. From the 418 collocation candidates corresponding to 38 bases, we manually filtered those which were proper collocations. Out of these, only 113 collocations (from 25 bases) that were considered cross-disciplinary have been selected. The other 305 collocation candidates have been discarded mainly because they were considered free phrases or terminological. In order to improve the efficiency of this manual filtering, we compiled a bigger domain-specific corpus using WebBootCat (Baroni et al., 2006) with four main domains and 12 subdomains. After some experiments, we applied the Inverse Document Frequency model, a dispersion measure, to verify if a collocation is significantly more frequent in a given subdomain. If so, it will not be considered cross-disciplinary. We will present the results of these experiments, as well as the current state of the collocational tool.

Keywords: cross-disciplinary; academic vocabulary; collocations

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), pp. 235-247.

Alonso-Ramos, M., García Salido, M. & Garcia, M. (2017). Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations. In I. Kosem et al. (eds) *Proceedings of 2017 eLex Conference*, Leiden. pp. 571-586.

Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *Proceedings of EAMT 2006*. Oslo, pp. 247-252.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), pp. 213-238.

Drouin, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, *12*(2), pp. 45-64.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P. & Sharma, N. (2018). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), pp. 1-17.

García-Salido, M., Villayandre, M. & Alonso-Ramos, M. (2018). A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), pp. 260-265.

Gardner, D, & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, *35*(3), pp. 305-327.

Hyland, K. & Tse's, P. (2007). Is there an "academic vocabulary"?. *TESOL quarterly*, *41*(2), pp. 235-253.

Jacques, M. P. & Tutin, A. (2018). *Lexique transversal et formules discursives des sciences humaines.* London: ISTE Editions.

Lexicography and the Semantic Web: A Demo with LexO

Andrea Bellandi, Fahad Khan

Institute for Computational Linguistics, Via G. Moruzzi, Pisa - Italy
E-mail: name.surname@ilc.cnr.it

1. Introduction

The purpose of this contribution is to present LexO[8], the first version of a collaborative web editor for easily building and managing of lexical and terminological resources in the context of the Semantic Web. The adoption of Semantic Web technologies and the Linked Data paradigm has been driven by the need to ensure the construction of resources that are interoperable and can be shared and reused by the scientific community. LexO's primary objective is to enable terminologists and lexicographers to create a resource ex novo this is by means of the adoption of a lexical model that allows the association of detailed and structured lexical information (Bellandi et al., 2018); (Khan et al., 2016) to ontological concepts. In this respect, the lemon lexical model (McCrae et al., 2012), later renamed OntoLex-lemon (McCrae et al., 2017), is currently regarded as the de facto standard for enriching Semantic Web ontologies with lexical information. LexO can provide a support for creating, managing, publishing lexical and terminological resources as Linked Open Data, that is typically a complex task, especially for those who have not yet mastered Semantic Web-based standards and technologies, such as RDF and OWL. However, the long-term ambition of LexO would be to make a deeper contribution to e-lexicography.

2. LexO

LexO arises out of several DH research projects[9] that aimed to construct lexical resources and terminologies, and also takes into account experiences coming from international projects dealing with e-lexicography, such as ELEXIS[10]. It tries to

[8] The source code is available at https://github.com/cnr-ilc/LexO-lite.

[9] LexO has been mainly developed in the context of the DiTMAO project - multilingual resource of medico-botanical terminology focussed on Old Occitan, and it has already been used in several projects, such as, i) Ferdinand De Saussure (http://www.ilc.cnr.it/en/content/saussure) - multilingual diachronic resource of Saussure's terminology, ii) Totus Mundus - bilingual chinese-italian resource on Matteo Ricci's Atlas (http://www.ilc.cnr.it/en/content/ilc-iit-totus-mundus).

[10] ELEXIS (https://elex.is) is an ambitious project within the domains of NLP and e-lexicography with the aim of creating a European wide lexicographic infrastructure. It is based on a previous Cost Action ENeL - aiming to establish a European network of lexicographers and a common approach to e-lexicography that forms the basis for a new

consider the necessity of making the main formats (TEI, LMF (Francopoulo et al., 2006), OntoLex-lemon) interoperable, and in particular looking at a broader perspective of an ecosystem, where different standards can coexist and mutually enrich each other. In the following, we list the main features of LexO, also in the light of the results presented in the "User Needs" task[11] of the ELEXIS project: i) it hides all the technical complexities related to markup languages, language formalities and other technology issues, facilitating access to the Semantic Web technologies to non-expert users; ii) it allows for a team of users, each one with his/her own role (lexicographers, domain experts, scholars, etc.) to work on the same resource collaboratively; iii) it adheres to international standards for representing lexica and ontologies in the Semantic Web (such as OntoLex-lemon and OWL), so that lexical resources can be shared easily or specific entities can be linked to existing datasets; iv) it provides a set of services implemented by means of RESTful Web Services that allow software agents to access to the resources managed by means of LexO.

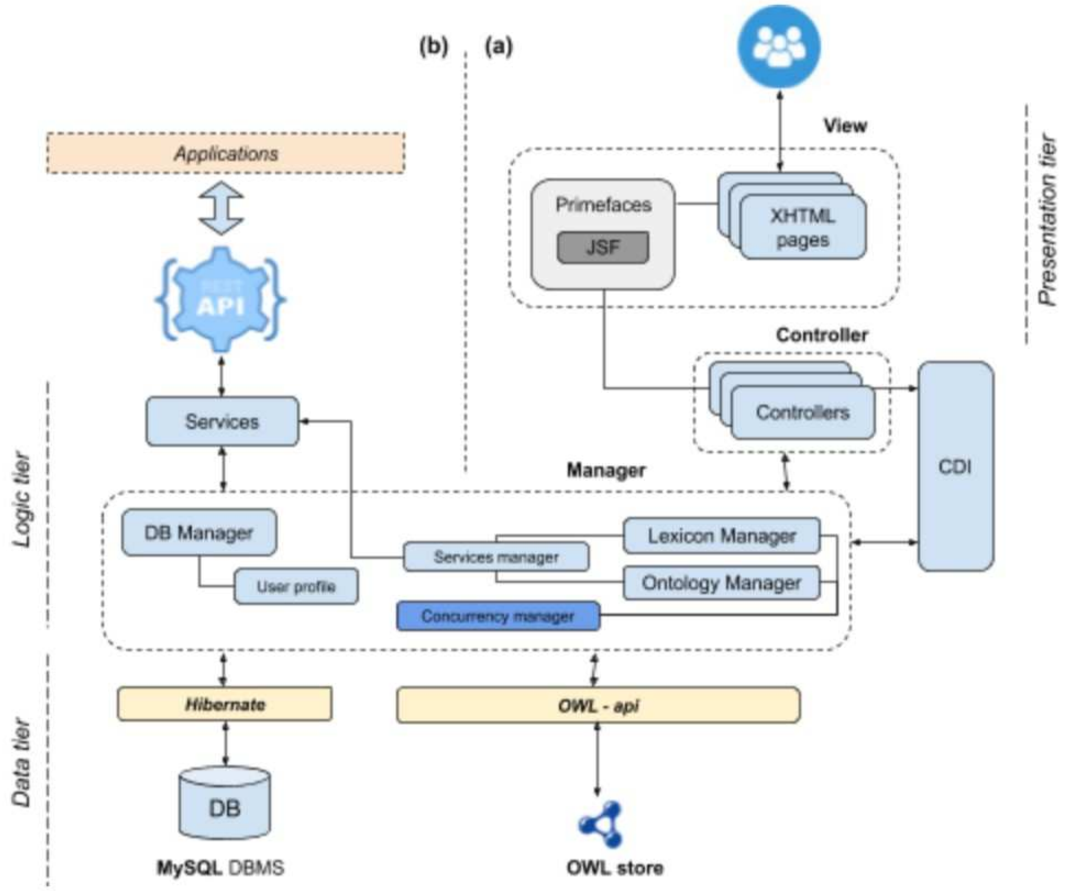
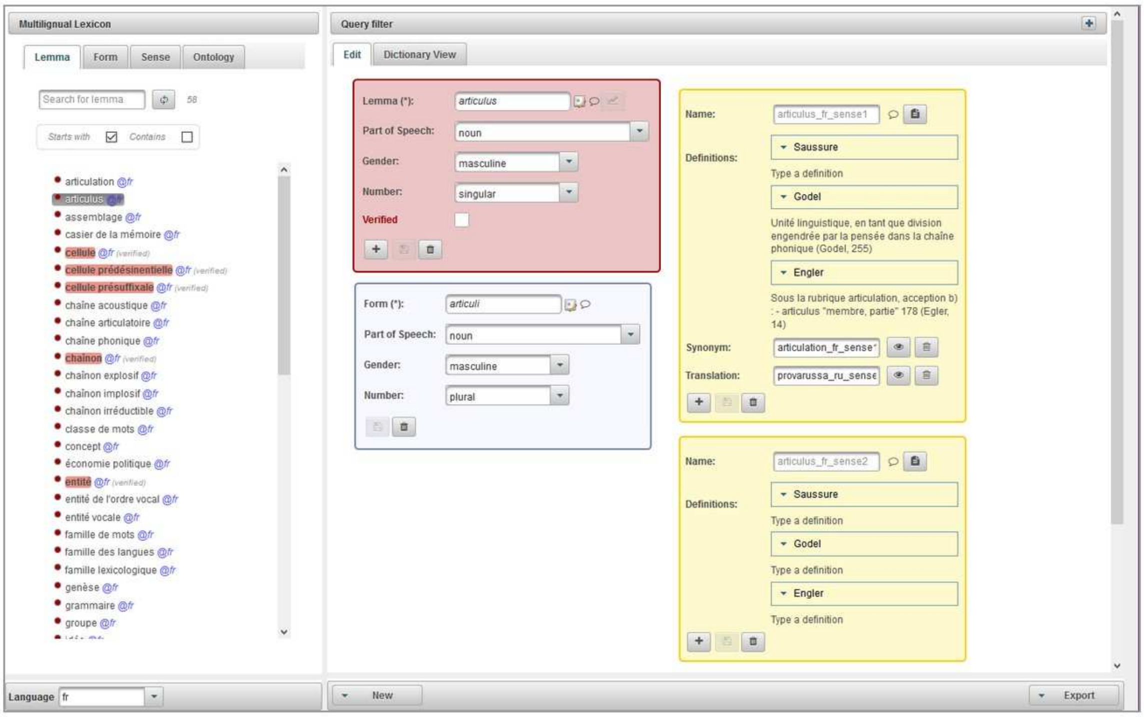


Figure 1: Multi-tier architecture of LexO. It is based on the software design pattern known as "three-tier architecture", and which uses Apache Tomcat v8.0 as the webserver. The component-based architectural structure was implemented by the object-oriented J2SE framework, enhanced with Contexts and Dependency Injection annotations (CDI). (a) LexO

type of lexicography (http://www.elexicography.eu/).

[11] See deliverable D1.1 "LEXICOGRAPHIC PRACTICES IN EUROPE: A SURVEY OF USER NEEDS" at https://elex.is/deliverables/

as web application - (b) LexO as rest services.



- articulation @fr
- articulus @fr
- assemblage @fr
- casier de la mémoire @fr
- cellule @fr (verified)
- cellule prédésinentielle @fr (verified)
- cellule présuffixale @fr (verified)
- chaîne acoustique @fr
- chaîne articulatoire @fr
- chaîne phonique @fr
- chaînon @fr (verified)
- chaînon explosif @fr
- chaînon implosif @fr
- chaînon irréductible @fr
- classe de mots @fr
- concept @fr
- économie politique @fr
- entité @fr (verified)
- entité de l'ordre vocal @fr
- entité vocale @fr
- famille de mots @fr
- famille des langues @fr
- famille lexicologique @fr
- genèse @fr
- grammaire @fr
- groupe @fr

Lemma (*): articulus
Part of Speech: noun
Gender: masculine
Number: singular
Verified □

Form (*): articuli
Part of Speech: noun
Gender: masculine
Number: plural

Name: articulus_fr_sense1
Definitions:
Saussure
Type a definition
Godel
Unité linguistique, en tant que division engendrée par la pensée dans la chaîne phonique (Godel, 255)
Engler
Sous la rubrique articulation, acception b) : - articulus "membre, partie" 178 (Egler, 14)
Synonym: articulation_fr_sense
Translation: provarussa_ru_sense

Name: articulus_fr_sense2
Definitions:
Saussure
Type a definition
Godel
Type a definition
Engler
Type a definition

Figure 2: Main interface of LexO. On the left, a column shows the list of lemmas composing the resource, the forms, the lexical senses, and the concepts belonging to the ontology of reference. The information related to the selected entry is shown in the central panel. It is possible to link senses to the related ontological concept.

The demo session will be organized in two parts. First, we will introduce LexO and how it falls in the framework of e-lexicography, its architecture, the data model, and how the tool can help users on the basis of their needs. Finally, we will start with the demo software by means of different real use cases.

Keywords: Semantic web; Lexicography; LexO; OntoLex-Lemon; Linked data

3. Acknowledgements

This work has been conducted in the context of the cooperation agreement between Guido Mensching, director of the DiTMAO project at the Seminar für Romanische Philologie of the Georg-August-Universität Göttingen, and the Istituto di Linguistica Computazionale “A. Zampolli” of the Italian National Research Council (29 August 2016).

The authors are also supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure).

4. References

Bellandi, A., Giovannetti, E. & Weingart, A. (2018). Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information*, 9(3), p. 52.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). *Lexical markup framework* (LMF).

Khan, F., Bellandi, A. & Monachini, M. (2016). Tools and Instruments for Building and Querying Diachronic Computational Lexica. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pp. 164– 171.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gömez-Përez, A., G.J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging Lexical Resources on the Semantic web. In *Proceedings of the Language Resources and Evaluation (LREC) 46(4)*, pp. 701–719.

McCrae, J., P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem et al. (eds.) *Proceedings of eLex 2017 conference,* pp. 19–21.

The WBÖ Lexicographic Editor – A tool for TEI-born Articles

Philipp Stöckle, Ludwig M. Breuer

Austrian Academy of Sciences, Austrian Centre for Digital Humanities (ACDH)

The Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich 'WBÖ') is a long-term project, whose main goal is the comprehensive lexicographic documentation of the manifold Bavarian base dialects in Austria and South Tyrol. The data used for writing the dictionary articles were recorded by so-called "Sammler" (gatherers) on paper-slips and collected in the "Hauptkatalog" (main catalogue). With the main purpose of facilitating and accelerating the process of writing dictionary articles, the hand-written paper slips were entered manually into a TUSTEP system and, subsequently, converted into TEI-XML (cf. Bowers & Stöckle, 2018). Since 2016, the WBÖ is (re)located at the department "Variation and Change of German in Austria" at the Austrian Centre for Digital Humanities (ACDH) at the Austrian Academy for Sciences (ÖAW), where a new staff of researchers has been working on a modernized concept.

Besides creating a web-based research platform where the database shall be made accessible for researchers, the focus lies on continuing the lexicographic work and writing dictionary articles. Up to now, the dictionary articles were written by using common word processing software (like Microsoft Word). While using Word to write texts seems the easiest and most straightforward way, it holds some significant disadvantages, especially since the articles will be published in different formats, i.e. print and online.

In order to provide a standardized structure, the articles will be composed directly in TEI. For this purpose, a new lexicographic editor system has been created, which will be the focus of our presentation. The editor system provides a clearly arranged and easy-to-use user interface for the lexicographers and, at the same time, uses XML-TEI structure to store the articles. Since the database uses the same structure, the editor tool communicates with the database, which facilitates and accelerates the lexicographic work, e.g. by copying entries directly into the lexicographic editor. It also allows for direct linking between information in the articles (e.g. with respect to a certain sense) and the database entries which were used as source. This way, users can have access to the raw data which contain more detailed information (e.g. with respect to the exact geographic locations of the entries) than the dictionary articles. Moreover, the standardized TEI structure of the articles makes it possible to easily generate HTML data for internet publishing as well as for printing.

The tool has been developed to meet the specific requirements of the WBÖ, but as it has an open source code, so it can be adapted by other lexicographic projects.

Keywords: lexicographic editor; TEI; open source

References

Bowers, J. T. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In: A. U. Frank., C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (eds.) *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*. Wien: Gerastree proceedings, pp. 45-54.

It Takes a CROWD to Raise Awareness: Expert Patients as Co-creators of Lexicographic Resources

Sara Carvalho[1] [2], Rute Costa[1], Raquel Silva[1]

[1] NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa,
Avenida de Berna 26-C, 1069-061 Lisboa – Portugal;
[2] CLLC, Centro de Línguas, Literaturas e Culturas da Universidade de Aveiro, Campus
Universitário de Santiago, 3810-193 Aveiro – Portugal
E-mail: sara.carvalho@ua.pt, rute.costa@fcsh.unl.pt, raq.silva@fcsh.unl.pt

This article aims to explore the synergies resulting from the conjugation of crowdsourcing, lexicography, and patient engagement by proposing the creation of EndoCrowdLex, an online, corpus-based lexicographic resource related to endometriosis[12], in which patients will be directly involved as co-creators in a multidisciplinary team comprising lexicographers, terminologists and subject field experts.

The unprecedented technological innovation that has characterized the last few decades, as well as the exponential increase of available data, have fostered the development of online platforms that have enabled a growing 'participatory culture' among citizens worldwide. In this respect, crowdsourcing (Howe, 2006) and its potential in the co-creation of products and services have been widely acknowledged across different sectors, such as healthcare (Wazny, 2017; Créquit et al., 2018) and lexicography (Kosem et al., 2013; Fišer et al., 2014; Čibej et al., 2015). Furthermore, health literacy (Nutbeam, 2000) and the increasingly active – and empowered – role of patients within the healthcare setting have been at the forefront of most eHealth[13] initiatives. Patient education has recently taken one step further with the advent of the expert patient, i.e. a patient with a medical condition whose acquired knowledge[14] and experience regarding that condition enable him/her to play an active role in its management (Carvalho, 2018). Given their know-how and status within the patient community, such expert patients could constitute valuable 'crowds' to work with from a lexicographic standpoint. Moreover, and despite its worldwide prevalence and significant economic and social impact (Kolterman et al., 2017), endometriosis-related resources are practically non-existent in the current lexicographic landscape.

[12] A chronic gynecological disease, which arises when tissue similar to the endometrium (i.e. the lining of the uterus) is found outside the uterine cavity (Dunselman et al., 2014).

[13] Defined as "the use of information and communication technologies (ICT) for health." (https://www.who.int/ehealth/en/).

[14] Often via formal and tailored education.

Therefore, this article focuses on the importance of these expert patients in the creation of EndoCrowdLex. On the one hand, they constitute an invaluable source of input concerning the actual needs of the endometriosis patient community and their families, as well as of other potential end users, which could help fine-tune both content and medium. They may also play a crucial role in the testing of the resource's main features throughout the development stages. On the other hand, and due to their unique knowledge of this condition, it is believed that after some training, this 'crowd' could actively participate in a set of microtasks within EndoCrowdLex, thereby facilitating the work of both terminologists and lexicographers. One scenario could include analysis of corpus examples so as to extract and categorize candidate terms and relations, with the latter categorization being supported by a conceptual framework based on the top-level concepts and relations put forward by the UMLS[15] and SNOMED CT[16]. In addition, these patients could be involved in the drafting of popularized natural language definitions based on previously created – and validated – scientific ones (Costa et al., 2019). Although expert patients would likely be highly motivated by knowing they would be supporting health literacy and patient education concerning this underrepresented disease, gamification in microtask design might provide a further incentive.

In short, by combining the added value of crowdsourcing, lexicography, and expert patients within a multidisciplinary setting, EndoCrowdLex aims to contribute towards endometriosis awareness, as well as to current research in lexicography.

Keywords: corpus-based lexicographic resource; crowdsourcing; expert patient; eHealth; terminology

Acknowledgements

This research has been financed by Portuguese National Funding through the FCT – *Fundação para a Ciência e Tecnologia* as part of the project Centro de Linguística da Universidade NOVA de Lisboa - UID/LIN/03213/2019 and by the European Union's Horizon 2020 research and innovation program under grant agreement No. 731015 (ELEXIS).

References

Carvalho, S. (2018). A terminological approach to knowledge organization within the scope of endometriosis: the EndoTerm project. PhD thesis. Faculdade de Ciências Sociais e Humanas – Universidade NOVA de Lisboa/Communauté Université Grenoble Alpes. Available at: https://run.unl.pt/handle/10362/49745.

[15] Short form for the Unified Medical Language System. It was developed by the United States National Library of Medicine to support and enhance the retrieval of machine-readable biomedical information (NLM, 2009).

[16] A comprehensive and multilingual clinical healthcare terminology widely used for the registration of Electronic Health Records (IHTSDO, 2018).

Čibej, J., Fišer, D. & Kosem, I. (2015). The role of crowdsourcing in lexicography. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (eds.). *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies / Lexical Computing Ltd., pp. 70-83.

Costa, R., Silva, R., Ramos, M. & Carvalho, S. (2019). Science with and for society: the popularization of terminology within the domain of nutrition. In *Proceedings of 22nd Conference on Languages for Specific Purposes (LSP 2019) – Mediating Specialized Knowledge: Challenges and Oportunities for LSP Communication, Translation and Research*. Padua, 10-12 July 2019 (forthcoming).

Créquit, P., Mansouri, G., Benchoufi, M., Vivot, A. & Ravaud, P. (2018). Mapping of Crowdsourcing in Health: Systematic Review. *Journal of Medical Internet Research*, 20(5), e187. doi:10.2196/jmir.9330.

Dunselman, G., Vermeulen, N., Becker, C., Calhaz-Jorge, C. & D'Hooghe, T. (2014). ESHRE guideline: management of women with endometriosis. *Human Reproduction*, 29(3), pp. 400–412.

Fišer, D., Tavčar, A. & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. In *Proceedings of LREC 2014*, pp. 4371–4375.

Howe, J. (2006). The Rise of Crowdsourcing. Wired, 14. Retrieved on February 1, 2019 from: http://www.wired.com/wired/archive/14.06/crowds.html.

IHTSDO. (2018). SNOMED CT Starter Guide. International Health Terminology Standards Development Organisation. Retrieved on February 1, 2019 from https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide.

Koltermann, K. C., Dornquast, C., Ebert, A. D. & Reinhold, T. (2017). Economic Burden of Endometriosis: A Systematic Review. *Annals of Reproductive Medicine and Treatment*, 2(2), 1015. Retrieved on February 1, 2019 from https://pdfs.semanticscholar.org/e5a9/e0134a372beb0620e742b7ec12f85751dc2b.pdf.

Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 Conference, 17-19 October 2013 Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies & Eesti Keele Instituut, pp. 32-48.

NLM (2009). *UMLS® Reference Manual.* Bethesda: National Library of Medicine.

Nutbeam, D. (2000). Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century, *Health Promotion International*, 15(3), September 2000, pp. 259–267, Retrieved on February 1, 2019 from https://doi.org/10.1093/heapro/15.3.259.

Wazny K. (2017). "Crowdsourcing" ten years in: A review. *Journal of Global Health*, 7(2), 020602.

Modelling and Analysis of Dialect Data in the Database of the Southern Dutch Dialects (DSDD)

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

[1] Ghent University, Ghent, Belgium
[2] Instituut voor de Nederlandse Taal (Dutch Language Institute), Leiden, The Netherlands
E-mail: Veronique.DeTier@UGent.be, jesse.dedoes@ivdnt.org, katrien.depuydt@ivdnt.org, tanneke.schoonheim@ivdnt.org, sally.chambers@ugent.be, Jacques.Vankeymeulen@ugent.be

The project *Database of the Southern Dutch Dialects (DSDD)* aims to aggregate and standardize three existing comprehensive dialect lexicographic databases of the Flemish, Brabantic and Limburgian dialects into one integrated dataset. The project, which runs from 2017 to 2020, will result in a harmonized dataset, an API for researchers and a portal application in which the outcome of the project will be available to a broad audience.

Although the dictionaries were explicitly set up in parallel in order to make future aggregation possible, there are differences in methodology and they have not used the same set of concepts. There is also a significant degree of heterogeneity from a technical point of view, in terms of file formats and logical structure.

The project has set up a workflow to harmonize the data structures and to interconnect the dictionaries by adding an overarching layer of concepts. The integrated database aims to enable innovative research, especially in the field of quantitative lexicology and dialect-geographical analysis.

In this contribution, we focus on two aspects:

- The development of a data model for the integrated database, both from a technical and a linguistic point of view. The three dictionaries cover similar but not necessarily equivalent concepts; moreover, the standard language term chosen to represent a concept may differ between dictionaries. We are faced with the technical problem of grouping similar concepts without an extensive manual search, and a methodological problem of dealing with closely related but non-equivalent concepts. We evaluate graph-based concept linking methods to deal with the technical issue and propose a data model to incorporate the lexico-semantic relations, in order to do justice to the complexities inherent in the data.

- The exploitation of the integrated database for computational dialectology,

spatial statistical analysis and visualization of the dialect data in an interactive research environment. The aim is to develop a research environment which is both feature-rich and flexible enough to allow researchers to answer their research questions based on this extremely valuable but not altogether unproblematic data set. Prototypes of the API and the portal application will be demonstrated.

Keywords: Southern Dutch Dialects; dialectology; computational dialectology; dialect maps; integration

References

Heeringa, W. & Prokić, J. (2018). Computational Dialectology. *The Handbook of Dialectology*, Wiley Blackwell, 2018, pp. 330-347.

Van Keymeulen, J. et. Al. (2018). "Sustaining the Dictionary of Southern Dutch Dialects (DSDD): a case study for CLARIN and DARIAH." In *CLARIN Annual Conference 2018*, p. 128.

Van Keymeulen J. & Devos, M. (2007). Lexicale dialectometrie op basis van het Woordenboek van de Vlaamse dialecten. In M. Devos & R. van Hout (eds.) *Taal en Tongval, Themanummer 20: Dialectlexicografie*, pp. 9-33.

Van Keymeulen, J. & De Tier, V. (2010). Towards the completion of the Dictionary of the Flemish Dialects. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress.* Fryske Akademy, Leeuwarden, pp. 764-773. (issued on CD-ROM).

McCrae, J. & Buitelaar, P. (2018), Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp 109-123.

Perozzi, B. et. al (2014). Deepwalk: Online learning of social representations. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2014.

Rabanus, S. (2018). Dialect maps. In*The Handbook of Dialectology*, Wiley Blackwell, pp. 348-367.

Rothe, S. & Schütze, H. (2014). Cosimrank: A flexible & efficient graph-theoretic similarity measure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers).

De Tier, V. & Van Keymeulen, J. (2010). Software Demonstration of the Dictionary of the Flemish Dialects and the pilot project Dictionary of the Dutch Dialects. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress. Fryske Akademy*, Leeuwarden, pp. 620-627 (issued on CD-ROM).

Towards a More Efficient Workflow for the Lexical Description of the Dutch Language

Katrien Depuydt, Tanneke Schoonheim, Jesse de Does

Instituut voor de Nederlandse Taal (Dutch Language Institute), Leiden, The Netherlands
E-mail: katrien.depuydt@ivdnt.org, tanneke.schoonheim@ivdnt.org, jesse.dedoes@ivdnt.org

The linguistic description of the Dutch Language, both historical and contemporary, has been a task of the Dutch Language Institute (Instituut voor de Nederlandse Taal) since its foundation in 1967[17]. This has in the course of time resulted in corpora, computational lexica and scholarly dictionaries of historical and contemporary Dutch.

Around 2005, we found ourselves working on a range of historical and contemporary lexicographical projects, without investing in the synergy between them. This was inefficient, and since then efforts have been made to streamline the corpus-based description of the Dutch language.

First, the development of an integrated lexical database was initiated, the GiGaNT lexicon[18]. The aim of this lexicon is to get a well-structured overview of the described Dutch vocabulary, and to allow the systematic screening of the Dutch word stock for omissions in the lexicographic description for both historical and contemporary Dutch. The core of the historical component is the material from the Dutch historical scholarly dictionaries. The contemporary component was primarily developed for the 2015 release of the Dutch spelling reference lexicon and has been expanded ever since. A recent step in data integration of the contemporary projects was the linking of ANW[19] and neologism entries to the GiGaNT lexicon.

With regard to the corpora, efforts were made to integrate the several internal and external corpora for Dutch into one basic corpus, the Corpus Hedendaags Nederlands (CHN, Corpus of Contemporary Dutch). The external part of this corpus has been made available in the CLARIN infrastructure (chn.inl.nl). A common metadata scheme was developed and the structural encoding was done according to a basic encoding scheme in TEI, developed for all INT corpora, both historical and contemporary.

To ensure the corpus is up-to-date, a corpus workflow tool was developed (“duct”) which implements continuous integration in the context of corpus processing, including all steps from fetching files from an FTP server, via basic conversion to TEI to linguistic

[17] Up to 2016 the *Instituut voor Nederlandse Lexicologie* (INL, Institute for Dutch Lexicology).

[18] GiGaNT, *Groot Geïntegreerd lexicon van de Nederlandse Taal* (Large Integrated Lexicon of the Dutch Language).

[19] ANW, *Algemeen Nederlands Woordenboek* (Dictionary of Contemporary Dutch)

annotation and indexing the material in the BlackLab corpus retrieval system.

Finally, the workflow for the description of contemporary Dutch has been improved. New words (not necessarily always neologisms) are automatically detected in the new corpus material and forwarded to the lexicographical environment for selection. Selected new words are first added to the central database, where information on part of speech is added, as well as a complete paradigm including hyphenation. This means that for all projects this basic information is equally and easily available. In the future, morphological and etymological components are also foreseen, as well as phonetic transcriptions. It is the intention to link all lexicographical products of the INT to this central database so as to share the systematic linguistic description of Dutch in all (past, present and future) projects.

Keywords: corpus based lexicography; modern Dutch; workflow

References

Borin, L. et al. (2012), "The open lexical infrastructure of Språkbanken". In: Nicoletta Calzolari et al. (eds), *Proceedings of the 8th International Conference on Language Resources and Evaluation: May 23-25*, 2012. Istanbul.

Does, J. de & Depuydt, K. A. C (2009), "Computational Tools and Lexica to Improve Access to Text". In E. Beijk & L. Colman (eds.) *Fons Verborum. Feestbundel voor prof. Dr. A.M.F.J. (Fons) Moerdijk, aangeboden door vrienden en collega's bij zijn afscheid van het Instituut voor Nederlandse Lexicologie.* Leiden/Amsterdam, pp. 187-199.

Depuydt, K. & Dutilh-Ruitenberg, M. W. F. (2002) TEI-encoding for the Integrated Language Database of 8th-21st-Century Dutch. In A. Braasch & C. Povlsen (eds.) *Proceedings of the Tenth Euralex International Congress, Euralex 2002*, pp. 683-688.

Depuydt, K. A. C & Does, J. de (2018), The Diachronic Semantic Lexicon of Dutch as Linked Open Data. In I. Kernerman & S. Krek, *Proceedings of the LREC 2018 Workshop "Globalex 2018 –Lexicography & WordNets"*. [Miyazaki], 2018, pp. 23-28.

Ruitenberg, M. W. F et al. (2010) "Developing GiGaNT, a lexical infrastructure covering 16 centuries In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*, Leeuwarden: Fryske Akademy – Afûk 2010.

Schoonheim, T. & Tempelaars, R. (2010). Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW). *Proceedings of the XIV Euralex International Congress*, Ljouwert, Fryske Akademy/Afûk, abstract.

Tiberius, C. & Schoonheim, T. (2016). The Algemeen Nederlands Woordenboek (ANW) and its lexicographical process. *Online publizierte Arbeiten zur Linguistik* (2016): 20.

A Web of Loans: Multilingual Loanword Lexicography with Property Graphs

Peter Meyer, Mirjam Eppinger

Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: meyer@ids-mannheim.de, eppinger@ids-mannheim.de

Introduction

The Lehnwortportal Deutsch (2012 seqq.) serves as an integrated online information system on German lexical borrowings into other languages, synthesizing an increasing number of lexicographical dictionaries and providing basic cross-resource search options. The paper discusses the far-reaching revision of the system's conceptual, lexicographical and technological underpinnings currently under way, focussing on their relevance for multilingual loanword lexicography.

Data structures

Roughly in the spirit of Měchura (2016), the new Lehnwortportal system combines traditional XML-based digital lexicography with graph-based interlinking of information. It is special in relying on a property graph database, using the Turing-complete Apache Tinkerpop Gremlin query language (Rodriguez, 2015). The overall architecture is best suited for manually curated resources and is to be distinguished from a more NLP-oriented RDF-based Linked Data approach as discussed, for example, in (Gracia, Kernerman & Bosque-Gil, 2017). The graph features three main types of nodes (lexical units, their senses, entries in lexicographical resources). It expresses relations between lexical units such as borrowing, derivation, diachronic development, and variation; in addition, it relates lexical units to their word senses and to the entries in the portal's resources they appear in. Distinct but interconnected subgraphs cleanly separate the data found in the individual original sources from the manually edited, homogenized and interlinked portal data layer that is supposed to represent genuinely linguistic information (Meyer & Eppinger, 2018).

Accessing and editing the graph

Apart from a traditional entry/XML-based presentation, the revised Lehnwortportal provides intuitive and powerful real-time access to the entire graph via an innovative visual query builder (Meyer, 2018) that allows users to find arbitrarily complex graph constellations (n-tuples of words). Amongst other things, the query system offers Boolean operators to describe alternative or non-existing paths (e.g. borrowing histories) as well as comparisons such as “is homonymous to” or “has same PoS as”. Users can easily navigate the curated graph by following edges between graph nodes visually or hypertextually. Authorized lexicographers may directly edit the subgraphs around search results.

Lexicographical process

The non-linear nature of graph editing and the requirement to keep the portal’s XML resources in sync with the graph creates the need for intertwined ‘graph-augmented’ lexicographical tools and practices. Amongst other things, bookkeeping of changes and additions to the data, including versioning, should itself be graph-driven, taking advantage of the graph query system and the possibility to add temporary editing attributes to search results. Changes in the underlying resources must percolate in a well-defined manner in the graph and lead to the automatic flagging of nodes and edges whose modifications might be lexicographically incorrect and must therefore be checked manually. Especially for the digitization/integration of new lexicographical data, the Lehnwortportal features a ‘graph-first’ input editor.

New resources

The revised Lehnwortportal, freely accessible from 2021, will host a number of new dictionaries of German loanwords in European languages, amongst others, English (Pfeffer & Cannon, 1994), French (Le Trésor de la Langue Française Informatisé) and Hungarian (Benkő & Büky, 1994). The paper includes a conceptual comparison to two related projects on Dutch (van der Sijs, 2015) and Italian (Heinz, 2017) loanwords in other languages.

Keywords: property graph; loanword lexicography; lexicographical process

References

Benkő, L. & Büky, B. (1993). *Etymologisches Wörterbuch des Ungarischen.* Budapest: Akad. Kiadó.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward Linked Data-Native Dictionaries. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (eds.) *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.* Brno: Lexical Computing, pp. 550-559. Available at: https://elex.link/elex2017/proceedings-download.

Heinz, M. (ed.) (2017). *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti. Atti dell’incontro OIM (Firenze, 20 giugno 2014).* Firenze: Accademia della Crusca.

Le Trésor de la Langue Française Informatisé. Accessed at: http://atilf.atilf.fr/

Lehnwortportal Deutsch. Ed. by Institut für Deutsche Sprache, Mannheim. Accessed at: http://lwp.ids-mannheim.de. (16 August 2019)

Měchura, M. (2016). Data structures in lexicography: from trees to graphs. In A. Horák, P. Rychlý & A. Rambousek (eds.) *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2016.* Brno: Tribun EU, pp. 97-104.

Meyer, P. (2019). Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein interaktiv-visueller Query Builder für Property-Graphen. In P. Sahle (ed.) *DHd 2019 Digital Humanities: multimedial & multimodal. Konferenzabstracts.* Frankfurt am Main, pp. 312-314. Available at: https://zenodo.org/record/2600812.

Meyer, P. & Eppinger, M. (2018). fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17-21 July, Ljubljana.* Ljubljana: Znanstvena založba, pp. 1017-1022.

Pfeffer, J. A. & Cannon, G. (1994). *German Loanwords in English. An Historical Dictionary.* Cambridge, New York, Melbourne: Cambridge University Press.

Rodriguez, M. A. (2015). The Gremlin Graph Traversal Machine and Language. In J. Cheney & Th. Neumann (eds.) *Proceedings of the 15th Symposium on Database Programming Languages (DBPL 2015).* New York: The Association for Computing Machinery, pp. 1-10.

van der Sijs, N. (2015). *Uitleenwoordenbank, hosted by the the Meertens Instituut.* Accessed at: http://www.meertens.knaw.nl/uitleenwoordenbank. (16 August 2019)

The New Digital Edition of the Dictionnaire de l'Académie Française

Laurent Catach

Académie française / 4H Conseil
E-mail: laurent.catach@academie-francaise.fr

This presentation deals with the new digital edition of the Dictionnaire de l'Académie française, the oldest and most respected dictionary of the French language, which was first published in 1694. This new digital portal has been available to access freely at www.dictionnaire-academie.fr since February 2019.

The first release of the portal includes the full 8th edition (1935), and the current 9th edition up to the letter S (the rest will be made available shortly). It should be emphasized that this 9th edition represents a major evolution in the long history of the Dictionnaire, as it is twice the size of the 8th edition, with 25,000 new words and new features such as etymology.

The next version of the portal, to be released by the end of 2019, will contain the full corpus of all nine editions of the Dictionnaire, giving unprecedent access to the full lexicographical work of the Académie française since its inception in 1635. In the end, the portal will give access to more than 250,000 entries from the nine editions.

In addition to from some functionalities which can now be found in most digital dictionaries (such as autocompletion, spelling correction, hypertext navigation, etc.), this new edition has some interesting and more original characteristics:

- the "word history" feature, giving the list of all entries of the same word in the nine editions, whatever spelling modifications have occurred;
- full conjugation of all the verbs in the 9th edition of the Dictionnaire;
- integration of lexical notes, published by the Académie, about difficulties or curiosities of the French language;
- linked lexical data pointing to the official website of the Délégation à la langue française, which contains a large database of recent terms and recommendations in every field of terminology;
- linked lexical data pointing to the BDLP, a large database describing in detail diatopic variations of the French language, in 20 countries (Quebec, Belgium, French-speaking countries in Africa, etc.);

- links to the digitalized versions (in image mode) of the past editions of the Dictionnaire, thanks to a partnership with the Bibliothèque nationale de France.

The new website is also fully built using responsive design, making the Dictionnaire a free and very simple-to-use resource available on every kind of smartphone, and similar to a mobile application.

The presentation will describe in detail the principles, the objectives and the functionalities of the portal. With 300 million French-speaking people all around the world, and expectations that this could grow to 700 million by 2070, this new portal aims to become a new point of reference in the digital sphere for the French language, and also a key resource for education. It is also intended to become the basis for new extensions, new content and new links to other external lexical data available on the web, showing the will of the Académie française to provide high-quality and evolving content to the French-speaking and French-learning communities. Future possibilities, such as linking to the semantic web and open data, will also be discussed.

Keywords: Académie française; dictionary content; French dictionary; digital dictionaries; education; Francophonie; Français langue étrangère (FLE); linked lexical data

Integrating the Etymological Dimension into the OntoLex-Lemon Model: A Case Study

Pascale Renders

University of Lille, Rue du Barreau BP 60149, 59650 Villeneuve d'Ascq
E-mail: pascale.renders@uliege.be

Nowadays, lexicographical resources increasingly integrate, partially or completely, the Semantic Web and, in particular, the Linguistic Linked Open Data network (http://linguistic-lod.org/llod-cloud). More recently, the Elexis project (European Lexicographic Infrastructure) set itself the goal of promoting cooperation among research teams, in particular by pushing the development of standards, methods and tools linking lexicographical resources with open access to their data (http://www.elex.is). The need for data linking is particularly strong in the historical lexicography of Romance languages, where numerous reference books complete the *Französisches Etymologisches Wörterbuch* (FEW), a thesaurus gathering all lexical units of Galloromance languages and dialects. These resources would benefit from a digital linking with the FEW by allowing their users to have direct access to the corrections they propose.

OntoLex-Lemon (http://www.w3.org/2016/05/ontolex/) is one of the standards made available to the community of lexicographers for the linking of digital resources. It has already been tested with several digitized dictionaries. This standard was not initially designed with the purpose of linking dictionaries, but rather to enrich the Semantic Web's existing ontologies with lexical information. This explains why the model does not yet allow the complexity of all the data from lexicographical resources to be represented. OntoLex-Lemon's applicability to lexicography quickly led to improvements, for instance through the addition of morphological and syntactical information. However, it is only at a later stage that the model started to take into account the etymological, and more generally the historical dimensions. Scholars first added the possibility to point out etymons when the dictionaries contained the information (see Declerck & Wandl-Vogt, 2015), then Khan proposed a more general extension to the model (Khan, 2018). Still, the historical dimension remains rather poorly represented in the Linked Open Data (see for example Tittel & Chiarcos 2018, who try to integrate the DEAF in this network).

Our presentation studies the possibility to integrate the etymological data of Romance linguistics into the Linked Open Data. We focus on the Galloromance field and analyse examples coming from various types of resources (dictionaries, thesauri and linguistic atlases, including the FEW). In this way, we show whether the model proposed by Khan 2018 enables the addition of the "etymology-history" dimension as commonly

encountered in Romance linguistics. Finally, we propose several ways in which the model can be adapted to fully integrate the addressed resources into the Semantic Web.

Keywords: OntoLex-Lemon; Romance lexicography; Etymology

References

DEAF: *Dictionnaire étymologique de l'ancien français.* (1974-). Tübingen: Niemeyer.

FEW: *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes.* (1922-2002). Bonn/Heidelberg/Leipzig-Berlin/Bâle: Klopp/Winter/Teubner/Zbinden.

Declerck, T. & Wandl-Vogt, E. (2015). Towards a Pan European Lexicography by Means of Linked Open Data. In Kosem, I. et al. (eds.) *Proceedings of the eLex 2015 conference.* Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd, pp. 342-355.

Khan, F. (2018). Towards the representation of Etymological Data on the Semantic Web. *Information* 9 (12), p. 304.

Tittel, S. & Chiarcos, C. (2018). Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the Dictionnaire étymologique de l'ancien français with OntoLex-Lemon. In Kernerman, I. & Krek, S. (eds.) *Proceedings of the LREC 2018 Globalex Workshop (Globalex 2018 - Lexicography and WordNets).* Miyazaki, Japan.

Representation of Collocability in Russian Dictionaries

Maria Khokhlova

St. Petersburg State University
E-mail: m.khokhlova@gmail.com

The rapid development of the internet and computer technologies has influenced various spheres of linguistics, but it can be said that lexicographers have profited most of all. Russian dictionaries have a long tradition, but when it comes to computational lexicography they are lagging behind. In the present study we address ourselves to the wide range of printed lexicographic sources that describe collocations.

Explanatory dictionaries occupy the most prominent part among other Russian reference books representing the results of fundamental work on describing lexis. One can find set phrases not only in special sections of the entries but also in the examples, sayings and quotations. For example, the diamond symbol is used to designate set expressions and phraseological units in some works (*Dictionary of Contemporary Literary Russian Language*, 1948-1965; *Dictionary of the Russian Language*, 1981–1984), while the latter is defined in the *Big Academic Dictionary of Russian* (2004–) with the tilde symbol ~.

The *Dictionary of Set Verb-Noun Phrases in Russian* (Deribas, 1983) comprises 5,197 collocations for 744 verbs and 1,345 nouns. The majority of phrases consist of bigrams including verbs and nouns as direct or indirect objects. The authors emphasize that the dictionary is focused on language learners and does not provide any definitions or explanations, merely listing the collocations of literary language.

The *Explanatory Combinatorial Dictionary of Modern Russian* (Mel'čuk & Zholkovsky, 1984) represents the result of a unique approach to the formal description of collocability. The authors characterize their dictionary as an active one, i.e. that can be used not to understand but to produce speech. The core notion of the dictionary is a lexical function which associates a word (or an argument) with a set of words and phrases expressing the meaning or role which correspond to the function.

The *Dictionary of Russian Collocations with English-Russian Dictionary of Keywords* proved to be the first Russian dictionary that used the notion of "collocation" (Borisova, 1995), structuring them according to their semantics with numbers that correspond to lexical functions.

The *Dictionary of the Collocability of the Words of the Russian Language* (Denisov, Morkovkin, 2002) is the most famous and comprehensive collocations dictionaries of Russian. The authors distinguish between lexical and semantic collocability (according to Yu. Apresyan), and also define syntactic collocability as a set of semantic and syntactic positions available for a word, in other words, its valency frame.

The *Active Dictionary of the Russian Language* (Apresyan, 2014-) includes extensive information on collocability, being the most ambitious project of the last few decades. Each zone is labelled separately in the dictionary entries, e.g. meaning, collocability, government model, synonyms, etc.

The printed lexicographical sources are often neglected because of the complications with their digitalization or obsolescence of their data, and some of them are not so well-known to a mass audience. As further development of the tools that represent collocability we can see crowdsourcing and user feedback as a way of improvement.

Keywords: collocations; dictionaries; entry structure; Russian

Acknowledgements

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. MK-2513.2018.6).

References

Apresyan, Ju. D. (ed.) (2014-). *Active Dictionary of the Russian Language* [Aktivnyy slovar’ russkogo yazyka]. Vol. 1-3. Moscow: Yazyki slavyanskoy kul’tury.

Big Academic Dictionary of Russian [Bolshoy akademicheskiy slovar v 30 tomakh]. (2004–2016). Moscow-St. Petersburg: Nauka.

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

Denisov, P. N. & Morkovkin, V. V. (2002). *Dictionary of the Collocability of the Words of the Russian Language* [Slovar' sochetaemosti slov russkogo iazyka]. Moscow: Astrel'.

Deribas V. M. (1983). *Verb-Noun Collocations in Russian* [Ustojchivyje glagol’no-imennyje slovosochetanija russkogo jazyka]. Moscow: Russian language.

Dictionary of Contemporary Literary Russian Language [Slovar sovremennogo russkogo literaturnogo yazyka v 17 tomakh], (1948–1965). Chernyshev, V.I. (ed.). Moscow-Leningrad: Izd-vo Akademii nauk SSSR.

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

Mel’čuk, I. & Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian* [Tolkovo-kombinatornyj slovar russkogo jazyka]. Vienna.

Linked Lexical Data of Different Resources and Countries in the Alpine Region: The Project VerbaAlpina as an Example of Good Practice in “Smart Lexicography”

Beatrice Colcuc, Christina Mutter

Ludwig-Maximilians-Universität München
E-mail: beatrice.colcuc@romanistik.uni-muenchen.de, christina.mutter@lmu.de

In the pre-digital world there were two lexicographic forms of publication, dictionaries and linguistic atlases, which looked at vocabulary from two mutually exclusive perspectives: the onomasiological and the semasiological. The research project VerbaAlpina (https://www.verba-alpina.gwi.uni-muenchen.de/), funded by the German Research Foundation (DFG) as a long-term project since 2014, investigates the linguistic and cultural area of the Alpine region on a transnational basis and aims to combine these two opposing perspectives with the possibilities of digitization in an innovative lexicographic online platform. Thus, VerbaAlpina makes lexicographic data digitally accessible and with its responsive surface the project’s web page can also be accessed on mobile devices such as smartphones. The database is primarily based on the traditional linguistic atlases and dictionaries of the past hundred years available in the Alpine region. This historical linguistic material is supplemented with current linguistic data by means of crowdsourcing via a crowdsourcing tool designed by VerbaAlpina.

One of the major challenges regarding the integration of linguistic data from different language resources consists in the lack of uniformity of the data from the individual data sources (linguistic atlases, dictionaries, crowdsourcing) which are not structured in the same way. VerbaAlpina must first unify the different transcription systems of the atlases and dictionaries. For this purpose, the already existing data in both digital and analogue forms undergo a process of systematic data processing to fit the unified structure of the relational database (MySQL) in which all project data is stored. This process can be subdivided into three major steps: the transcription of the linguistic data, their tokenization and the allocation of the single tokens to different lexicographic units (VerbaAlpina distinguishes between so-called base types, morpho-lexical and phonetic types) and concepts. The project’s web interface offers the possibility to visualize these data on an interactive map. The linguistic data can be displayed in both directions, onomasiologically and semasiologically, by using appropriate filters. In addition to this qualitative representation of the linguistic material, the portal also offers the possibility of a quantifying representation of the data in the sense of an aggregation within certain geographical regions.

In order to link this unified data with other (language) resources beyond the project boundaries, the data must be unambiguously marked to enable targeted referencing to individual data. This is achieved by assigning identifiers such as DOIs (Digital Object Identifiers), the Q- and L-IDs used by Wikidata, and GNDs (Gemeinsame Normdatei 'Integrated Authority File'). While QIDs provide a reference to concepts and LIDs and GNDs one to morpho-lexical types, DOIs can be used to identify universal content.

This talk is intended to take a closer look at the processes of data processing, the associated challenges and the approaches VerbaAlpina is taking to unify lexical data from different sources, to make them linkable in an interdisciplinary way and to visualize them.

Keywords: digitization; crowdsourcing; linked lexical data

Frames and the Digital Identity of Online Dictionaries: A Comparative Analysis Between Conventional Online Dictionaries of Portuguese and the Frame-based Online Dictionary Dicionário Olímpico

Bruna da Silva, Rove Chishman

Unisinos University, São Leopoldo, Brazil
E-mail: broonamoraes@gmail.com, rove@unisinos.br

This study, inserted in the interface between Frame Semantics (Fillmore, 1982, 1985) and Electronic Lexicography, aims at investigating the contribution of the notion of frames to the development of online dictionaries, based on (i) a metalexicographic analysis of Dicionário Olímpico (Chishman, 2016) and a set of conventional online dictionaries and (ii) a comparative analysis between Dicionário Olímpico and the other dictionaries, based on the theoretical framework of Electronic Lexicography, traditional Lexicography and Cognitive Semantics. Therefore, we turn our attention to the bibliography of Frames Semantics, in order to highlight the precepts underlying the notion of frames and the theoretical potential of this concept for the description of lexical meaning, and then the applied theory bias in relation to Lexicography (Fillmore, 2003; Fillmore & Atkins, 1992).

However, we propose an extension of the original interface, as we turn to the literature of Electronic Lexicography. In this sense, we offer a panorama related to the area dedicated to the development of digital dictionaries, in order to deal with challenges, expectations and demands that are outlined from the historical trajectory of the field and its current configuration. The methodology can be divided into four parts that correspond to: (i) the selection of the set of conventional online dictionaries to be analysed; (ii) the developments related to the metalexicographic analysis; (iii) the elaboration of metalexicographic analysis forms, concerning the composition and nature of online dictionaries and the research on the history of Electronic Lexicography; and (iv) the description of the analysis strategies, including the procedures for completing the forms, the preparation of observation sheets and the assembly of dictionary rankings.

The analysis and discussion of the data were based on the information obtained through the analysis of the materials and was intended to serve the discussion about the way in which the dictionaries show the online digital identity and about the role of the notion of frames in the construction of the online digital identity from Dicionário Olímpico. The results showed that: the lack of clarity about ways to assume the online digital identity efficiently is the main obstacle to the development of online dictionaries;

the online digital identity presents itself at different levels in the tools analysed, and is related to the fact that the dictionaries have or do not have a printed counterpart; and the notion of frames played the role of providing the guidelines for an efficient presentation of the digital elements in the Dicionário Olímpico. Lastly, we find that the notion of frames contributes to the development of online dictionaries insofar as the notions of "encyclopaedic knowledge", "empiricism" and "continuities between language and experience", which underlie the notion of frames, are responsible for the fact that, in a dictionary based on frames, the digital elements take on more central functions in the construction of meaning.

Keywords: Frame semantics; electronic lexicography; digital identity; digital elements

Acknowledgements

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Finance Code 001.

References

Chishman et al. (2016). Dicionário Olímpico. São Leopoldo: Unisinos.

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

Fillmore, C. J. (1985). Frames and the Semantics of Understanding. *Quaderni di Semantica*, 6(2), pp. 222-254.

Fillmore, C. J. (2003). Double-Decker Definitions: The Role of Frames in Meaning Explanations. *Sign Language Studies*, 3(3), pp. 263-295.

Fillmore, C. J. & Atkins, S. (1992). Toward a Frame-based Lexicon: The Semantics of RISK and its Neighbors. In A. Lehrer & E. Kittay (eds.) *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*. Hillsdale: Erlbaum, pp. 75-102.

Exploration of the Evolution of Lexical Units: Experiments with Methods in French Diachronic Corpora

Emmanuel Cartier

University Paris 13, Sorbonne, Paris, France
E-mail: emmanuel.cartier@lipn.univ-paris13.fr

This presentation will explore several methods and techniques to automatically track and characterize the life-cycle of lexical units (LU) in diachronic corpora. We can identify from several linguistic studies (Meillet, 1904; Coseriu, 1981; Lipka et al., 2004; Blythe & Croft, 2012; Traugott and Trousdale, 2013; Nevalainen, 2015) that lexical units can be in five main time phases: *emergence*, when they first appear in discourse; *diffusion*, when they begin to be used by more and more people throughout the community; *adoption*, when they are integrated in the lexical stock of most members of the community; *degenerescence*, when their use begins to drop; *death*, when they are not used anymore, or considered as “archaisms”.

The most obvious criterium to assess the state of a LU in the lexical memory of users is its frequency of usage at a given moment in time and its evolution through time. This factual frequency in corpora is one of the main determinants of the *entrenchment* of concepts (Langacker, 1989; Schmid, 2017). Several metrics have been devised to tune the raw frequency and associated concepts (Hilpert & Gries, 2016), and time series analysis gives us several tools to partition through time the main phases of the life of an LU (Rogers, 2010; Blythe & Croft, 2012; Gries & Hilpert, 2008). We will present, with the help of Google Ngrams corpus (Michel et al, 2010) and the Timestamped JSI web corpus 2014-2018 (Trampus & Novak, 2012), several experiments to evaluate the accuracy of these approaches.

We will show that the sole frequency notably does not permit us to track new meanings emerging from an existing formal unit nor semantic field reorganizations, and that, aside from emergence and disappearance, it is not accurate to assess the paths to diffusion and adoption.

In the second part, we will propose complementary techniques to deal with these cases, taking into account not only the frequency of the formal LU, but its combinatorial profile (Gries, 2012) and its distributional profile (Baroni & Lenci, 2010; Cartier 2016; Mikolov et al., 2013). We will present some first experiments for each of these approaches.

In the third part, we will present the necessity to take into account the socio-pragmatic parameters (Schmid, 2015) implied in any communication situation where LU are encountered, as these permit us to better assess their diffusion and/or adoption. This approach has been developed in relation to Social Networks Analysis (Fagyal et al, 2010) and the sociolinguistic *community of practice* approach (Milroy & Milroy, 1985; Eckert, 2012). We will show that its computational counterpart (Clem, 2016) should be part of a scientific program to follow the diffusion of LU through social networks.

We will end this presentation with some hints on a program to develop research and applications in the area of the diachronic evolution of lexical units.

Keywords: lexical innovation; lexical evolution; life-cycle tracking of lexical units; quantitative linguistics; S-Curve; social networks analysis; combinatorial profile; distributional profile

References

Blythe, R. A. & Croft, W. (2012). S-Curves and the Mechanisms of Propagation in Language Change. *Language*, pp 269–304.

Baroni, M. & Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4), pp. 673–721.

Cartier, E. (2016). Distributionnalisme et sémantique: état des lieux en traitement automatique des langues, pp. 288–313. Paris: CRL.

Clem, E. (2016). *Social Network Structure, Accommodation, and Language Change.* UC Berkeley Phonetics and Phonology Lab Annual Report.

Coseriu, E. (1981). Los conceptos de 'dialecto', 'nivel' y 'estilo de lengua' y el sentido propio de la dialectología. *Linguistica española actual*, III/1(1), pp. 1–33.

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual review of Anthropology*, 41, pp. 87–100.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L. & Lakkaraju, K. (2010). Centers and Peripheries: Network Roles in Language Change. *Lingua*, 120(8), pp. 2061–2079.

Gries, S. T. (2012). Behavioral Profiles: a Fine-Grained and Quantitative Approach in Corpus-based Lexical Semantics. In G. Jarema, G. Libben & C. Westbury (eds.) *Methodological and analytic frontiers in lexical research.* Amsterdam & Philadelphia: John Benjamins, 57-80. [reprint of 2010h].

Gries, S. T. & Hilpert M. (2008). The Identification of Stages in Diachronic Data: Variability-based Neighbor Clustering. *Corpora* 3(1), pp. 59-81.

Hilpert, M. & Gries, S. T. (2016). Quantitative Approaches to Diachronic Corpus Linguistics. In M. Kytö & P. Pahta (eds.) *The Cambridge Handbook of English Historical Linguistics.* Cambridge: Cambridge University Press, pp. 36-53.

Langacker, R. W. (1987). *Foundations of cognitive grammar.* Volume 1. Theoretical prerequisites. Stanford University Press Stanford.

Lipka, L., Handl, S. & Falkner, W. (2004). Lexicalization & institutionalization: the state of the art in 2004. *SKASE Journal of Theoretical Linguistics*, 1(1), pp. 1–18.

Meillet, A. (1904). Comment les mots changent de sens. *L'Année sociologique* (1896/1897-1924/1925), 9, pp. 1–38.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. et al. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science.*

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781.

Milroy, J. & Milroy, L. (1985). Linguistic Change, Social Network and Speaker Innovation. *Journal of Linguistics*, 21(2), pp. 339–384.

Nevalainen, T. (2015). Descriptive Adequacy of the S-Curve Model in Diachronic Studies of Language Change. *Varieng*, 16.

Rogers, E. M. (2010). Diffusion of innovations, fourth edition [1953]. Simon and Schuster.

Schmid, H.-J. (2015). A Blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association*, 3(1), pp. 1–27.

Schmid, H.-J. (2017). *A Framework for Understanding Linguistic Entrenchment and its Psychological Foundations in Memory and Automatization.* Mouton de Gruyter.

Trampus, M. & Novak, B. (2012). The Internals Of An Aggregated Web News Feed. In *Proceedings of 15th Multiconference on Information Society 2012* (IS-2012).

Traugott, E. C. & Trousdale, G. (2013). *Constructionalization and Constructional Changes.* Oxford University Press, Oxford Studies in Diachronic and Historical Linguistics.

Weinreich, U., Labov, W. & Herzog, M. (1968). Empirical Foundations for a Theory of Language Change. In Winfred, P. Lehmann and M. Yakov (eds) *Directions for Historical Linguistics*, pp. 95-195.

Lexonomy Clinic

Michal Měchura[1], Miloš Jakubíček[2], Adam Rambousek[1]

[1] Natural Language Processing Centre, Masaryk University, Czech Republic
[2] Lexical Computing CZ s.r.o
E-mail: valselob@gmail.com, milos.jakubicek@sketchengine.co.uk, xrambous@fi.muni.cz

This demo will be organized as a 'clinic' for users of the open-source dictionary writing system Lexonomy (www.lexonomy.eu, Měchura, 2017). Both new and experienced users will be welcome to come by and ask specific questions or ask to have specific features explained to them. In addition, we have prepared a set of short mini-presentations on specific topics which we will be ready to present to the audience depending on their interests. These presentations will cover both new and existing Lexonomy features, including:

- New, recently introduced features:
    - Embedding images, sound files and videos in entries.
    - Creating entry-to-entry cross-references.
    - Indexing entries by things other than headwords.
- Existing but less well-known features which may need some explanation:
    - Sharing XML fragments among multiple entries (Měchura, 2018).
    - Headwords and alphabetical sorting.
    - Configuring search.
    - Entry flagging.
    - Working with Lexonomy's XML editor.
    - Keyboard navigation.
- Sketch Engine integration features (Jakubíček et al., 2018, 2017):
    - Generating a dictionary draft in Sketch Engine and post-editing it in Lexonomy.
    - Pulling individual example sentences, translations, thesaurus items and other data from Sketch Engine.

- Advanced customization features (with some hand-coding required):
- Uploading your own DTD.
- Uploading your own XSL and CSS stylesheets.
- Hand-coding your own editing widgets to override Lexonomy's default XML editor.

After the conference these mini-presentations be made available to the public as help files on the Lexonomy website.

Keywords: dictionary-writing systems; dictionary editing; Lexonomy

References

Jakubíček, M., Kovář, V., Měchura, M. & Rychlý, P. (2017). "One-Click Dictionary." presented at the Electronic lexicography in the 21st century (eLex 2017) conference.

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). "Practical Post- Editing Lexicography with Lexonomy and Sketch Engine." presented at the XVIII EURALEX International Congress: Lexicography in Global Contexts.

Měchura, M. (2017). Introducing Lexonomy: An Open-Source Dictionary Writing and Publishing System. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pp. 662–679. http://www.lexonomy.eu/docs/elex2017.pdf.

Měchura, M. (2018). Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement" In J. Čibej, V. Gorjanc, I. Kosem & S. Krek. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*. Ljubljana, Slovenia.

Corpus Filtering Via Crowdsourcing for Developing a Learner's Dictionary

Peter Dekker[1], Tanara Zingano Kuhn[2], Branislava Šandrih[3], Rina Zviel-Girshin[4], Špela Arhar Holdt[5], Tanneke Schoonheim[1]

[1] Dutch Language Institute, Netherlands
[2] Universidade de Coimbra, CELGA-ILTEC, Portugal
[3] University of Belgrade, Faculty of Philology, Serbia
[4] Ruppin Academic Center, Israel
[5] University of Ljubljana, Slovenia

Web corpora are valuable sources for the development of language learning material. They can be consulted directly by language learners, to find out how a word is used in practice, via systems such as SkeLL (Sketch Engine for Language Learning; https://skell.sketchengine.co.uk/). Web corpora can also serve as a basis for creating a learner's dictionary, where the concordances reflect contemporary language as it is used on the web. However, when using web corpora for pedagogical and lexicographical purposes, the quality of the data has to be taken into account. The data may contain inappropriate or even offensive language, thus requiring data checking and potential filtering.

In previous projects, filtering of such sensitive words, so-called PARSNIPs (Politics - Alcohol - Racism - Sex - Narcotics - Isms - Pork) (Kilgarriff et al., 2014), was done automatically by using predefined seedwords. However, this approach removes a great deal of data in a somewhat non-controlled way, while on the other hand many inappropriate sentences remain unidentified. We propose a crowdsourcing approach to filter corpora for Portuguese, Dutch, Serbian and Slovene, to make them suitable as a basis for a learner's dictionary. Here, we use samples of web corpora from the Sketch Engine corpus management system (Kilgarriff et al., 2004) as a basis, but our approach could be applied to any web corpus.

We present sentences to a crowd, consisting of native speakers of those languages, through the PYBOSSA (https://pybossa.com) platform. The sentences are selected from a sample corpus and consist of potentially good and "bad" (inappropriate) sentences. The inappropriate sentences are included as ground truth for analysis. A feature from the Sketch Engine that we use in our approach is the Good Dictionary Examples (GDEX) function (Kilgarriff et al., 2008), which ranks the concordances in the corpus according to pre-defined criteria, with the best examples at the top of the list. Potentially good sentences are extracted from the corpus, with Sketch Engine GDEX filtering on, and then filtered using a blacklist of offensive and controversial words. The bad sentences are obtained from the corpus, without GDEX filtering, and then filtered using a short blacklist of offensive words, where the remainder is kept. In both cases, words in the corpus are matched to a blacklist.

After performing the crowdsourcing experiment, the contributor judgments can be fed to a Machine Learning classification model, later applied for the automatic filtering of the remaining corpus. As can be seen, this approach puts forward challenges of a different nature. On the one hand, the efficiency of crowdsourcing for large-scale data processing needs to be evaluated. On the other hand, the crowdsourcing project has to be properly designed, so that not only valuable and reliable results can be collected, but the crowd also feels motivated to participate. Nevertheless, it should be highlighted that one of the key values of this approach lies in the possibility to obtain empirical data on the actual opinions of the community pertaining to (non)offensiveness – at least in the context of language learning.

Keywords: crowdsourcing; corpus cleaning; learner's dictionary

References

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. et al. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7-36.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the XI EURALEX International Congress*. Lorient.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*. Barcelona.

Mapping the Dictionary: Researching Design Characteristics of a New Graphical User Interface

Ryan Dawson

Universitat de Barcelona, Barcelona, Spain
E-mail: Ryan.James.Dawson@gmail.com

While technology is advancing rapidly in the field (e.g. corpus and computational linguistics, Natural Language Processing), these advances have yet to culminate in any significant changes to digital adaptations of the dictionary. In fact, the layout of the dictionary has remained relatively unchanged for centuries. In order to take lexicography beyond documentation to implementation, we must overcome one of the lingering limitations of the previous era, as the book format that has been retained throughout its digital iterations presents problems in how new resources can be integrated effectively.

With a focus on users as learners and the medium of smartphones/tablets, this research proposes a new format better suited to incorporating pertinent information for various language tasks. Information pulled from resources such as customizable corpora, lexical datasets adaptable for difference proficiencies, WordNets, and others can be simultaneously displayed using a new Graphical User Interface: an automatically generating, map/diagram format.

Preliminary tests have focused on the ability of users, without prior instruction, to correctly identify key components of the design concept, such as the implied meaning of size, colour, and spatial organization across different contexts. Tailored to multilingual individuals across a range of language levels, this new GUI offers an intuitive way to visualize, navigate, and process data by utilizing design characteristics and touchscreen gestures already familiar to those with smartphones/tablets.

Keywords: integrated resources; data visualization; graphical user interface; customization; smartphone; tablet; dictionaries; Lexical data sets

Access Routes to BODY PART Multiword Expressions in the "Big Five" MELDs: Use of Cross-referencing

Sylwia Wojciechowska

Faculty of English, Adam Mickiewicz University in Poznań
E-mail: swojciechowska@wa.amu.edu.pl

The treatment of multiword expressions (MWEs) in dictionaries has not received much attention in metalexicography (Oppentocht & Schutz, 2003: 219; Lew, 2012).The paper aims to analyse access routes to body part MWEs containing such nominals as *hand*, *head*, *shoulder*, *eye* and *ear* in the "Big Five" monolingual English learners' dictionaries online (MELDs). MWEs with body parts names have been chosen because they constitute a fairly large homogenous group, related by means of metonymic motivation (Goossens, 1990: 349-377). The examined MWEs have been selected from the entries for the above-mentioned body parts, and only the ones shared by all the five MELDs have been included. Among the investigated items are: *give somebody a hand*, *lose your head*, *a shoulder to cry on*, *see eye to eye* and *reach somebody's ears.* The study focuses on the analysis of access routes from the page of the headword to the target MWEs containing the headword. It is assumed that taking into account an extensive use of hyperlinks in online lexicography, dictionaries will exploit this lexicographic device as a way of cross-referencing, and hence facilitate access to related lexical items.

The research investigates the number and type of body part MWEs crossed-referenced in the entry proper and in separate boxes including related lexical data. It is also checked if the hyperlinks direct the user to a separate dictionary entry with the searched expression as the headword or if the cross-referenced page scrolls to the target sense which does not constitute a separate entry. Differences in the distribution of cross-references between the dictionaries under scrutiny are envisaged due to lack of consensus concerning MWEs terminology (Moon, 1998). While proposing her own typology of fixed expressions and idioms, Moon (1998: 19-20) observes that there is no generally agreed set of categories in the literature subsumed under MWEs, and clear classifications are impossible. Collocations, idioms, phrases, similes, metaphors and sayings overlap, and it is often hard to assign an MWE to a single category. Therefore, to give a broader picture, all of the above categories of MWEs are analysed in the study.

The results of the analysis demonstrate similarities and differences in the access structure between the five MELDs. All the dictionaries define some body part MWEs within the entry, and hyperlink the other ones. What differs is the extent of this cross-referencing and its position on the page: within the entry, at the bottom of the entry or in separate boxes. Each dictionary features "related words" boxes with the searched item MWEs, but again the content of these boxes differs between the dictionaries. The variety of access routes is evaluated in the study with a view to offering a more homogeneous presentation of hyperlinked related MWEs.

Keywords: multi-word expressions; body part fixed expressions and idioms; cross-references; online dictionaries; access structure; hyperlinks; related lexical data; metonymy

References

Dictionaries:

CALD: *Cambridge Advanced Learner's Dictionary.* Accessed at: http://dictionary.cambridge.org. (18 February 2019)

COBUILD: *COBUILD Advanced English Dictionary.* Accessed at: http://www.collinsdictionary.com. (18 February 2019)

LDOCE: *Longman Dictionary of Contemporary English.* Accessed at: http://www.ldoceonline.com. (18 February 2019)

MEDO: *Macmillan English Dictionary Online.* Accessed at: http://www.macmillandictionary.com. (18 February 2019)

OALD: *Oxford Advanced Learner′s Dictionary.* Accessed at: http://www.oxfordlearnersdictionaries.com. (18 February 2019)

Other references:

Goossens, L. (2002 [1990]). Metaphtonymy: The interaction of metaphor and metonymy in expressions for linguistic action. In R. Dirven & R. Pörings (eds.) *Metaphor and Metonymy in Comparison and Contrast.* Berlin: Mouton de Gruyter, pp. 349–377.

Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (eds.) *Electronic Lexicography.* Oxford: Oxford University Press, pp. 343-361.

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

Oppentocht, L. & Schutz, R. (2003). Developments in electronic dictionary design. In P. van Sterkenburg (ed.) *A Practical Guide to Lexicography*, *Terminology and Lexicography Research and Practice 6.* Amsterdam: John Benjamins, pp. 215-227.

Duden Online: Automatic Linking of Words in Definitions

Thorsten Frank, Dr. Ilka Pescheck

Bibliographisches Institut, Mecklenburgische Straße 53, Berlin, Germany
E-mail: thorsten.frank@duden.de, ilka.pescheck@duden.de

Launched in 2011, Duden online has become the most important German online dictionary with over 235,000 entries, and 27 million user contacts per month.

Initially, the content of the online dictionary came from our printed dictionaries with few automatic adaptations. Since then, our editors and our computational linguists have been working on adapting the data and structure to the differing needs of online dictionaries (cf. Müller-Spitzer, 2014).

This paper deals with one central aspect of the development of the dictionary: The implementation of links in definitions. How can we automatically select words or phrases that are significant for the understanding of a lemma definition and that should lead to other entries via hyperlinks?

For those definitions consisting of one word only, links to the full entries can be implemented easily, e.g.:

Aar, *der: Adler*

Solution: link to *Adler [= eagle]*

But other examples show that not every word can be linked and there must be a smarter process of selection, e. g.:

absichtslos*: ...; nicht absichtlich; unabsichtlich [= not intended; unintended].*

Linking to *nicht [=not]* is certainly not helpful, but do we want to link to *absichtlich,* which is the antonym of the *absichtslos*? This is still under discussion, any feedback regarding this would be welcome.

The aim of our project is to identify different morphological and semantic patterns in the multi-word definitions by using the DME (Duden morphology engine) or other free programs to be able to insert links to other dictionary entries automatically.

Currently we expect to find the following patterns:

GROUP 1: Synonymy

a) Definitions containing only one word (see above) or multiple words of the same word class, separated by commas (followed by a period or semicolon)

> ***Aktionsradius, der: 1.*** *Wirkungsbereich, Reichweite. [= range, reach]* ***2.*** *Entfernung, die ein Schiff … [distance, that a ship …]*
>
> ***angriffig: a)*** *kämpferisch, streitbar; draufgängerisch; [= militant]* ***b)*** *aggressiv [= aggressive]*

b) Single words at the end of a definition, separated by a semicolon:

> ***aasig: 1. … 2.*** *von Niedertracht, Infamie erfüllt; gemein [= imbued with malice, infamy; mean]*

GROUP 2: Morphological resemblance

a) A verb, defined by a definite article plus the nominalized verb (that means the same word stem beginning with a capital letter and ending on a derivational morpheme like *-ung*).

> ***abändern:*** *die Abänderung [=to alter: the alteration]*

b) A nominalization ending on the derivational morpheme *-heit [=-ness]* defined by a nominalization with the derivational morpheme *-sein [= being]*. Solution: link to the root word (adjective) of the *-sein*-expression

> ***Beschaffenheit, die:*** *das Beschaffensein [=the nature, quality of something]*

A general problem is that automatically inserted links can only point to lemmata, not to specific readings of the lemmata.

Keywords: word definition; linking; PoS-tagging

References

Müller-Spitzer, C. (ed.) (2014). Using Online-Dictionaries. Berlin/Boston: de Gruyter.

Abstracts of papers

The Lexicographer's Voice: Word Classes in the Digital Era

Geda Paulsen, Ene Vainik, Maria Tuulik and Ahti Lohk

Institute of the Estonian Language, Estonia
E-mail: geda.paulsen@eki.ee, ene.vainik@eki.ee, maria.tuulik@eki.ee, ahti.lohk@eki.ee

The present study examines the role of word classes in contemporary lexicography using examples from Estonian. Since Estonian is a morphologically rich language, the results may be extendable to other languages with abundant morphology. Two research questions are examined: i) What are the problems and practices of lexicographers when determining word classes? and ii) What are the needs and expectations of lexicographers for a possible digital tool that would facilitate word class identification? The results of a metalexicographic survey carried out among 23 Estonian lexicographers show the relevance of word classes as a categorial frame in their lexicographic work. There is a need to improve or reconsider the (theoretical and technical) factors influencing the process of PoS tagging. A reliable software application (provisionally a PoS evaluator) easing the decision making process would be welcome. According to the ideas suggested by the respondents, the solution would be an improved morphological and syntactic parsing system with respect to the present solutions, and a corpus-driven application presenting statistics with regard to the morphosyntactic distribution of an ambiguous word with access to the data source.

Keywords: lexicography; word classes; metalexicographic survey; Estonian

TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the *Academia das Ciências de Lisboa*

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

[1] NOVA CLUNL, Universidade NOVA de Lisboa
[2] Belgrade Center for Digital Humanities, Serbia
[3] 2Ai – Instituto Politécnico do Cávado e do Ave / Algoritmi, Universidade do Minho
E-mail: anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt; ttasovac@humanistika.org; asimoes@ipca.pt

This paper describes some experiments made while encoding the first complete dictionary of the *Academia das Ciências de Lisboa* (DACL) in the context of TEI Lex-0, a community-based interchange format for lexical data aimed at facilitating the interoperability and reusability of lexical resources. Even though the original encoding of the DACL was based on TEI, we decided to switch to TEI Lex-0 because it allowed us to streamline our encoding. Our experiments show that even though TEI Lex-0 is stricter than TEI itself (allowing fewer elements and imposing certain constraints that are not present in plain TEI), it is fully capable of representing the complexities of the entry structure of the DACL. In the paper, we discuss the TEI Lex-0 encoding of the DACL, as well as the conversion methodology and the tools used for the automatic conversion from the original encoding. We are currently focusing on the macrostructural level, more precisely on the types of lexical units and on the written and spoken forms of the lemma, providing a set of modelling principles and representation forms of every type of entry in the DACL. This paper is part of ongoing work and a contribution to the efforts of the DARIAH-ERIC Lexical Resources working group.

Keywords: dictionary encoding; lexicography; TEI; XML; TEI Lex-0

Practice of Smart LSP Lexicography: The Case of a New Botanical Dictionary with Latvian as a Basic Language

Silga Sviķe, Karina Šķirmante

Ventspils University of Applied Sciences, Inženieru Street 101, Ventspils, LV-3601, Latvia
E-mail: silga.svike@gmail.com, karina.krinkele@gmail.com

The article provides an insight into the project "A New Botanical Dictionary: Terms in Latvian, Latin, English, Russian, and German" implemented in the second half of 2017 and in 2018 within the Ventspils University of Applied Sciences (VUAS) internal call for proposals "Development of Scientific Activity at the VUAS". The VUAS Faculty of Translation Studies in collaboration with the Faculty of Information Technologies in their scientific and research work along with other Latvian universities aim to occupy a niche in the branch of applied linguistics, therefore the research is related to this discipline and offers solutions in practical lexicography.

The study describes a new botanical dictionary (NBD) – a mobile application prototype – with Latvian as a basic language. An insight into the macrostructure of the dictionary and the structure of entries is given. The research deals with questions concerning IT solutions in general (simple) and semantic search in particular. It also introduces a general search – a morphological approach developed by the authors of the research specifically for the Latvian language; this approach is used to search for Latvian botanical terms in both singular and plural forms. The extracted and linked data methodology developed by the authors is described in detail, as well as the NBD technical solutions and architecture, technologies used, database model, and additional features.

Keywords: LSP lexicography; botanical dictionary; mobile application

Challenges in the Semi-automatic Reversion of a Latvian-English Dictionary

Daiga Deksne[1], Andrejs Veisbergs[2]

[1] Tilde, Vienības gatve 75a, Riga, Latvia, LV-1004
[2] University of Latvia, Visvalža iela 4a, Riga, Latvia, LV-1050
E-mail: daiga.deksne@tilde.lv, andrejs.veisbergs@lu.lv

The electronic version of the Latvian-English dictionary has been significantly supplemented over the last year with new linguistic material from corpora, databases and other sources. In contrast, the English-Latvian dictionary can be considered outdated as its electronic version was updated 10 years ago. This motivated us to create a semi-automatic process for reversion of the Latvian-English dictionary in order to supplement the English-Latvian dictionary with missing entries. Some of the major challenges for automatic reversion were as follows: grouping translations by part of speech, deciding to which entry the example should be attached, and ordering translations with similar meaning. By using automatic scripts it was possible to create reversed entries of quite good quality within a short time. Three groups of entries were prepared for manual post-editing: new entries with a single translation, new entries with a more complex structure, and existing entries with additional new content. The tasks for post-editing are: to check the suitability of the chosen headword, part of speech and translation order, to group the translations having the same meaning, and to move examples after appropriate translations.

Keywords: electronic dictionaries; bilingual dictionary reversing; phraseology

Zapotec Language Activism and Talking Dictionaries

K. David Harrison[1], Brook Danielle Lillehaugen[2],

Jeremy Fahringer[3], Felipe H. Lopez[4]

[1] Swarthmore College, 500 College Ave., Swarthmore PA 19081, and New York Botanical Garden, 2900 Southern Blvd, The Bronx, NY 10458

[2] Haverford College, 370 Lancaster Ave., Haverford PA 19041

[3] Swarthmore College, 500 College Ave., Swarthmore PA 19081

[4] University of California, San Diego, 9500 Gilman Dr., Literature Dept 0410, La Jolla CA 92093

E-mail: harrison@swarthmore.edu, blilleha@haverford.edu, jfahrin1@swarthmore.edu, lieb@ucla.edu

Online dictionaries have become a key tool for some indigenous communities to promote and preserve their languages, often in collaboration with linguists. They can provide a pathway for crossing the digital divide and for establishing a first-ever presence on the internet. Many questions around digital lexicography have been explored, although primarily in relation to large and well-resourced languages. Lexical projects on small and under-resourced languages can provide an opportunity to examine these questions from a different perspective and to raise new questions (Mosel, 2011). In this paper, linguists, technical experts, and Zapotec language activists, who have worked together in Mexico and the United States to create a multimedia platform to showcase and preserve lexical, cultural, and environmental knowledge, share their experience and insight in creating trilingual online Talking Dictionaries in several Zapotec languages. These dictionaries sit opposite from big data mining and illustrate the value of dictionary projects based on small corpora, including having the flexibility to make design decisions to maximize community impact and elevate the status of marginalized languages.

Keywords: lexicography; collaboration; endangered languages; Zapotec

Resource Interoperability: Exploiting Lexicographic Data to Automatically Generate Dictionary Examples

María José Domínguez Vázquez[1], Miguel Anxo Solla Portela[2], Carlos Valcárcel Riveiro[3]

[1] Department of English and German Philology and Galician Language Institute, University of Santiago de Compostela

[2] Department of English and German Philology, University of Santiago de Compostela

[3] Department of English, French and German Philology, University of Vigo

Email: majo.dominguez@usc.es, miguel.solla@usc.es, carlos.valcarcel.riveiro@uvigo.es

This paper describes the different design and development stages of the MultiGenera and MultiComb prototypes for the multilingual automatic generation of dictionary examples that contain nominal argument patterns at the phrasal and sentence levels. The main objective of MultiGenera is the development of a simulator for the automatic generation of phrases in Spanish, German and French, which is based on the argument patterns of ten valency nouns. The second one, MultiComb, aims to automatically generate the phrasal and sentence contexts of the previously selected nouns in MultiGenera. In the present study we focus on the description of resource interoperability and a set of tools developed to support the methodology of both projects.

Keywords: Valency Dictionary; Argument Patterns; Natural Language Generation; WordNet; Semantics and Ontologies

Croatian Web Dictionary – Mrežnik – Linking with Other Language Resources

Lana Hudeček, Milica Mihaljević

Institute of Croatian Language and Linguistics
Republike Austrije 16, 10000 Zagreb

E-mail: lhudecek@ihjj.hr, mmihalj@ihjj.hr

The *Croatian Web Dictionary – Mrežnik* will be a free, monolingual, hypertext online dictionary consisting of three modules (general module for adult native speakers and older schoolchildren, the module for schoolchildren aged 6 to 10, and the module for non-native speakers of Croatian). *Mrežnik* is a corpus-based dictionary, not a corpus-driven dictionary, i.e. the corpus and all data extracted from it serve only as guidelines. The project started on the 1st of March 2017 and the duration of the project is four years. *Mrežnik* is based on these two Croatian corpora: *Croatian Web Repository* (http://riznica.ihjj.hr/index.hr.html) and hrWaC – the *Croatian Web Corpus* (http://nlp.ffzg.hr/resources/corpora/hrwac/). The paper will focus on the possibilities of linking the *Croatian Web Dictionary – Mrežnik* with other language resources of the Institute of Croatian Language and Linguistics and examples of entries connected to these resources will be shown.

Keywords: Mrežnik; e-lexicography; dictionary links; hyperlinks; Croatian

Representation and Classification of Polyfunctional Synsemantic Words in Monolingual Dictionaries and Language Corpora: The Case of the Croatian Lexeme *Dakle*

Virna Karlić[1], Petra Bago[2]

[1] Department of South Slavic Languages and Literatures
[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, HR-10000
Email: {vkarlic, pbago}@ffzg.hr

The paper will discuss the central issues concerning lexicographic descriptions of synsemantic words, with special regard to those with multiple syntactic and pragmatic functions. This topic will be exemplified through a description of a representative example, the Croatian lexeme *dakle* (Eng. *well*, *now*; *consequently*; *accordingly*, *so*, *then*, *therefore*, *thus*). We will focus on the shortcomings of lexicographic descriptions of such words in four contemporary monolingual dictionaries of the Croatian (standard) language. We pay particular attention to the inconsistent part of speech classification in these dictionaries, as well as to the type and content of their definitions, which generally do not take into account multiple syntactic and pragmatic functions of the word. This paper will analyse the functions and the use of lexeme *dakle*, an analysis based on language material extracted from the Croatian web corpus hrWaC, and processed by two independent annotators. We have attained fair agreement between annotators for the first task of determining the (supra)syntactic function (Cohen's κ is 0.4332), and poor agreement for the second task of determining the semantic-pragmatic function (Cohen's κ is 0.2908). Ultimately, the data collected, when compared to dictionary content, can serve as a starting point for a general discussion of an adequate methodology for lexicographic description of polyfunctional synsemantic words.

Keywords: monolingual lexicography; language corpora; pragmatics; synsemantic words; polyfunctionality; Croatian language; lexeme *dakle*

Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French

Sabine Tittel[1], Frances Gillis-Webber[2]

[1] Heidelberg Academy of Sciences and Humanities, Seminarstraße 3,
D–69117 Heidelberg, Germany
[2] Department of Computer Science, University of Cape Town, Cape Town, South Africa
E-mail: sabine.tittel@urz.uni-heidelberg.de, fran@fynbosch.com

When modelling linguistic resources as Linked Data, the identification of languages using language tags and language codes is a mandatory task. IETF's BCP 47 defines the standard for tags, and ISO 639 provides the codes. However, these codes are insufficient for the identification of diatopic variation within a language and, also, for different historical language stages. This weakness hampers the accurate identification of data, which in turn leads to ambiguity when extending, aggregating and re-using this data—a key notion of Linked Open Data and the Semantic Web. We show the limitations of language identification with a case study of French linguistic data from both a diachronic and a diatopic perspective. Our exemplary data derives from dictionaries of Old French, Middle French, and of Modern French dialects, and from a Modern French linguistic atlas. For each exemplar, we propose a solution using the *privateuse* sub-tag of BCP 47's language tag, staying within the boundaries of existing standards. Using a predefined pattern for the *privateuse* sub-tag, the solutions enable a dialect, a patois, in combination with a time period, to be defined and identified. This can lead to shared agreement of language tags that will increase interoperability within the context of Linked Data.

Keywords: language codes; language tags; language annotation; Linked Open Data; French dialects

Reengineering an Online Historical Dictionary for Readers of Specific Texts

Tarrin Wills, Ellert Þór Jóhannsson

Dictionary of Old Norse Prose, Njalsgade 136, DK-2300 Copenhagen S,
University of Copenhagen
E-mail: tarrin@hum.ku.dk, ellert@hum.ku.dk

This paper presents an example of how a digital historical dictionary can be reengineered for new uses and new audiences, without changing the underlying data and editing processes. We start from the premise that a large proportion of users of historical dictionaries will be using them to read specific old texts as part of their studies or research in fields that use the texts as source material (literature, history, religion, etc.). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* (ONP) has a vast archive of digitized texts, together with detailed referencing sufficient, in theory, to generate a glossary for each page and line of the texts. For the feature demonstrated here we reverse the normal dynamic dictionary-generation process. Instead of generating dictionary entries, the application searches for citations on an edition page and generates a running glossary to the edition, displaying it alongside the edition text. In this paper we present the new public interface to the dictionary (currently at onp.ku.dk) and the contextual glossaries that are generated from the dictionary's data. These have been developed using adaptive web technologies for use on a range of devices, including tablets and phones.

Keywords: Old Norse; lexicography; reading aids

Assessing EcoLexiCAT: Terminology Enhancement and Post-editing

Pilar León-Araúz, Arianne Reimerink, Pamela Faber

Department of Translation and Interpreting
University of Granada
E-mail: {pleon, arianne, pfaber}@ugr.es

EcoLexiCAT is a freely available online application, which integrates all features of the professional translation workflow in a stand-alone interface where a source text is interactively enriched with terminological information (i.e. definitions, translations, images, compound terms, corpus access, etc.) from different external resources. EcoLexiCAT is powered by MateCat and the external sources include EcoLexicon, BabelNet, the EcoLexicon English Corpus (powered by Sketch Engine) and IATE, as well as other common resources (e.g. Wordreference, Wikipedia, Linguee, etc.). Machine translation (MT) can also be optionally added. In order to evaluate the functionalities and performance of the tool, two experiments were carried out. In the first, one subject group used EcoLexiCAT and the other used MateCat, acting as the control group. In the second, both subject groups used EcoLexiCAT and only one used MT. Both experiments shed interesting light on user behaviour, performance and satisfaction while using EcoLexiCAT.

Keywords: EcoLexiCAT; CAT tools; terminology management; MT post-editing

Lexical Tools for Low-Resource Languages: A Livonian Case-Study

Valts Ernštreits

The University of Latvia Livonian Institute, Kronvalda 4-220, Riga LV1010, Latvia
E-mail: valts.ernstreits@lu.lv

This article focuses on the empirical experience and conclusions, resulting from the creation of language research and acquisition tools for Livonian – one of the smallest languages in Europe.

A cluster was created for Livonian containing three interconnected databases, each with distinct types of data – lexical, morphological, and a corpus. The lexical database contains the lemmas and their data, the morphological database stores morphological forms, while all textual material, including the dictionary examples, is in the corpus. When indexing the corpus, every word refers to a lemma in the lexical database and its morphological information (new lemmas are added prior to indexation), ensuring consistency of the language data, and from each database the full data set of the other databases can be accessed.

The function of each cluster is to extract the maximum amount of information from limited data sources. While technologies designed for languages with a large number of speakers focus on using quantitative methods and automation to extract qualitative information from a large and constantly expanding amount of linguistic data, the main function of technologies designed for small languages is to extract the same type of information from a limited and largely static data set.

This article also examines a string of problems faced when working with a small amount of resources (inadequate language data, insufficient personnel, lack of rules for automating processes, etc.) and methods for resolving these problems in the case of Livonian.

Keywords: Livonian; low-resource languages; lexicography; corpora; data collection

Challenges for the Representation of Morphology in Ontology Lexicons

Bettina Klimek[1], John P. McCrae[2], Julia Bosque-Gil[3],

Maxim Ionov[4], James K. Tauber[5], Christian Chiarcos[4]

[1] Institute for Applied Informatics (InfAI), Leipzig University
[2] Data Science Institute, National University of Ireland Galway
[3] Ontology Engineering Group, Universidad Politécnica de Madrid
[4] Goethe-Universität Frankfurt am Main
[5] Open Greek and Latin Project

Recent years have experienced a growing trend in the publication of language resources as Linguistic Linked Data (LLD) to enhance their discovery, reuse and the interoperability of tools that consume language data. To this aim, the OntoLex-*lemon* model has emerged as a *de facto* standard to represent lexical data on the Web. However, traditional dictionaries contain a considerable amount of morphological information which is not straightforwardly representable as LLD within the current model. In order to fill this gap a new Morphology Module of OntoLex-*lemon* is currently being developed. This paper presents the results of this model as on-going work as well as the underlying challenges that emerged during the module development. Based on the MMoOn Core ontology, it aims to account for a wide range of morphological information, ranging from endings to derive whole paradigms to the decomposition and generation of lexical entries which is in compliance to other OntoLex-*lemon* modules and facilitates the encoding of complex morphological data in ontology lexicons.

Keywords: morphology; RDF; OntoLex-*lemon*; MMoOn; inflection; derivation

The ELEXIS Interface for Interoperable Lexical Resources

John P. McCrae[1], Carole Tiberius[2], Anas Fahad Khan[3], Ilan Kernerman[4], Thierry Declerck[5,7], Simon Krek[6], Monica Monachini[3] and Sina Ahmadi[1]

[1] Data Science Institute, National University of Ireland Galway
[2] Instituut voor de Nederlandse Taal
[3] CNR- Istituto di Linguistica Computazionale «A. Zampolli»
[4] K Dictionaries
[5] Austrian Centre for Digital Humanities, Austrian Academy of Sciences
[6] Jožef Stefan Institute/University of Ljubljana
[7] DFKI GmbH, Multilinguality and Language Technology Lab

ELEXIS is a project that aims to create a European network of lexical resources, and one of the key challenges for this is the development of an interoperable interface for different lexical resources so that further tools may improve the data. This paper describes this interface and in particular describes the five methods of entrance into the infrastructure, through retrodigitization, by conversion to TEI-Lex0, by the TEI-Lex0 format, by the OntoLex format or through the REST interface described in this paper. The interface has the role of allowing dictionaries to be ingested into the ELEXIS system, so that they can be linked to each other, used by NLP tools and made available through tools to Sketch Engine and Lexonomy. Most importantly, these dictionaries will all be linked to each other through the Dictionary Matrix, a collection of linked dictionaries that will be created by the project. There are five principal ways that a dictionary maybe entered into the Matrix Dictionary: either through retrodigitization; by conversion to TEI Lex-0 by means of the forthcoming ELEXIS conversion tool; by directly providing TEI Lex-0 data; by providing data in a compatible format (including OntoLex); or by implementing the REST interface described in this paper.

Keywords: lexicography; linked data; infrastructure; ELEXIS; REST; RDF; TEI; JSON

Ontological Knowledge Enhancement in EcoLexicon

Juan Carlos Gil-Berrozpe, Pilar León-Araúz, Pamela Faber

University of Granada
Department of Translation and Interpreting, Buensuceso 11, 18071 Granada, Spain
E-mail: jcgilberrozpe@ugr.es, pleon@ugr.es, pfaber@ugr.es

Contemporary research has focused on how concepts are represented and organized in the mind, leading to neurocognitive theories such as grounded cognition or embodied cognition. These theories have greatly influenced further studies in linguistics and terminology. In this way, conceptualization, categorization, and knowledge organization are the foundation of cognitive-oriented terminology theories which highlight the relevance of situated knowledge structures, such as Frame-based Terminology. Accordingly, the practical application of Frame-based Terminology is EcoLexicon, a dynamic terminological knowledge base on environmental science. Concepts in this terminological resource are domain-specific within the Environmental Event, a model that interrelates concepts by assigning them different roles. However, the Environmental Event does not include specific category types to annotate these concepts ontologically. Therefore, this paper presents a process of ontological knowledge enhancement in EcoLexicon. This process was mainly based on the categorization of its concepts in semantic classes with a multidimensional approach. As a result, EcoLexicon was ontologically enhanced not only in terms of this categorization, but also through a redesign of the conceptual categories module, which involved modifying the existing category hierarchy and implementing new features focused on describing the combinatorial potential of concepts and categories (i.e. the conceptual combinations function and the ontological view).

Keywords: conceptual categories; conceptualization; categorization; ontology; environmental knowledge

Aggregating Dictionaries into the Language Portal Sõnaveeb: Issues With and Without Solutions

Kristina Koppel, Arvi Tavast, Margit Langemets, Jelena Kallas

Institute of the Estonian Language, Estonia
E-mail: kristina.koppel@eki.ee, arvi@tavast.ee, margit.langemets@eki.ee, jelena.kallas@eki.ee

In this paper we present Sõnaveeb, a new type of language portal of the Institute of the Estonian Language containing data from a growing number of dictionaries and termbases. Sõnaveeb currently displays a total of 200,000 Estonian headwords, obtained from many databases, with many new types of lexicographic information: collocations, etymology, multi-word expressions, etc.

The paper reports on problems encountered so far: the consistency of information and avoiding duplicates when unifying the dictionaries, turning dictionary-specific information into customizations of the central service, deciding on deliberate ambiguities, parsing data fields containing more than one data element, including textual condensation, moving from annotating form (e.g. italics) to annotating content (e.g. a citation), moving from (near) duplicates to sensible information fragments, deciding between an app and a responsive web page, and possible legal problems regarding the authorship of the new central resource, as it may become difficult to show who authored which part of the published resource.

The development of Sõnaveeb continues in the direction of both the tighter aggregation of existing datasets and the addition of new data from other dictionaries and termbases, as well as compiling new data in the new DWS Ekilex.

Keywords: lexicographic database; data aggregation; unified dictionary; Dictionary Writing System; user needs; Estonian

Automating Dictionary Production: a Tagalog-English-Korean Dictionary from Scratch

Vít Baisa[20,2], Marek Blahuš[1], Michal Cukr[1], Ondřej Herman[1,2], Miloš Jakubíček[1,2], Vojtěch Kovář[1,2], Marek Medveď[1,2], Michal Měchura[1,2], Pavel Rychlý[1,2], Vít Suchomel[1,2]

[1] Lexical Computing
[2] Masaryk University
Brno, Czech Republic
{firstname.lastname}@sketchengine.eu

In this paper we present lexicographic work on a Tagalog-English-Korean dictionary. The dictionary is created entirely from scratch and all of its content (besides audio pronunciation) is initially generated fully automatically from a large web corpus that we built for these purposes, and then post-edited by human editors. The full size of the dictionary is 45,000 entries, out of which 15,000 most frequent entries are manually post-edited, while the remaining 30,000 entries are left only as automated. The project is currently ongoing and will be finished in December 2019. The dictionary will be part of the online platform run by the Naver Corporation[1] and freely available.

Keywords: Sketch Engine; Lexonomy; post-editing lexicography; dictionary; corpus; Tagalog; Filipino; English; Korean

[20] Available at https://dict.naver.com/

Planning a Domain-specific Electronic Dictionary for the Mathematical Field of Graph Theory: Definitional Patterns and Term Variation

Theresa Kruse[1], Laura Giacomini[1,2]

[1] Institute for Information Science and Natural Language Processing (IwiSt),
Universität Hildesheim, Universitätsplatz 1, D-31141 Hildesheim
[2] Institute for Translation and Interpreting (IÜD), University of Heidelberg,
Plöck 57a, D-69117 Heidelberg
E-mail: theresa.kruse@uni-hildesheim.de, laura.giacomini@uni-hildesheim.de

We plan to create an electronic dictionary for the mathematical field of graph theory. The dictionary should help students to improve their usage of the mathematical terminology. Besides the alphabetical access, the dictionary will also provide thematic, onomasiological access; it will contain lemmas in German and English, related terms and equivalence statements. Presently, such a dictionary does not exist. The dictionary basis is formed by two corpora composed of textbooks, scientific papers and lecture notes, containing all the texts the students use in their graph theory course in German and English. In the current pre-lexicographic stage, our focus is on relations between terms and on patterns used in the corpus to express them. We collect the definition patterns in the corpus and plan to use them for term extraction. Thereby, we can extract the semantic relations at the same time. In this paper we explore in particular the synonymy relations from an orthographical, morphological and syntactic perspective and draw conclusions for data acquisition. It might be possible to apply our extraction methods later for creating dictionaries in other mathematical domains.

Keywords: terminology, mathematical; patterns; relations; term variation

Smart Lexicography for Low-Resource Languages: Lessons Learned from Buddhist Sanskrit and Classical Tibetan

Ligeia Lugli

SOAS University of London, Thornhaugh Street, London WC1H 0XG, room 339
E-mail: ll34@soas.ac.uk

Traditional lexicography requires titanic efforts and enormous resources. For many languages, such resources have never been available. As a result, they have received only limited lexicographic coverage. Today, these languages can take advantage of many of the same digital tools and strategies that have simplified and expedited dictionary-making for mainstream languages. However, the resource gap remains evident even in the digital era, with basic corpus processing tasks that lie at the foundation of contemporary 'smart lexicography' still constituting a challenge for many under-resourced languages.

Drawing on my own experience in Sanskrit and Tibetan lexicography, this paper aims to offer some guidance as to the advantages and limitations of the application of smart lexicography to under-resourced languages. In particular, this paper suggests that in order to optimize resources, it may be advisable to prioritize high-quality lexical annotation of the corpus over highly curated dictionary entries, and to let digital tools take care of the lexicographic representation of the annotated linguistic information.

Keywords: automated lexicography; GDEX; Buddhist Hybrid Sanskrit; Tibetan

An Open Online Dictionary for Endangered Uralic Languages

Mika Hämäläinen, Jack Rueter

Department of Digital Humanities, University of Helsinki
E-mail: mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi

We describe a MediaWiki-based online dictionary for endangered Uralic languages. The system makes it possible to synchronize edits done in XML-based dictionaries and edits done in the MediaWiki system. This makes it possible to integrate the system with the existing open-source Giellatekno infrastructure that provides and utilizes XML formatted dictionaries for use in a variety of NLP tasks. As our system provides an online dictionary, the XML-based dictionaries become available for a wider audience and the dictionary editing process can be crowdsourced for community engagement with a full integration to the existing XML dictionaries. We present how new automatically produced data is encoded and incorporated into our system in addition to our preliminary experiences with crowdsourcing.

Keywords: online dictionary; endangered languages; Uralic languages

Text Visualization for the Support of Lexicography-Based Scholarly Work

Shane Sheehan, Saturnino Luz

Usher Institute of Population Health Sciences & Informatics,
The University of Edinburgh, UK
E-mail: Shane.Sheehan@ed.ac.uk, S.luz@ed.ac.uk

We discuss three visualisation techniques for corpus analysis, Concordance Mosaic, Metafacet and ComFre, and explore the design rationale based on a characterization of the corpus linguistic domain. The Concordance Mosaic visualization is designed for the investigation of collocation patterns. It encodes word positions in a concordance list in a manner that emphasizes quantitative analysis of frequency or collocation statistics. Metafacet provides an interface for investigating concordance lists through the lens of meta-data. When combined with the Mosaic it provides a powerful technique for investigating collocations in the context of meta-data. ComFre can be used to compare word frequencies between two corpora of different size, it has potential use as a technique for identifying terms which are representative of the corpora under investigation. The domain characterization shows how the visualizations were designed with corpus linguistic methodologies at the core. It consists of a task analysis based on the methodology outlined in Sinclairs' *Reading Concordances: An Introduction*, and the analysis of methodology case studies from language scholars.

Keywords: visualization; concordance; frequency; meta-data; collocation

LeXmart: A Smart Tool for Lexicographers

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

José João Almeida[2]

[1] 2Ai – Instituto Politécnico do Cávado e do Ave
[2] Algoritmi, Universidade do Minho
[3] NOVA CLUNL, Universidade NOVA de Lisboa
E-mails: asimoes@ipca.pt; anasalgado@campus.fcsh.unl.pt; rute.costa@fcsh.unl.pt;
jj@di.uminho.pt

The digital era has brought some challenges to lexicographers, but it has also brought new opportunities as part of the rise of information technology and, more recently, the emergence of digital humanities. This paper provides a description of LeXmart, the framework that supports the digital development of the Portuguese Academy of Sciences Dictionary. LeXmart is a smart tool framework to support lexicographers' work that offers different types of tools, ranging from a structural editor to a set of validation tools.

Given that the dictionary is stored in eXist-DB, LeXmart is developed on top of its ecosystem, using W3C standard languages, and offering default functionalities offered by eXist-DB, namely a RESTful API.

Keywords: e-lexicography; dictionary; lexical databases; lexicographic framework; XML

Validating the OntoLex-*lemon* Lexicography Module with K Dictionaries' Multilingual Data

Julia Bosque-Gil[1,2], Dorielle Lonke[2], Jorge Gracia[3], Ilan Kernerman[2]

[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[2] K Dictionaries, Tel Aviv, Israel
[3] Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain
E-mail: jbosque@unizar.es, dorielle@kdictionaries.com, jogracia@unizar.es, ilan@kdictionaries.com

The OntoLex-*lemon* model has gradually acquired the status of *de-facto* standard for the representation of lexical information according to the principles of Linked Data (LD). Exposing the content of lexicographic resources as LD brings both benefits for their easier sharing, discovery, reusability and enrichment at a Web scale, as well as for their internal linking and better reuse of their components. However, with *lemon* being originally devised for the lexicalization of ontologies, a 1:1 mapping between its elements and those of a lexicographic resource is not always attainable. In this paper we report our experience of validating the new *lexicog* module of OntoLex-*lemon*, which aims at paving the way to bridge those gaps. To that end, we have applied the module to represent lexicographic data coming from the Global multilingual series of K Dictionaries (KD) as a real use case scenario of this module. Attention is drawn to the structures and annotations that lead to modelling challenges, the ways the *lexicog* module tackles them, and where this modelling phase stands as regards the conversion process and design decisions for KD's Global series.

Keywords: Linguistic Linked Data; RDF; multilingual; OntoLex-*lemon*; K Dictionaries

Modelling Specialized Knowledge With Conceptual Frames: The TermFrame Approach to a Structured Visual Domain Representation

Špela Vintar, Amanda Saksida, Katarina Vrtovec, Uroš Stepišnik

Faculty of Arts, University of Ljubljana, Aškerčeva 2, SI – 1000 Ljubljana
E-mail: {spela.vintar, amanda.saksida, katarina.vrtovec, uros.stepisnik} @ff.uni-lj.si

We describe an emerging knowledge base for karstology developed in line with the frame-based approach with data for three languages, English, Slovene and Croatian. An annotation framework was developed to identify the definition elements, semantic categories, relations and relation definitors in definitions of karst concepts extracted from specialized corpora. A multi-layered annotation was performed for sets of validated English and Slovene definitions. We present the distribution of semantic categories and typical definition frames for the most prominent semantic categories: surface and underground landforms, hydrological forms and geomes, for English and Slovene. The definition frames specify the typical properties of concepts we expect to be described, and in our case they were initialized by domain experts and then verified through corpus data. The structured domain representation resulting from the annotated corpus allows us to compare knowledge structures between languages, generate ideal definitions and experiment with domain visualisations, graphs and maps of geolocations.

Keywords: frame-based terminology; knowledge modelling; karstology; semantic annotation

The *Russian Academic Neography* Information Retrieval Resource

Marina N. Priemysheva, Yulia S. Ridetskaya, Kira I. Kovalenko

Institute for Linguistic Studies, Russian Academy of Sciences,
199053, 9 Tuchkov pereulok, St. Petersburg, Russia
E-mail: mn.priemysheva@yandex.ru, vjs_neolex@mail.ru, kira.kovalenko@gmail.com

Creation of electronic dictionaries and retrodigitalization are very popular trends in modern lexicography. The idea to use computer techniques in Russian neology appeared in 2013, but only recently has the *Russian Academic Neography* information retrieval resource been created. It represents both published dictionaries (annual, decadal and thirty-year dictionaries), which include about 116,000 words and collocations that had not been registered by normative explanatory dictionaries of the Russian language, and new materials that were not included in published volumes or that are being prepared for publication. Simple and advanced types of search give an opportunity to find words by various parameters (word, word component, year or time period, labels, etc.). It is also intended to include chronological and frequency parameters in the future. The aim of the *Russian Academic Neography* information retrieval resource is to represent the newest Russian vocabulary and to make it available for a wide spectrum of users.

Keywords: new-word dictionary; neography; electronic dictionary; *Russian Academic Neography*; information retrieval resource

Towards the Automatic Construction of a Multilingual Dictionary of Collocations using Distributional Semantics

Marcos Garcia, Marcos García-Salido, Margarita Alonso-Ramos

Universidade da Coruña, CITIC, Grupo LyS, Dpto de Letras, Fac. de Filoloxía. 15071, A Coruña

E-mail: {marcos.garcia.gonzalez,marcos.garcias,margarita.alonso}@udc.gal

This paper presents the method used to create a multilingual online dictionary of collocations of English, Portuguese, and Spanish. This resource is built automatically and contains three types of collocations: verb–object (e.g., "[to] issue [an] invoice"), adjective–noun ("deep shame"), and nominal compounds ("cigarette packet"). We take advantage of dependency parsing and statistical association measures to compile collocations of each language, and then we align them with their equivalents in the other languages by means of compositional methods which use cross-lingual models of distributional semantics. Collocations are extracted from large and assorted corpora, and the cross-lingual models are mapped using unsupervised approaches. For each collocation in a given language, the system shows different equivalents in the other languages, ranked by a confidence value. Besides the multilingual perspective, the resulting dictionary can also serve as a monolingual resource to retrieve the collocates of a given base, thus being a useful application to both native speakers and language learners. The dictionary will be published as an online tool, and all the resources generated in this research will be freely available.

Keywords: collocations; distributional semantics; dictionary; multilinguality

A Thesaurus of Old English as Linguistic Linked Data: Using OntoLex, SKOS and *lemon-tree* to Bring Topical Thesauri to the Semantic Web

Sander Stolk

Leiden University, Leiden, the Netherlands
E-mail: s.s.stolk@hum.leidenuniv.nl

An increasing number of dictionaries are represented on the Web in the form of linguistic linked data, utilizing OntoLex-Lemon for this purpose. Lexicographic resources other than dictionaries, however, have thus far not been the main focus of efforts surrounding this model. In this paper, we discuss porting a topical thesaurus to the Web: *A Thesaurus of Old English*. By means of this case study, this paper discusses how this thesaurus – and topical thesauri in general – can be represented with OntoLex-Lemon, SKOS and *lemon-tree* through a fully automated process.

Along with discussing the terminology required for expressing *A Thesaurus of Old English* as linguistic linked data, this paper indicates challenges encountered in the conversion process. These challenges range from material that is not meant to be made available to the general public to distinctions and relations that have been left implicit in the legacy form but are of much value and, indeed, required to be expressed explicitly in its linked data form. The aim of this paper, thus, is to provide recommendations for representing topical thesauri on the Web and to grant insight into aspects that may be encountered in porting similar lexicographic resources in the future.

Keywords: thesaurus; linguistic linked data; conversion; automation

The Semantic Network of the Spanish Dictionary During the Last Century: Structural Stability and Resilience

Camilo Garrido[1,2], Claudio Gutierrez[1,2], Guillermo Soto[3]

[1] Department of Computer Science, Universidad de Chile
[2] Millennium Institute of Foundational Research on Data
[3] Department of Linguistics, Universidad de Chile
E-mail: cgarrido@dcc.uchile.cl, cgutierr@dcc.uchile.cl, gsoto@uchile.cl

The semantic network of a dictionary is a mathematical structure that represents relationships among words of a language. In this work, we study the evolution of the semantic network of the Spanish dictionary during the last century, beginning in 1925 until 2014. We analysed the permanence and changes of its structural properties, such as size of components, average shortest path length, and degree distribution. We found that global structural properties of the Spanish dictionary network are remarkably stable. In fact, if we remove all the labels from the network, networks from different editions of the Spanish dictionary are practically indistinguishable. On the other hand, local properties change over the years offering insights about the evolution of lexicon. For instance, the neighbourhood of a single word or the shared neighbourhood between a pair of words. This paper presents preliminary evidence that dictionary networks are an interesting language tool and good proxies to study semantic clouds of words and their evolution in a given language.

Keywords: semantic networks; dictionary networks; Spanish language

Towards a Graded Dictionary of Spanish Collocations

Marcos García Salido, Marcos Garcia, Margarita Alonso-Ramos

Universidade da Coruña, CITIC, Grupo LyS, Dpto. de Letras,
Fac. de Filoloxía. 15071, A Coruña
E-mail: {marcos.garcias, marcos.garcia.gonzalez, margarita.alonso}@udc.gal

Several recent studies have observed that texts of different quality and written by learners at different proficiency levels also vary in the lexical combinations they contain. Such variation can be operationalized by quantitatively measuring the association between the components of these lexical combinations. In particular, pointwise mutual information (MI) has proved to be a good predictor of proficiency development, as several studies on English learners' writing have shown. This paper examines whether association measures are also a good predictor for the proficiency level of texts written by learners of Spanish, with a view to using such information for grading lexical combinations in order to include them in a collocation dictionary of Spanish. The study also investigates whether the association measures that correlate with learners' proficiency level can discriminate between phraseological collocations and non-collocations. Our results show that, whereas the MI of learner texts' lexical combinations is a better predictor of author proficiency than frequency, the latter performs better in identifying phraseological collocations among the whole set of lexical combinations.

Keywords: graded collocation dictionary; CEFR proficiency level; association measures

Porting a Crowd-Sourced German Lexical Semantics Resource to Ontolex-Lemon

Thierry Declerck[1,2], Melanie Siegel[3]

[1] German Research Center for Artificial Intelligence, Stuhlsatzenhausweg 3,
66123 Saarbrücken, Germany
[2] Austrian Centre for Digital Humanities, Sonnenfelsgasse 19, 1010 Vienna, Austria
[3] Darmstadt University of Applied Science, Max-Planck-Str. 2, 64807 Dieburg, Germany
E-mail: declerck@dfki.de, melanie.siegel@h-da.de

In this paper we present our work consisting of mapping the recently created open source German lexical semantics resource "Open-de-WordNet" (OdeNet) into the OntoLex-Lemon format. OdeNet was originally created in order to be integrated in the Open Multilingual Wordnet initiative. One motivation for porting OdeNet to OntoLex-Lemon is to publish in the Linguistic Linked Open Data cloud this new WordNet-compliant resource for German. At the same time we can with the help of OntoLex-Lemon link the lemmas of OdeNet to full lexical descriptions and so extend the linguistic coverage of this new WordNet resource, as we did for French, Italian and Spanish wordnets included in the Open Multilingual Wordnet collection. As a side effect, the porting of OdeNet to OntoLex-Lemon helped in discovering some issues in the original data.

Keywords: Open Multilingual Wordnet; OntoLex-Lemon; OdeNet; Lexical Semantics

SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian

Ranka Stanković[1], Branislava Šandrih[1], Rada Stijović[2], Cvetana Krstev[1], Duško Vitas[1], Aleksandra Marković[2]

[1] University of Belgrade, Studentski trg 1, Belgrade, Serbia
[2] Institute for Serbian Language, SASA, Knez Mihailova 36, Belgrade, Serbia
E-mail: ranka@rgf.rs, branislava.sandrih@fil.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs, cvetana@matf.bg.ac.rs, vitas@matf.bg.ac.rs, aleksandra.markovic@isj.sanu.ac.r

In this paper we present a model for selection of good dictionary examples for Serbian and the development of initial model components. The method used is based on a thorough analysis of various lexical and syntactic features in a corpus compiled of examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary. The initial set of features was inspired by a similar approach for other languages. The feature distribution of examples from this corpus is compared with the feature distribution of sentence samples extracted from corpora comprising various texts. The analysis showed that there is a group of features which are strong indicators that a sentence should not be used as an example. The remaining features, including detection of non-standard and other marked lexis from the SASA dictionary, are used for ranking. The selected candidate examples, represented as feature-vectors, are used with the GDEX ranking tool for Serbian candidate examples and a supervised machine learning model for classification on standard and non-standard Serbian sentences, for further integration into a solution for present and future dictionary production projects.

Keywords: Serbian; good dictionary examples; automatization of dictionary-making; feature extraction; machine learning

eDictionary: the Good, the Bad and the Ugly

Marijana Janjić, Dario Poljak, Kristina Kocijan

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, 10 000 Zagreb (Croatia)
E-mail: marijanajanjic@yahoo.com, poljak5@gmail.com, krkocijan@ffzg.hr

On its own, learning a new language is an inherently daunting task. Combined with lacking or simply non-existent language resources, the task itself seems almost impossible. For some languages, this scarcity of available resources is even more obvious and further complicates the issue.

With an interdisciplinary approach, a team of linguists, language teachers, information scientists, and students themselves undertook a task of developing a learner's dictionary of Asian languages. With a great deal of care and discussion, an online e-dictionary was chosen as a platform for its ease of use, accessibility, and expandability, in lieu of a traditional printed dictionary.

Since *eDictionary* is built as a website, it is established as a platform, agnostic and available to everyone with Internet access. Furthermore, such a design allows a link to resources hosted on other web portals. To that end, cooperation was initiated with *Croatian Language Portal* and their Croatian dictionary with the aim of hyperlinking all of our Croatian lemmas to their word definitions. With the added benefits of giving users the ability to request new resources while keeping track of the request internally and allowing the updates of the whole language database seamlessly, the proposed solution to *eDictionary* provides user engagement and continuous integration that should benefit us all.

Keywords: e-dictionary; learner's dictionary; user engagement; Asian languages; Croatian

A Corpus-Based Lexical Resource of Spoken German in Interaction

Meike Meliss[1], Christine Möhrs[2],

Maria Ribeiro Silveira[2], Thomas Schmidt[2]

[1] Leibniz-Institut für Deutsche Sprache, Mannheim /
Universität Santiago de Compostela (Spanien)
[2] Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: meliss@ids-mannheim.de / meike.meliss@usc.es, moehrs@ids-mannheim.de,
silveira@ids-mannheim.de, thomas.schmidt@ids-mannheim.de

This paper presents the prototype of a lexicographic resource for spoken German in interaction, which was conceived within the framework of the LeGeDe-project (LeGeDe=Lexik des gesprochenen Deutsch). First of all, it summarizes the theoretical and methodological approaches that were used for the initial planning of the resource. The headword candidates were selected by analyzing corpus-based data. Therefore, the data of two corpora (written and spoken German) were compared with quantitative methods. The information that was gathered on the selected headword candidates can be assigned to two different sections: *meanings* and *functions in interaction.*

Additionally, two studies on the expectations of future users towards the resource were carried out. The results of these two studies were also taken into account in the development of the prototype. Focusing on the presentation of the resource's content, the paper shows both the different lexicographical information in selected dictionary entries, and the information offered by the provided hyperlinks and external texts. As a conclusion, it summarizes the most important innovative aspects that were specifically developed for the implementation of such a resource.

Keywords: online lexicography; spoken German; corpus-based

Make My (Czechoslovak Word of the) Day

Michal Škrabal[1], Vladimír Benko[2]

[1] Charles University, Institute of the Czech National Corpus,
Panská 7, 110 00 Praha 1, Czech Republic
[2] Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics,
Panská 26, 811 01 Bratislava, Slovakia
E-mail: michal.skrabal@ff.cuni.cz, vladimir.benko@juls.savba.sk

Our paper introduces an experiment aimed at creating a database to be used as the source for a *Word of the Day* (*WotD*) application. Using a database of translation equivalents derived from a Czech-Slovak parallel corpus as a point of departure, semi-automated procedures are described that would preprocess the raw data so that the size of the lexicon to be processed manually is minimized. A by-product of this experiment is a list containing Czech to Slovak translation equivalents of differing levels of similarity, which could be an interesting source of information for Czech and Slovak contrastive studies.

In the last chapter the lexicographical application of acquired data is described. The criteria for selecting individual headwords remain an open question at the moment. Personally, we lean towards a combination of different aspects so that the final selection is as diverse and user-attractive as possible. The intended microstructure of the *WotD* dictionary entry is also presented. Its first peculiarity is the dual metalanguage, making it two explanatory dictionaries in one rather than a translation dictionary. Secondly, the content of the entries is closely related to the digital-born and corpus-based nature of the dictionary. Thus, some elements presented in traditional explanatory dictionaries are reduced or completely omitted in our microstructure – while others are highlighted.

Keywords: Word of the Day; translation equivalent; Czech; Slovak; *Treq* database

DiCoEnviro, a Multilingual Terminological Resource on the Environment: The Brazilian Portuguese Experience

Flávia Cristina Cruz Lamberti Arraes

Departamento de Línguas Estrangeiras e Tradução
Instituto de Letras, Universidade de Brasília, ICC – Ala Sul – Sala B1 167/63 -
Campus Universitário Darcy Ribeiro – Asa Norte – Brasília/DF CEP: 70910-900
E-mail: flavialamberti@gmail.com

DiCoEnviro is a multilingual terminological resource that contains terms in the field of the environment in different languages, i.e. French, English, Spanish, Portuguese, Italian and more recently Chinese. The present paper focuses on the Portuguese version of the resource in order to show how the terminological work has been developed particularly with the use of a Brazilian Portuguese corpus. More specifically the paper presents how DiCoEnviro i) represents the specialized meaning of the terms, ii) represents terminological structures within the environmental domain, and iii) uses lexical functions to establish connections between the terms within a lexical relation. The results show a selection of terms that belong to the environmental domain in Portuguese, particularly to deforestation, their analysis, linguistic description and representation of the most preferred lexical relations the terms establish among themselves. Terms and terminological relations for Portuguese in DiCoEnviro are under construction and our purpose is to increase the number of entries and relations that represent deforestation, as well as to expand the corpus to include other topics associated with the environment.

Keywords: environment; terminology; lexical-semantic approach

Towards Electronic Lexicography for the Kurdish Language

Sina Ahmadi[1], Hossein Hassani[2], John P. M^{c}Crae[1]

[1] Insight Centre for Data Analytics, National University of Ireland Galway
[2] Department of Computer Science and Engineering, University of Kurdistan Hewlêr
E-mail: sina.ahmadi, john.mccrae@insight-centre.org, hosseinh@ukh.edu.krd

This paper describes the development of lexicographic resources for Kurdish and provides a lexical model for this language. Kurdish is considered a less-resourced language, and currently, lacks machine-readable lexical resources. The unique potential which Linked Data and the Semantic Web offer to e-lexicography enables interoperability across lexical resources by elevating the traditional linguistic data to machine-processable semantic formats. Therefore, we present our lexicon in Ontolex-Lemon ontology as a standard model for sharing lexical information on the Semantic Web. The research covers the Sorani, Kurmanji, and Hawrami dialects of Kurdish. This research suggests that although Kurdish is a less-resourced language, in terms of documented lexicons, it has a wide range of resources, but because they are not machine-readable they could not contribute to the language processing. The outcome of this project, which is made publicly available, assists scholars in their efforts towards making Kurdish a resource-rich language.

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

SkELL Corpora as a Part of the Language Portal Sõnaveeb: Problems and Perspectives

Kristina Koppel[1], Jelena Kallas[1], Maria Khokhlova[2], Vít Suchomel[3,4], Vít Baisa[3,4], Jan Michelfeit[3]

[1] Institute of the Estonian Language, Estonia
[2] St. Petersburg State University, Russia
[3] Lexical Computing Ltd., Czech Republic
[4] Masaryk University, Czech Republic
E-mail: kristina.koppel@eki.ee, jelena.kallas@eki.ee, m.khokhlova@spbu.ru, vit.suchomel@sketchengine.co.uk, vit.baisa@sketchengine.co.uk, jan.michelfeit@sketchengine.co.uk

The paper provides an analysis of the quality and presentation of authentic corpus sentences from Sketch Engine for Language Learning (SkELL) corpora (Baisa & Suchomel 2014), based on the example of Sõnaveeb (Wordweb), a new language portal being developed in the Institute of the Estonian Language. Currently Sõnaveeb contains a total of 150,000 Estonian headwords; about 70,000 of them have Russian equivalents. Authentic corpus sentences are displayed for both languages. In some cases (e.g. terms, derived forms, compounds and multi-word expressions), corpus sentences are the only source of usage examples that are available on the portal.

We describe the parameters of Good Dictionary Examples (GDEX) (Kilgarriff et al., 2008) configurations for Estonian and for Russian used for the compilation of etSkELL 2018 and ruSkELL 1.6 corpora, give an overview of an evaluation of the GDEX configuration for Estonian, and outline the requirements for the user-friendly presentation of SkELL corpora as a part of the language portal.

Keywords: GDEX; SkELL; learner corpus; Estonian; Russian

Proto-Indo-European Lexicon and the Next Generation of Smart Etymological Dictionaries: The Technical Issues of the Preparation

Jouna Pyysalo[1], Fedu Kotiranta[2], Aleksi Sahala[1], Mans Hulden[3]

[1] University of Helsinki, Faculty of Arts, PL 24, 00014 Helsingin yliopisto
[2] Independent
[3] University of Colorado Boulder, Department of linguistics, Hellems 290,
295 UCB Boulder, CO 80309
E-mail: jouna.pyysalo@helsinki.fi, fedu@mediamoguli.fi, aleksi.sahala@helsinki.fi,
mans.hulden@gmail.com

Proto-Indo-European Lexicon (PIELex) is the generative etymological dictionary of Indo-European (IE) languages at http://pielexicon.hum.helsinki.fi. It is the first dictionary in the world capable of mechanically generating its data entries, i.e. the lexical stems of more than 120 of the most archaic IE languages. In addition, in order to solve the reverse process work has already begun on the problem of the mechanical generation of Proto-Indo-European (PIE) from the IE data,. The plan of the project as a whole is to run PIE Lexicon using an operating system (OS), a computer, under which the dictionary and its data are exclusively governed by smart features ranging from semantics to morphology, and the very root structure of Proto-Indo-European itself.

In principle PIE Lexicon is compatible with all digitized etymological dictionaries of IE languages, and as the operating system is scientifically neutral, material of any language or language family can be implemented onto the platform. By outlining the key features of the future coding plan we hope to offer ideas, assistance and support for other enterprises in the field of electronic lexicography.

Keywords: Indo-European linguistics; Proto-Indo-European; electronic lexicography; finite-state technology; historical linguistics

Repel the Syntruders! A Crowdsourcing Cleanup of the Thesaurus of Modern Slovene

Jaka Čibej, Špela Arhar Holdt

Centre for Language Resources and Technologies (Faculty of Arts, Faculty of Computer and Information Science), University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
E-mail: jaka.cibej@cjvt.si, spela.arhar@cjvt.si

The Thesaurus of Modern Slovene is the largest open-source digital collection of Slovene synonyms, published in March 2018 by the Centre of Language Resources and Technologies of the University of Ljubljana. The Thesaurus was initially compiled entirely automatically and allows users to contribute toward improving the resource by adding suggestions for missing synonyms and/or by evaluating both the synonym candidates from the initial database as well as the suggestions added by other users. As an automatically generated language resource, however, the initial database of the Thesaurus includes a certain degree of noise. In the paper, we present two crowdsourcing activities aimed at cleaning up the database. The first is a targeted annotation campaign aimed at evaluating multi-word synonym candidates in the Thesaurus, and the second is an analysis of user votes provided directly in the Thesaurus interface. Both scenarios are examples of an effective postprocessing method for an automatically generated language resource and demonstrate that crowdsourcing can play an important role in smart lexicography, especially in the case of less-resourced languages.

Keywords: crowdsourcing; synonyms; Slovene; thesaurus; digital lexicography

Communities of Related Terms in a Karst Terminology Co-occurrence Network

Dragana Miljkovic[1], Jan Kralj[1], Uroš Stepišnik[2], Senja Pollak[1,3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] University of Ljubljana, Ljubljana, Slovenia
[3] University of Edinburgh, UK
E-mail: dragana.miljkovic@ijs.si, jan.kralj@ijs.si, uros.stepisnik@gmail.com, senja.pollak@ijs.si

Karst science is an attractive field of interdisciplinary research with rich terminology. This study was performed as part of a project aiming at developing novel approaches to terminology extraction and visualization, in line with the understanding of knowledge, as represented in texts, as conceptually dynamic and linguistically varied. The aim of this paper is to investigate how powerful graph-based methods can be used for visualizing and analysing domain terminology. In order to detect communities in karst terminology, we analyse the frequently co-occurring karst terms in a scientific corpus of karstologic literature. The most frequent co-occurrence pairs, which included ten or more co-occurrences within the whole corpus, are delivered as input to the Louvain community detection algorithm and visualized as a domain graph. The resulting data was evaluated by domain experts who found that the detected term groups are meaningful and correspond to different types of karst phenomena. The results are further discussed in relation to more standard topic modelling approaches, using Latent Dirichlet Allocation and Non-negative Matrix Factorization algorithms.

Keywords: karstology; co-occurrence network; community detection algorithm; network visualization; topic modelling

Collecting Collocations for the Albanian Language

Besim Kabashi [1,2]

[1] Corpus und Computational Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Albanology
Ludwig-Maximilians-Universität München, Germany
E-mail: besim.kabashi@{fau,lmu}.de

The presented paper describes the collecting of data from different sources to build a collocation data set with the aim of compiling the first contemporary collocation dictionary for the Albanian language. The work is based (1) on the analysis of empirical data, i. e. linguistic corpora, using the computational methods and tools, as well as (2) on traditional dictionaries. As empirical data we use the AlCo (Albanian Text Corpus), the AlCoPress 2017-2019, N-Grams extracted from both, methods like Log-likelihood and Dice coefficient using the IMS Open Corpus Workbench (CWB) and the Corpus Query Processor, Web version (CQPweb). Despite the enormous support, an unsupervised automated compilation of a collocation dictionary of high quality, like those created by lexicographers, seems to be impossible without intervention. In order to complete the collection of the data we additionally use lexical information extracted from traditional dictionaries. The primary goal is to create a language resource that can be used among others also for Natural Language Processing purposes. The presented work is still in progress and, of course, will change until its final version.

Keywords: Albanian; collocations; NLP lexicography; corpus linguistics; language resources

How Can App Design Improve Lexicographic Outcomes? Examples from an Italian Idiom Dictionary

Valeria Caruso[1], Barbara Balbi[2], Johanna Monti[1], Roberta Presta[2]

[1] Università degli Studi di Napoli 'L'Orientale', Via Duomo 219 - 80138, Napoli (Italia)
[2] Università degli Studi Suor Orsola Benincasa, Via Suor Orsola 10 – 80135, Napoli (Italia)
E-mail: vcaruso@unior.it, barbara.balbi@centroscienzanuova.it,
roberta.presta@centroscienzanuova.it, jmonti@unior.it

Despite the growing number of smartphone apps used in everyday tasks, lexicographic applications are still rarely discussed. Studies focus mainly on the usability of available tools, but contributions concerning the development of dictionary apps are almost non-existent.

In this paper, three different design solutions are presented to implement a dictionary app for Italian idioms, having foreign learners as prospective users. Prototypes were sketched according to *Human-centred design* principles and by applying a participatory approach in which users contribute to the design process.

To offer a trustworthy tool, special attention was also paid to the lexicographic data provided. To this end, the *OWID Sprichwörterbuch* model was enriched with specific information to support foreign speakers, whose communicative needs had been tested in a production task with Italian idioms.

The presentation of three prototypes is specifically addressed to highlight design solutions which can guarantee descriptive richness.

Keywords: dictionary Apps; electronic lexicography; *Human-centred design*; lexicographical functions; interactive systems

Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study

Eva Sassolini[1], Anas Fahad Khan[1], Marco Biffi[2,3], Monica Monachini[1], Simonetta Montemagni[1]

[1] Istituto di Linguistica Computazionale "A. Zampolli" - CNR (Pisa, Italy)
[2] Accademia della Crusca (Firenze, Italy)
[3] Università degli Studi di Firenze (Italy)
E-mail: {eva.sassolini, fahad.khan, monica.monachini, simonetta.montemagni}@ilc.cnr.it, marco.biffi@unifi.it

The paper describes ongoing work on the digitization of an authoritative historical Italian dictionary, namely *Il Grande Dizionario della Lingua Italiana* (GDLI), with a specific view to creating the prerequisites for advanced human-oriented querying. After discussing the general approach taken to extract and structure the GDLI contents, in the paper we report the encouraging results of a case study carried out against two volumes which have been selected for the different conversion issues raised. Dictionary content extraction and structuring is being carried out through an iterative process based on hand coded patterns: starting from the recognition of the entry headword, a series of truth conditions are tested which allow the building and progressive structuring, in successive steps, of the whole lexical entry. We also started to design the representation of extracted and structured entries in a standard format, encoded in TEI. An outline of an example entry is also provided and illustrated in order to show what the end result will look like.

Keywords: historical dictionaries; automatic acquisition; TEI representation

Challenges and Difficulties in the Development of Dicionário Olímpico (2016)

Rove Chishman, Aline Nardes dos Santos, Bruna da Silva, Larissa Brangel

Unisinos University, São Leopoldo, Brazil
E-mail: rove@unisinos.br, aline.nardes@gmail.com, broonamoraes@gmail.com, larissabrangel@gmail.com

This paper discusses some theoretical and practical implications arising from the development of the Dicionário Olímpico (2016), created by the SemanTec (Semantics & Technology) research group. The Dicionário Olímpico (available at http://www.dicionarioolimpico.com.br/) is a bilingual lexicographic resource (Portuguese-English) which describes the lexicon of 40 Olympic sports. The dictionary is based on the theoretical-methodological framework of Frame Semantics, developed by Charles J. Fillmore. The paper brings some background to the Dicionário Olímpico's methodological approach. In addition, it describes the lexicographical structure of the resource and the way frame-semantic features were incorporated and adapted in this context. Finally, it explores two kinds of challenges faced by the project: the identification and description of semantic frames, and the design of a template for frame definitions. These stages of development have included some adaptations of frame-semantic concepts with the purpose of building a user-friendly, frame-based dictionary. Such challenges have enriched the lexicographic work and impacted subsequent projects that are yet to be developed by the authors.

Keywords: Frame Semantics; Frame-based dictionary; Dicionário Olímpico

Introducing Kosh, a Framework for Creating and Maintaining APIs for Lexical Data

Francisco Mondaca[1], Philip Schildkamp[2], Felix Rau[2]

[1] Cologne Center for eHumanities, University of Cologne
[2] Data Center for the Humanities, University of Cologne
E-mail: f.mondaca@uni-koeln.de, philip.schildkamp@uni-koeln.de, f.rau@uni-koeln.de

In recent years, the use of application programming interfaces (APIs) throughout the Internet has increased significantly. The main reason for this growth is the multiplicity of scenarios where APIs can be employed. In the case of APIs for lexical data, their use varies from applications for mobile devices, desktop applications to natural language processing (NLP) applications, among others. While some publishers offer their data via APIs, for most small or medium size publishers implementing and providing an API is still an obstacle due to the costs and technical expertise required for their development and maintenance. Against this background, we have developed Kosh, an open-source framework for creating and maintaining APIs for lexical data. Kosh has been conceived to minimize the technical expertise required for its use, while offering generic, flexible and efficient data management. In this article, we present the methodology employed in Kosh's development and describe its architecture and functionality.

Keywords: API; Elasticsearch; framework; GraphQL; rest

Improving Dictionaries by Measuring Atypical Relative Word-form Frequencies

Kristian Blensenius, Monica von Martens

University of Gothenburg, Dept. of Swedish, Lundgrensgatan 1B, Gothenburg (Sweden)
E-mail: kristian.blensenius@gu.se, monica.von.martens@gu.se

In this article, we discuss and give examples of how word-form frequency information derived from existing corpora statistics can be used to improve dictionary content. The frequency information is used in combination with rule-based morphological data based on derivational and inflectional information from the Swedish Morphological Database compiled at the University of Gothenburg, and the lexical database owned by the Swedish Academy. The method currently used in the ongoing project for updating the monolingual Contemporary Dictionary of the Swedish Academy is described, and some examples of dictionary entries identified as candidates for update based on frequency measures are given. Different aspects of morphological dictionary content are discussed and highlighted by comparison between the above-mentioned definition dictionary and a learner's dictionary. The role of headword or lemma as well as cross-referencing methods in a digital dictionary as compared to a printed dictionary is also discussed. Finally, a few examples of suggested modifications and enhancements are given.

Keywords: morphology; frequency; word forms

TEI Encoding of a Classical Mixtec Dictionary Using GROBID-Dictionaries

Jack Bowers[123], Mohamed Khemakhem[1456], Laurent Romary[1]

[1] Inria-ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities, Paris, France
[2] EPHE - École Pratique des Hautes Études, Paris, France
[3] ACDH - Austrian Center for Digital Humanities, Vienna, Austria
[4] UPD7 - Université Paris Diderot - Paris 7, Paris, France
[5] CMB – Centre Marc Bloch, Berlin, Germany
[6] BBAW – Berlin-Brandenburg Academy of Sciences and Humanities, Berlin, Germany
E-mail: iljackb@gmail.com, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

This paper presents the application of GROBID-Dictionaries (Khemakhem et al., 2017; Khemakhem et al., 2018a; Khemakhem et al., 2018b; Khemakhem et al., 2018c), an open source machine learning system for automatically structuring print dictionaries in digital format into TEI (Text Encoding Initiative) to a historical lexical resource of Colonial Mixtec 'Voces del Dzaha Dzahui' published by the Dominican Fray Francisco Alvarado in the year 1593. The GROBID-Dictionaries application was applied to a re-organized and modernized version of the historical resource published by Jansen and Perez Jiménez (2009). The TEI dictionary thus produced will be integrated into a language documentation project dealing with Mixtepec-Mixtec (ISO 639-3: mix) (Bowers & Romary, 2017, 2018a, 2018b), an under-resourced indigenous language native to the Juxtlahuaca district of Oaxaca Mexico.

Keywords: Mixtec; TEI; GROBID-Dictionaries

Investigating Semi-Automatic Procedures in Pattern-Based Lexicography

Laura Giacomini[12], Paolo DiMuccio-Failla

[1]Institute for Translation and Interpreting (IÜD), University of Heidelberg,
Plöck 57a, D-69117 Heidelberg
[2]Institute for Information Science and Natural Language Processing (IwiSt), University of
Hildesheim, Universitätsplatz 1, D-31141 Hildesheim
E-mail: laura.giacomini@iued.uni-heidelberg.de, paolodimuccio@gmail.com

In this contribution we present existing pattern description models with different degrees of computerization, discuss their potential from the perspective of the creation of an e-lexicographic resource for language learners, introduce the parameters of pattern accuracy and ontology reliability for a qualitative evaluation of the results, and make some proposals for a future quantitative evaluations. The models discussed are a) Hanks's CPA and the Pattern Dictionary of English Verbs (PDEV), b) methods employed by Tecling (Technologies for Linguistic Analysis, Pontifical Catholic University of Valparaiso, Chile) and Verbario, a pattern database of Spanish verbs, and c) an ongoing lexicographic project for the compilation of a learner's dictionary of Italian linked to a conceptual ontology. These approaches are founded in the tradition of theories focussing on the connection between lexis and grammar, especially in John Sinclair's view of *normal patterns of usage* as the true bearers of meaning of a language.

Keywords: pattern-based lexicography; semi-automatic procedures; ontology; pattern of usage; learner's dictionary

Enriching an Explanatory Dictionary with FrameNet and PropBank Corpus Examples

Pēteris Paikens[1], Normunds Grūzītis[2], Laura Rituma[2], Gunta Nešpore[2], Viktors Lipskis[2], Lauma Pretkalniņa[2], Andrejs Spektors[2]

[1] Latvian Information Agency LETA, Marijas street 2, Riga, Latvia
[2] Institute of Mathematics and Computer Science, University of Latvia,
Raina blvd. 29, Riga, Latvia
E-mail: peteris.paikens@leta.lv, normunds.gruzitis@ailab.lv

This paper describes ongoing work to extend an online dictionary of Latvian – Tezaurs.lv – with representative semantically annotated corpus examples according to the FrameNet and PropBank methodologies and word sense inventories. Tezaurs.lv is one of the largest open lexical resources for Latvian, combining information from more than 300 legacy dictionaries and other sources. The corpus examples are extracted from Latvian FrameNet and PropBank corpora, which are manually annotated parallel subsets of a balanced text corpus of contemporary Latvian. The proposed approach augments traditional lexicographic information with modern cross-lingually interpretable information and enables analysis of word senses from the perspective of frame semantics, which is substantially different from (complementary to) the traditional approach applied in Latvian lexicography. In cases where FrameNet and PropBank corpus evidence aligns well with the word sense split in legacy dictionaries, the frame-semantically annotated corpus examples supplement the word sense information with clarifying usage examples and commonly used semantic valence patterns. However, the annotated corpus examples often provide evidence of a different sense split, which is often more coarse-grained and, thus, may help dictionary users to cluster and comprehend a fine-grained sense split suggested by the legacy sources. This is particularly relevant in case of frequently used polysemous verbs.

Keywords: explanatory dictionary; FrameNet; PropBank; semantic annotation; Latvian

Karst Exploration: Extracting Terms and Definitions from Karst Domain Corpus

Senja Pollak[1,2], Andraž Repar[1], Matej Martinc[1], Vid Podpečan[1]

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2] Usher Institute of Population Health Sciences and Informatics,
Edinburgh Medical School, Edinburgh, UK
E-mail: senja.pollak@ijs.si, repar.andraz@gmail.com, matej.martinc@ijs.si,
vid.podpecan@ijs.si

In this paper, we present the extraction of specialized knowledge from a corpus of karstology literature. Domain terms are extracted by comparing the domain corpus to a reference corpus, and several heuristics to improve the extraction process are proposed (filtering based on nested terms, stopwords and fuzzy matching). We also use a word embedding model to extend the list of terms, and evaluate the potential of the approach from a term extraction perspective, as well as in terms of semantic relatedness. This step is followed by an automated term alignment and analysis of the Slovene and English karst terminology in terms of cognates. Finally, the corpus is used for extracting domain definitions, as well as triplets, where the latter can be considered as a potential resource for complementary knowledge-rich context extraction and visualization.

Keywords: karstology; term extraction; term embeddings; term alignment; definition extraction; triplets; specialized corpora

Designing an Electronic Reverse Dictionary Based on Two Word Association Norms of English Language

Jorge Reyes-Magaña[1,2], Gemma Bel-Enguix[1], Gerardo Sierra[1], Helena Gómez-Adorno[3]

[1] Instituto de Ingeniería, Universidad Nacional Autónoma de México, Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, Ciudad de México, México
[2] Facultad de Matemáticas, Universidad Autónoma de Yucatán, Anillo Periférico Norte, Tablaje Cat. 13615, Colonia Chuburná Hidalgo Inn, Mérida, Yucatán, México
[3] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, Ciudad de México, México
E-mail: jorge.reyes@correo.uady.mx, gbele@iingen.unam.mx, gsierram@iingen.unam.mx, helena.gomez@iimas.unam.mx

This work introduces the exploitation of some language resources, namely word association norms, for building lexical search engines. We used the Edinburgh Associative Thesaurus and the University of South Florida Free Association Norms for the construction of knowledge graphs that will let us execute algorithms over the nodes and edges in order to do a lexical search. The aim of the search is to perform an inverse dictionary search that, given the description of a concept as a query in natural language, will retrieve a target word. We evaluated two graph approaches, namely Betweenness Centrality and PageRank, using a corpus of human-definitions. The results are compared with the BM25 text-retrieval algorithm and also with an online reverse dictionary– OneLook Reverse Dictionary. The experiments show that our lexical search method is competitive with the IR models in our case study, even with a slight outperformance. This demonstrates that an inverse dictionary is possible to build with these kind of resources, no matter the language of the Word Association Norm.

Keywords: inverse dictionary; norm association words; graph theory

ELEXIFINDER:
A Tool for Searching Lexicographic Scientific Output

Iztok Kosem, Simon Krek

Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: iztok.kosem@ijs.si, simon.krek@ijs.si

Access to lexicographic research is highly important for lexicographers when conceptualizing and compiling dictionaries, and preparing their publications for presentation to the lexicographic community. There have been several attempts to offer a systematic record of lexicographic scientific output, and advanced search of it, but most of them are no longer updated, focus only on bibliographic data, and do not include works from other fields related to lexicography. The tool called Elexifinder has been developed within the European Infrastructure for Lexicography (ELEXIS) project in order to facilitate knowledge exchange in the lexicographic community and promote open access culture in lexicographic research. In this paper, we present the first version of the tool that contains 1,755 publications and 78 videos in 11 different languages, and offers various search options to users. We describe the Elexifinder architecture, the process of including content, and present the interface's features. The paper concludes with the presentation of future plans, including the various publications that will be included in the next version of the tool.

Keywords: Elexifinder; lexicographic research; ELEXIS; lexicography; online tool

Lexicographic Practices in Europe: Results of the ELEXIS Survey on User Needs

Jelena Kallas[1], Svetla Koeva[2],

Margit Langemets[1], Carole Tiberius[3], Iztok Kosem[4]

[1] Institute of the Estonian Language, jelena.kallas@eki.ee, margit.langements@eki.ee
[2] Institute for Bulgarian Language, Bulgarian Academy of Sciences, svetla@dcl.bas.bg
[3] The Dutch Language Institute, carole.tiberius@ivdnt.org
[4] Jožef Stefan Institute, iztok.kosem@ijs.si

The paper presents the results of a survey on lexicographic practices and lexicographers' needs across Europe (and beyond) both for born-digital and retrodigitized resources. The survey was conducted during the period from 11 July to 1 October 2018 in the context of the Horizon 2020 project ELEXIS (European Lexicographic Infrastructure). The survey was completed by 159 respondents from a total of 45 countries, comprising 36 European countries and nine countries outside Europe.

Looking in detail at the results of the survey, the paper focusses on determining what constitutes a job description of a modern lexicographer, including the training needed. One of more notable findings is that lexicographic training is still in most cases provided by the employer rather than obtained through formal education programmes. Furthermore, a list of various dictionary-writing systems and corpus-query systems is provided, including their features currently most often used by lexicographers. Accompanying this is information about the features lexicographer want or need in their tools. Also, the paper offers insights into current trends in lexicography and what lexicographers see as the most important emerging trends that will affect lexicography in the future. Overall, these results provide a detailed insight into what is needed in terms of tools and training and thus feed back into the ELEXIS project and will help to fine-tune resources within ELEXIS.

Keywords: e-lexicography; lexicographers' needs; survey; lexicographic practices

LexiCorp: Corpus Approach to Presentation of Lexicographic Data

Vladimír Benko

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
Panská 26, 811 01 Bratislava, Slovakia
E-mail: vladimir.benko@juls.savba.sk

We present an experiment aimed at integrating XML-encoded dictionary data with corpus processing tools. Tokenized, lemmatized and PoS-tagged, the dictionary data can be processed by a traditional corpus manager such as NoSketch Engine (NoSkE), with the main benefit being the availability of ad-hoc full-text queries, as well as queries restricted to certain structure elements, without having to know too much about the internals of the respective XML encoding. Loaded with data from several Slovak dictionaries, the beta version of the dictionary portal (referred to as LexiCorp) is already used by our lexicographers.

We demonstrate the LexiCorp operation in the "Simple Query" mode and the use of "Zone" attribute in queries. However, having in mind that all NoSkE functionalities are available, we can say that users of LexiCorp can now receive a powerful working tool.

As NoSkE is an open-source system and implementation of LexiCorp requires just a minor modification of dictionary data and NoSkE's CSS style(s), this approach is applicable to similar lexicographic projects as well. Though not intended to be a replacement of a fully-fledged Dictionary Writing System, it can be conveniently used to supplement functionalities that may be missing there, such as the use of regular expressions, statistics based on XML attributes, and queries related to morphological forms of search expressions.

Keywords: Dictionary writing system; corpus manager; full-text querying; *NoSketch Engine*

Language Varieties Meet One-Click Dictionary

Egon W. Stemle, Andrea Abel, Verena Lyding

Institute for Applied Linguistics, Eurac Research, Bolzano - Bozen, IT
E-mail: {egon.stemle,andrea.abel,verena.lyding}@eurac.edu

The goal of the STyrLogism Project is to semi-automatically extract neologism candidates (new lexemes) for the German standard variety used in South Tyrol, and generally to create the basis for long-term monitoring of its development. We use automatic lexico-semantic analytics for the lexicographic processing, but instead of continuing to develop our independent neologism detection application, we have recently become part of a thriving community of users and developers within the EU infrastructure project ELEXIS, which aims to harmonize efforts that relate to producing and making dictionary resources available, and to develop tools with consistent standards and increased interoperability. Consequently, we moved the development of our neologism application into Lexonomy, one of ELEXIS' promoted open-source projects. In the following, we report on the current state of this ongoing development by describing how we integrate our work with the Sketch Engine and Lexonomy tools, pointing out the challenges involved, and discussing how our work on language varieties can be evaluated.

Keywords: language variety; One-Click Dictionary; web corpus; dictionary of variants; ELEXIS

Abstracts of keynote talks

SIL's Language Data Collection

David Baines

SIL International, Dallas, USA
E-mail: david_baines@sil.org

SIL linguists have studied minority languages since 1934. This talk will describe the extent of SIL's language data and give a brief description of the history of its data collection methods and tools.

The translation of the Bible into many languages represents a multilingual parallel corpus. Complete translations of the New and Old Testaments exist in 690 languages. New Testament translations exist in an additional 1,550 languages. SIL is considering how to provide greater access to academic linguists to those translations for which they hold the copyright.

SIL has also published lexicons for 660 languages and vocabulary lists in an additional 200 languages, and is considering possibilities for sharing that data more widely. SIL's FieldWorks software has been used as a tool for managing lexical data and to create many of the more recent dictionaries.

Keywords: multilingual corpus; lexicon; FieldWorks; rapid word collection

Related sites

FieldWorks: Open-source dictionary editing software. https://software.sil.org/fieldworks/

FLEx Tools: Programs for manipulating FLEx data. https://github.com/cdfarrow/flextools

LanguageDepot: FieldWorks data hosting. https://public.languagedepot.org/

Language Forge: Online dictionary creation and collaboration. https://languageforge.org/

Rapid Word Collection: Create dictionaries in minority languages. http://rapidwords.net/

VOC, a Spelling Dictionary for the Portuguese Language: Role and Characteristics

Margarita Correia

CELGA-ILTEC, University of Coimbra
& University of Lisbon, Lisbon, Portugal

The Portuguese language is a pluricentric language, now spoken in eight countries from four continents. Since 1911, it has had two different spelling norms, one in Brazil and one in Portugal. During the 20th century, the authorities of these two countries struggled to get a set of spelling rules that could deal with the variations in these two national varieties, using a spelling system which is mostly phonemic. Since 1990, all Portuguese speaking countries have been bound by the Orthographic Agreement for the Portuguese Language (AOLP90). However, for more than two decades this set of spelling rules could not be shaped into a common spelling dictionary for official use in all these countries.

VOC (*Vocabulário Ortográfico Comum da Língua Portuguesa* – Common Spelling Dictionary for the Portuguese Language) was conceived and carried out within this was the context. It is a digital platform that hosts the spelling dictionaries of each Portuguese speaking country, all composed following common orthographic and lexicographic principles.

In this presentation, we will introduce VOC, highlighting its lexicographic challenges and issues, but also its political role and impact.

The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography

Alexander Geyken

Berlin-Brandenburg Academy of Sciences and the Humanities, Berlin, Germany
E-mail: geyken@bbaw.de

The Zentrum für digitale Lexikogaphie der deutschen Sprache (ZDL, Center for Digital Lexicography of the German Language) aims to provide a comprehensive and empirically reliable description of the German language from its origins to the present. To this end, four German academies in Berlin (BBAW, coordinator), Göttingen (AdGW), Leipzig (SAW), and Mainz (AdWL) have joined forces. The academies have a rich tradition of dictionary projects, encompassing historical as well as modern dictionaries and including the Grimmsches Wörterbuch, the dictionaries of Old High German, Middle High German, Early New High German and the Digital Dictionary (DWDS) of contemporary German. In addition, the Centre is cooperating with the Leibniz Institute for the German Language (IDS) for neologisms and contemporary text corpora. In order to provide a ubiquitous search interface to these diverse dictionary sources, a considerable amount of integration work will be necessary in the coming years, including work on common formats, lemma lists, as well as cross-linking references from dictionaries to corpora.

The Right Rhymes: Smart Lexicography in Full Effect

Matt Kohl

GeoPhy
E-mail: hlaford@gmail.com

This talk will introduce The Right Rhymes, an evidence-based dictionary of hip-hop language, and the survey methodologies used to build the underlying corpus, develop the dictionary data, and publish to the web. The talk will open with a demo of the rap lyrics corpus; this will include an overview of the tools used for assembly and maintenance, then dive into metadata integration and contextual data enrichment, explaining why investment in these areas improved the dictionary. Following that will be a brief examination of the dictionary data, touching on content modelling, and how that can enable flexibility and evolution in the final product. The talk will conclude with a backend-to-frontend tour of The Right Rhymes website, including API exposure, data visualization, and lessons learned in designing an interface for the modern user.

Wordnet as a Relational Semantic Dictionary Built on Corpus Data

Maciej Piasecki

Wroclaw University of Technology, Wroclaw, Poland
E-mail: maciej.piasecki@pwr.edu.pl

Princeton WordNet – the prototypical wordnet (”the mother of all wordnets”) – started off as a psycholinguistic experiment on language acquisition by children. Later, it developed into a lexico-semantic database. Thus, WordNet was not originally meant to be a dictionary, but at some point began to be treated as one. It is usually presented as a network of lexicalized concepts (represented as synsets – synonym sets). In addition, many people call it and use it as a kind of ontology. Contrary to such claims, we will argue that WordNet can be modelled (and constructed) as a relational semantic dictionary in which lexical meanings are the basic building blocks defined by a dense network of lexico-semantic relations as a primary means of their description.

From such a perspective, synsets are construed as sets of lexical meanings that share lexico-semantic relations of certain types. Thus, there is no need for assigning to them a special ontological status. Relations between synsets are just notational abbreviations for relations among lexical meanings. The whole construction of a wordnet is based on the Minimal Commitment Principle: minimizing the number of assumptions, and maximising the freedom of further interpretation of the wordnet structure.

In a way typical for dictionaries, all lexical properties are assigned to lexical meanings, especially non-relational elements of description such as usage examples, textual definitions or attributes like stylistic register. The properties, but also lexico-semantic relations, can be based on language data in a straightforward way, e.g. by various linguistic tests verified against usage examples, not only the intuitions of linguists.

In order to show the consequences of the model, we will refer to plWordNet – a wordnet of Polish – which was on its basis. A corpus-based wordnet development process has been applied in the construction of plWordNet, i.e. large text corpora were used as a source of lexical knowledge supporting the work of lexicographers to extract items such as lemmas, clusters of usage examples suggesting potential meanings, multi-word expressions, distributional models revealing the semantic relatedness or instances of lexico-semantic relations. The talk will be illustrated with examples and statistics zooming in on several details of the solution.