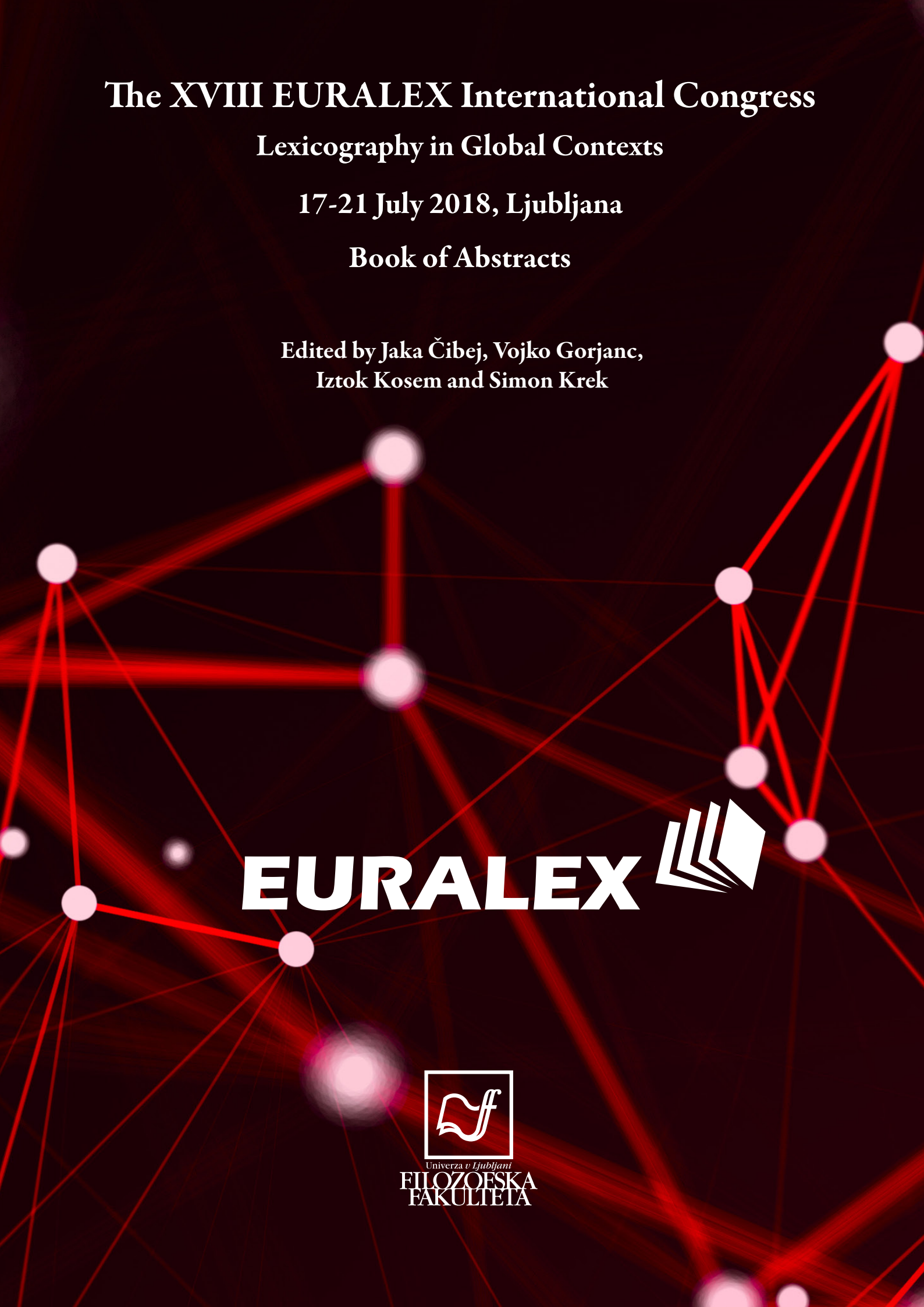
The XVIII EURALEX International Congress

Lexicography in Global Contexts

17-21 July 2018, Ljubljana

Book of Abstracts

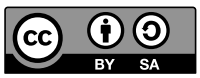
Edited by Jaka Čibej, Vojko Gorjanc,
Iztok Kosem and Simon Krek



Univerza v Ljubljani
FILOZOFSKA FAKULTETA

The XVIII EURALEX International Congress: Lexicography in Global Contexts
Book of Abstracts

Edited by: Jaka Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek
English language proofreading: Paul Steed
Technical editor: Aleš Cimprič



Published by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts
Issued by: University of Ljubljana, Centre for language resources and technologies
For the publisher: Roman Kuhar, Dean of the Faculty of Arts, University of Ljubljana

Ljubljana, 2018
First edition, e-edition
Publication is free of charge.

The editors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0215).

Acknowledgements

We would like to thank all those who have made the XVIII EURALEX International Congress possible, by contributing to the reviewing, to the logistics and by financially supporting the event. In particular, we would like to thank our sponsoring partners and patrons:

A.S. Hornby Educational Trust

Ingenierie Diffusion Multimedia Inc.

Oxford University Press

ELEXIS – European Lexicographic Infrastructure

CLARIN.SI – Common Language Resources and Technology Infrastructure, Slovenia

Alpineon d.o.o.

DELIGHT d.o.o.

TshwaneDJe

Faculty of Arts, University of Ljubljana

Programme Committee

Andrea Abel (European Academy of Bozen/Bolzano, EURAC)

Polona Gantar (University of Ljubljana, Faculty of Arts)

Zoe Gavriilidou (Democritus University of Thrace, Xanthi)

Vojko Gorjanc (University of Ljubljana, Faculty of Arts)

Iztok Kosem (University of Ljubljana, Faculty of Arts / Trojina)

Simon Krek (Chair) (University of Ljubljana, Center for Language Resources and Technologies / Jožef Stefan Institute, Artificial Intelligence Laboratory)

Robert Lew (Adam Mickiewicz University in Poznań, Faculty of English)

Tinatin Margalitadze (Ivane Javakhishvili Tbilisi State University)

Reviewers

Adam Rambousek, Agáta Karčová, Agnes Tutin, Ales Horak, Alexander Geyken, Amália Mendes. Amanda Laugesen, Andrea Abel, Anne Dykstra, Annette Klosa, Antton Gurrutxaga, Arvi Tavast, Carla Marello, Carole Tiberius, Carolin Müller-Spitzer, Chris Mulhall, Christine Moehrs, Corina Forascu, Danie Prinsloo, Edward Finegan, Egon Stemle, Elena Volodina, Francesca Frontini, Francis Bond, Frieda Steurs, Geoffrey Williams, Henrik Lorentzen, Ilan Kernerman, Iztok Kosem, Janet Decesaris, Jelena Kallas, Jette Hedegaard Kristoffersen, John McCrae, Jorge Gracia, Julia Bosque-Gil, Julia Miller, Klaas Ruppel, Kristina Strkalj Despot, Kseniya Egorova, Lars Trap-Jensen, Lionel Nicolas, Lothar Lemnitzer, Lut Colman, Magali Paquot, Margit Langemets, Maria Khokhlova, Marie-Claude L'Homme, Michal Měchura, Michal Kren, Miloš Jakubiček, Monica Monachini, Nataša Logar Berginc, Nicoletta Calzolari, Nikola Ljubešić, Nora Aranberri, Oddrun Grønvik,

Orin Hargraves, Orion Montoya, Patrick Hanks, Patrick Drouin, Paul Cook, Paz Battaner, Philipp Cimiano, Pilar León Araúz, Piotr Zmigrodzki, Pius ten Hacken, Polona Gantar, Radovan Garabík, Robert Lew, Roberto Navigli, Ruben Urizar, Rufus Gouws, Sass Bálint, Sara Može, Simon Krek, Stella Markantonatou, Svetla Koeva, Sylviane Granger, Špela Arhar Holdt, Tamás Váradi, Tanneke Schoonheim, Tatjana Gornostaja, Thierry Fontenelle, Tinatin Margalitadze, Tomaž Erjavec, Ulrich Heid, Valentina Apresjan, Vincent Ooi, Vojko Gorjanc, Xabier Artola Zubillaga, Xabier Saralegi, Yongwei Gao, Yukio Tono, Zoe Gavriilidou

european lexicographic infrastructure

Contents

EXTENDED ABSTRACTS

Extended abstracts

Lexicography and Theory

Arleta Adamska-Sałaciak
Adam Mickiewicz University
arleta@wa.amu.edu.pl

The impetus for this paper comes mainly from Tarp (2008) and the reactions the book has provoked, notably Piotrowski (2009 and 2013). In a nutshell, Tarp argues that lexicography must have a theory of its own, not least because it needs to prove its autonomy from linguistics. Therefore, to begin with, it is argued here that the autonomy of lexicography does not depend on its having its own specific theory. Afterwards, the more complex question is tackled of whether lexicography is (or can be) a science. That fundamental question cannnot be meaningfully asked, let alone answered, without acknowledging the substantial terminological confusion that clouds the issue. The (English) terms crucial to the debate – *science, lexicography, theory* – each have more than one recognised meaning, and the concepts behind them are far from universal across languages. Aside from the purely terminological concerns, that is, determining what a science *is*, one also needs to consider what a particular scientific discipline *does*. Accordingly, lexicography is also examined from the point of view of the questions that its practitioners routinely try to answer.

Keywords: lexicography, science, theory, linguistics

Updating and Expanding a Retro-digitised Dictionary: Some Insights from the *Dictionary of the Irish Language*

Sharon J Arbuthnot
Queen's University, Belfast
Sharon.Arbuthnot@qub.ac.uk

The *Dictionary of the Irish Language* (DIL), which covers the language from earliest evidence up to around 1650, was originally published in hardcopy between 1913 and 1976. In 2007, the entire dictionary contents were digitised, tagged in TEI-conformant XML and published free online. The resultant web resource (www.dil.ie) currently receives an average of around 18,000 unique visitors and 200,000 page downloads each month.

Since 2007, DIL has been the focus of two small-scale research projects, which have aimed (a) to expand the dictionary by incorporating words, senses, idiomatic usages, and diagnostic and other evidence which came to light after the original text was published, and (b) to improve the integrity of the contents by correcting errors of interpretation, inconsistences and questionable decisions regarding orthography and presentation in the original text. A partially revised version of DIL was released online in 2013 and a further set of revisions will be incorporated in 2019. In all, about 10,000 separate corrections and additions will have been made in the course of the two projects, affecting around 20% of the entries contained in the original dictionary.

As might be expected, the work of updating material has presented more complex challenges than that of drafting entirely new entries or inserting new citations, senses, and the like. Constrained by issues of finance and time, and with only two full-time members of staff, the project teams have had to work with (and sometimes around) the original text and to devise strategies for dealing with entries which are problematic in themselves but also tightly bound into the larger dictionary through cross-references. Deleting or emending such material can necessitate a series of secondary changes and lead to a situation where scholarly literature of the last century makes reference to DIL entries which no longer exist.

Based on experience, and on a working method of tackling problematic entries on a case-by-case basis, this paper takes a close look at some of the issues which arose when the original dictionary was found to be potentially misleading or in error, and outlines solutions adopted in the revised version to deal with these. Topics covered include: whether ghost-words originating within the dictionary or in published texts of manuscript materials ought to be deleted, whether it is advisable to standardise headwords, even if examples are attested only in inflected or case forms, and how to treat material which was tentatively placed under two or more different headwords in the original dictionary and is still not properly understood. Steps taken to emend definitions which are now either factually or culturally obsolete will be discussed also, and there will be some consideration of the more general complexities associated with revising a paper-born dictionary in which entries have little formal or consistent structure.

Keywords: online dictionaries, retro-digitising, content revision

The Attitude of Language Users Towards User Involvement in Dictionary Compilation

Špela Arhar Holdt[1], Jaka Čibej[2], Iztok Kosem[3]

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

[2] Jožef Stefan Institute and Faculty of Computer and Information Science

[3] Faculty of Arts, University of Ljubljana and Jožef Stefan Institute

spela.arharholdt@fri.uni-lj.si, jaka.cibej@ijs.si, iztok.kosem@ff.uni-lj.si

Between May and July 2017, a large-scale European survey on dictionary use was conducted aimed at exploring the attitude of language users towards general monolingual dictionaries. The survey involved 29 European countries and was conducted with the support of the European Network for E-lexicography (http://www.elexicography.eu/events/european-survey-on-dictionary-use/). In the survey, the users were asked about the dictionaries they use and the situations in which they use them. Additionally, every participating country was able to include questions to obtain data relevant for dictionary user research at the national level. As part of the efforts by the Centre for Language Resources and Technologies of the University of Ljubljana to implement crowdsourcing and other methods of user involvement into modern lexicographic workflows, the Slovene-specific subsection of the survey included two questions pertaining to user contributions. The questions were asked with the purpose of identifying the general attitude of language users towards the inclusion of non-experts in dictionary compilation, as well as to confirm (or reject) our own assumptions about the implementation of crowdsourcing in the process: [1] *In the digital environment, dictionaries can also be constructed with the help of language users. What is your attitude towards this type of collaboration?* [2] *Would you personally contribute to the compilation of a monolingual dictionary of Slovene?*

The survey was completed by 619 participants. For the first question, 32.1% of respondents chose the answer *I support user involvement both in the form of commenting on the existing dictionary content as well as supporting lexicographers in simpler tasks (to clean up and arrange language data, select dictionary examples, etc).* An additional 44.7% of respondents chose *I support user involvement, but only in the form of commenting on the existing dictionary content.* 9.2% of the respondents chose *It is not important to me whether users participate or not*, while 11.5% chose *I do not support user involvement. Dictionaries should be compiled by trained experts.* The response *Other* was chosen by 2.1%, while 0.3% of respondents provided no answer to the question.

For the second question, a total of 47.7% of respondents answered they would participate in the dictionary compilation: 16.3% chose *Yes, if the work were paid*; 17.6% chose *Yes, if the work was recognised by the community (e.g. for my CV or promotion)*; and 13.7% chose *Yes, as a volunteer, because I believe the task is important*. On the other hand, 37.5 % would not participate: 12.8% chose *No, because I have no time or am not interested*; and 24.7% chose *No, because I do not believe I am qualified*. 10.2% chose *I do not know*, while *Other* was chosen by 4.5%. 0.2% of respondents provided no answer to the question.

In order to identify possible trends, the answers to these two questions were juxtaposed with the answers obtained from the international section of the survey. A number of aspects were considered, such as frequency of dictionary use, preferred dictionary format (digital vs. printed), preferred type of financing for dictionary projects (private vs. public), general interest in language, etc. For example, we tested the correlation between the frequency of dictionary use and the willingness to

contribute towards dictionary construction, or between the most frequent type of dictionary format and the attitude towards user involvement. Furthermore, the wide range of metadata on respondents collected by the survey allowed us to observe the sample not only as a single group of respondents, but to pinpoint trends that are specific for various users depending on their views on different issues or other situational data (e.g. their age and profession). The exploratory analysis has shown that for a number of correlations, the sample size (N=619) is too small for statistically significant generalizations. However, it highlights several interesting trends that are useful for our future work. In some cases, the trends have confirmed our assumptions (e.g. the users who are more willing to contribute to dictionary compilation are those who profess a greater interest in language and those who prefer digital dictionaries), while others are less consistent with our expectations (e.g. younger generations demonstrate a lack of time and/or interest to contribute to dictionary compilation; on the other hand, older generations define themselves as unqualified, while middle generations are more likely to contribute provided they receive payment for their work).

The results provide a number of aspects to focus on in the second round of the survey, in which we aim to increase the sample size and provide more empirical evidence on the attitudes towards user involvement in dictionary compilation. We also plan to compare our findings with the results of other countries that have included questions on user involvement and crowdsourcing into the local part of the European dictionary survey.

Keywords: user survey, user involvement, crowdsourcing

How to connect linguistic atlases with digital lexicography ? First step for a dynamic network in the Galloromance field

Esther Baiwir, Pascale Renders
Université Lille3
Université de Liège
ebaiwir@ulg.ac.be, pascale.renders@ulg.ac.be

The digital revolution is at the origin of an international effort that started a couple of years ago towards digitizing and linking lexicographic resources. Historical French and Romance linguistics also took part in this dynamic, through initiatives such as the creation of new digital dictionaries (*Dictionnaire du Moyen Français* a.k.a. DMF, *Dictionnaire Étymologique Roman* a.k.a. DERom) and the digitization of printed work (*Trésor de la Langue Française informatisé* a.k.a. TLFi). In the future, it will be possible to link most of these key resources to the *Französisches Etymologisches Wörterbuch* (FEW), which they complete and often correct on specific diachronic and diatopic aspects (cf. Renders/Baiwir/Dethier 2015).

Among reference works addressing diatopic variations, the French linguistic atlases are not yet available in a digital form, although the preservation and visibility of the dialectal linguistic heritage would greatly benefit from its digitization and linking. Whether the integration of an atlantographic resource in a lexicographic network can be achieved, remains to be proved.

The APPI (*Atlas Pan-Picard Informatisé*) project aims to prepare the future integration of Galloromance linguistic materials into the network in the making. The study focuses on the picard dialect, which is particularly interesting because it spreads over France and Belgium and developed differently on either side of the border (cf. Baiwir 2017). The Belgian method, applied in the *Atlas Linguistique de la Wallonie* (ALW), defines a model for an atlantico-lexicographic structure which could, after a few adjustments, provide a solution to link French atlases to the FEW. In addition to this, the APPI project offers a global view on the picard heritage that goes beyond administrative borders, and its valorization through new modes of exploitation.

This new project involves three phases. The first one consists in the development of an atlantographic pan-picard corpus, unifying for the first time dialectal materials collected via surveys that were conducted in the twentieth century in France (*Atlas Linguistique et ethnographique picard* a.k.a. ALPic) and Belgium (ALW). The second step aims to turn this corpus into a digital atlanticolexicographic resource, replacing the dialectal data into the historical context of their lexical family. At this stage, an online user interface will give access to the new digital data, enabling researchers to analyse them with new methods. Finally, the third phase addresses the linking of the new digital resource with the FEW and explores the feasibility of extending the model to other linguistic areas.

The project will bring an important contribution to the fields of digital lexicography and French and Romance linguistics. The numerous projects ongoing in the latter will benefit from an easier access to picard data, in particular pan-roman initiatives (*DÉRom*, *ALiR*, etc.), whereas digital lexicography will gain a laboratory to test a model that links lexical data through their etymology. This paper exposes the scientific issues that need to be solved and explains how this new digital resource will help integrate linguistic atlases into a lexicographic network.

Keywords: dialectology, etymology, linguistic atlas, French lexicography

References

ALiR = Contini Michel (dir.), *Atlas Linguistique Roman* (1996), Vol. I, Tome 1 (Presentation), 232 pages; Tome 2 (Commentaires), 151 pages; Tome 3 (Atlas), 14 cartes, Rome, Istituto Poligrafico e Zecca dello Stato.

ALPic = Carton, Fernand & Maurice Lebegue (1989-1998). *Atlas linguistique et ethnographique picard*, 2 vol. Editions
du CNRS.

ALW = Remacle, Louis, Legros, Elisee, Lechanteur, Jean, Counet, Marie-Therese, Boutier, Marie-Guy, Baiwir, Esther
(1953 -). *Atlas linguistique de la Wallonie*, Liege, Universite de Liege (10 vol.).

Baiwir, Esther (2017). "La geographie linguistique au nord du domaine d'oil", *Bien dire et bien aprandre* 32, Le picard moderne: un etat de la recherche, 73-100.

DERom = *Dictionnaire étymologique roman*, http://www.atilf.fr/DERom

DMF = Martin, Robert et Bazin, Sylvie (dir.) (2015), *Dictionnaire du Moyen Français*, Nancy, ATILF/CNRS & Universite de Lorraine, http://www.atilf.fr/dmf.

FEW = Wartburg, Walther von *et al.* (1922–2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes*, 25 vol., Bonn/Heidelberg/Leipzig-Berlin/Bale, Klopp/Winter/Teubner/Zbinden.

Renders, Pascale, Baiwir, Esther et Dethier, Gerard (2015). "Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information", in Kosem, I., Jakubiček, M., Kallas, J. (Eds.) *et al.*, *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*, 452-460.

TLF, TLFi = P. Imbs, B. Quemada (dir.) (1971–1994), *Trésor de la langue française. Dictionnaire de la langue du XIXe et du XXe siècle (1789–1960)*, version informatisee : http://atilf.atilf.fr/tlf.htm

Collaborative Editing of Lexical and Termino-ontological Resources: a quick introduction to LexO

Andrea Bellandi, Emiliano Giovannetti, Silvia Piccini
Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa
name.surname@ilc.cnr.it

We here present LexO[1], a web collaborative editor of lexical and termino-ontological resources. As the underlying lexical model we adopted *lemon*[2] (Declerck et al. 2010), which appeared to be perfect for our purposes, in particular regarding the separation between the conceptual and linguistic dimensions (Cabré 1999; Depecker 2002). Several lexicon editors have already been proposed, among which we cite: Lexus[3] (Ringersma & Kemps-Snijders 2007), COLDIC (Bel et al. 2008), Wordnet Editor (Szymanski 2009), PoolParty (Schandl & Blumauer 2010), Lemon source[4], VocBench editor (Fiorelli et al. 2017; Stellato et al. 2015). Concerning editors of terminologies, it is worth mentioning LexGrid Editor (Johnson et al. 2005) and CoBaLT (Kenter et al. 2012). Tools more focussed on the editing of termino-ontological resources are very few: LabelTranslator (Espinoza et al. 2008), TextViz (Reymonet et al. 2007), TemaTres[5] and Tedi[6] (ontoTerminology EDItor), the latter aimed at the construction of so called "ontoterminologies" (Roche 2008).

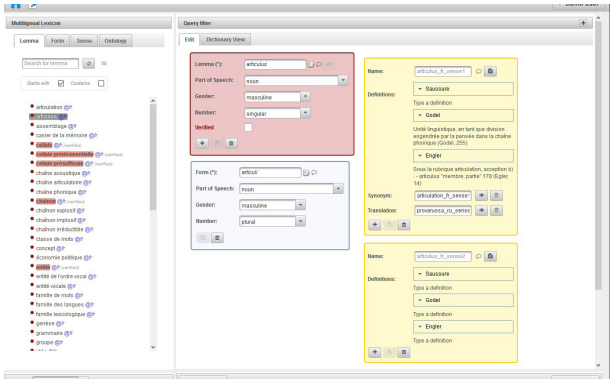


Figure 1: LexO's main interface.

The key point emerging from our review is the lack of a tool offering, at the same time, all the requirements needed by lexicographers and terminologists. As a result, we developed LexO having the following characteristics: i) *ease of use*: conceived mainly for humanists, LexO hides all the complexities related to the underlying models; ii) *collaborativeness*: LexO is collaborative and includes a locking system avoiding the concurrent editing of the same entries; iii) *sharing and linking*: the editor adheres to standards for representing and sharing lexica and ontologies in the Semantic Web; iv) *reference to texts*: LexO allows to link each entity of the resource to the relative attestations; v) *extensibility*: the system is flexible and extensible enough to formalize peculiar

1 A demo of LexO is available at the following address: http://ditmao-dev.ilc.cnr.it:8082/saussure. A selection of entries belonging to Ferdinand De Saussure's lexicon can be consulted.

2 http://lemon-model.net/

3 http://tla.mpi.nl/tools/tla-tools/lexus

4 http://lemon-model.net/download/source.html

5 http://www.vocabularyserver.com/

6 http://christophe-roche.fr/tedi

features of historical lexicons, such as diachrony (the modelling of which is inspired by (Khan et al. 2016)) and attestation, as already experimented in (Bellandi et al. 2017). These two latter features make LexO particularly suited to be applied in Humanities, although it may be used by lexicographers and terminologists in general.

LexO's interface is composed of two main sections (Fig. 1). On the left, a column shows, depending on the selected tab, the list of lemmas composing the resource, the forms, the lexical senses, and the concepts belonging to the ontology of reference (future versions will allow to manage more than one ontology).

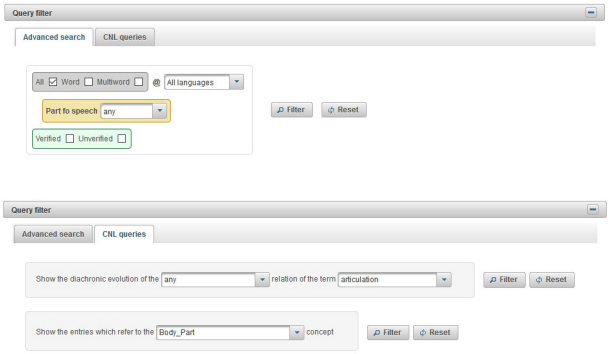


Figure 2: LexO's query panels.

If the resource is multilingual, lemmas, forms and senses can be filtered by language. The information related to the selected entry is shown in the central panel where the lemma appears in the upper part of the leftmost column on top of the relative forms. On the right, the lexical senses are shown. By selecting the "Dictionary View" tab, the central panel shows a dictionary-like rendering of all the information related to the selected entry (future versions will allow to export the whole lexicon in a dictionary style). On top of the central panel, a section can be expanded to query (also diachronically) the resource, both by filling a series of fields for advanced searches and by composing queries in a controlled natural language style (Fig. 2). The result of a diachronic query is shown in Fig. 3.

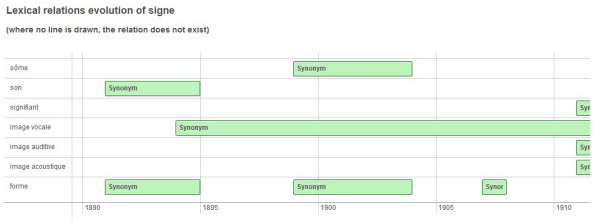


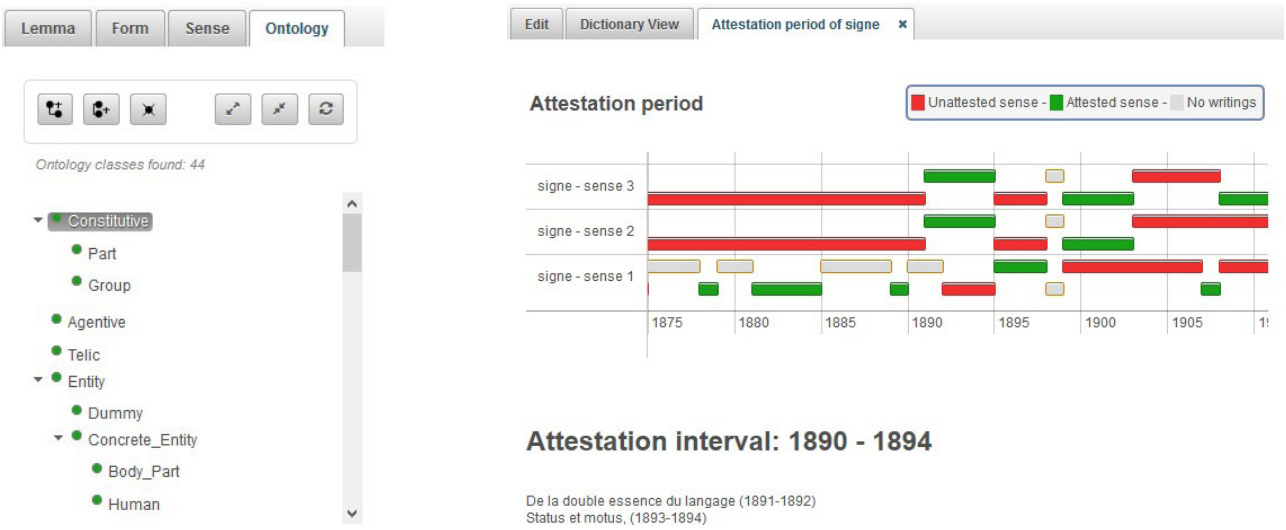
Figure 3: The result of a diachronic query[7].

A team of users can work simultaneously in LexO to create, modify or delete a lexical entry, form, sense, or concept, or to connect an entity to another entity, such as a sense to another sense via the "synonymy" property or a sense to a concept via the "ontological reference" property. The ontology can be built using a dedicated tab in the left column (Fig. 4a). To date, it is limited to a taxonomy implemented in OWL-DL. In addition, a user can: i) add a personal note to each entity, ii) view the attestation timeline of a lemma with diachronic information (Fig. 4b), and iii) view a list of concordances showing where and how a lemma, with a specific sense, is attested (Fig. 5).

7 To date, diachronic queries can be made for the following lexical entries: *antisôme*, *aposème*, *contresôme*, *parasème*, *parasôme*, *sème*, *signe*. These terms, extracted from the so-called *Notes Item* (Saussure 2002), constitute an interesting nucleus of neologisms forged by Saussure in the context of his theory of linguistic sign.

The interfaces for the creation of timelines and for the linking of entries to attestations are still in development.

LexO has been already adopted in the context of several research projects, such as: i) DiTMAO - multilingual resource of medico-botanical terminology focussed on Old Occitan (Weingart & Giovannetti 2016), ii) Ferdinand De Saussure - multilingual diachronic resource of Saussure's terminology (Piccini et al. 2016), iii) Totus Mundus - bilingual chinese-italian resource on Matteo Ricci's Atlas (Piccini 2016), iv) Clavius on the Web - diachronic resource of historical astronomy (Piccini et al. 2016). The development of LexO began in the context of the DiTMAO project and each further adaptation has been done with relative ease by extending the underlying lemon model and by updating the user interface accordingly. LexO will be kept up-to-date in accordance to the specifications defined by the OntoLex Community Group and released with an open source license.



(a) (b)

Figure 4: (a) The ontology editor tab, currently allowing to build a taxonomy of concepts; (b) Attestation interval of "signe": details of a specific selected period of sense 3 are shown at the bottom.

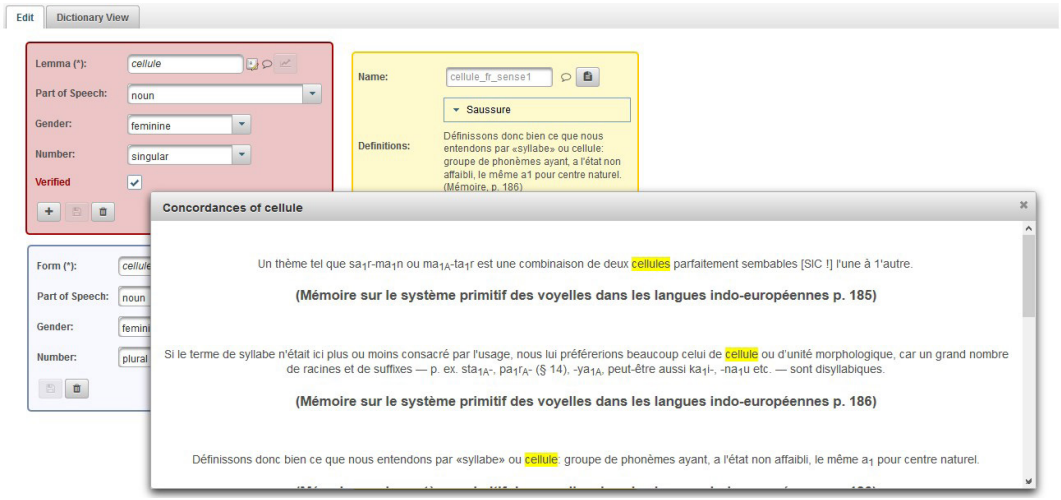


Figure 5: An example of concordances.

Keywords: computational lexicography, computational terminology, ontologies, termino-ontological resources, web tools, editing, lemon, semantic web

Acknowledgement

The development of LexO has been partially funded by the DFG in the context of the cooperation agreement between prof. 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 (August 29th, 2016).

References

Bel, N., Espeja, S., Marimon, M., and Villegas, M. 2008. COLDIC, a Lexicographic Platform for LMF compliant lexica. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco.

Bellandi, A., Boschetti, F., Khan, F., Del Grosso, A. M., Monachini, M. 2017. Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni. AIUCD 2017 Conference - Book of Abstracts, 119–125.

Cabré. M.T. 1999. Terminology. Theory, Methods and Applications. (edited by Juan C. Sager and translated by Anne DeCesaris), Amsterdam/Philadelphia, John Benjamins

Declerck, T., Buitelaar, P., Wunner, T., McCrae, J., Montiel-Ponsoda, E., and Aguado de Cea, G. 2010. Lemon: An Ontology-Lexicon model for the Multilingual Semantic Web. In Proceedings of the W3C Workshop: The Multilingual Web - Where Are We?, 26-27/10/2010, Madrid.

Depecker, L. (2002). *Entre signe et concept. Éléments de terminologie générale*. Paris : Presses Sorbonne Nouvelle.

Espinoza, M., Gómez-Pérez, A., and Mena, E. 2008. Labeltranslator - a tool to automatically localize an ontology. The Semantic Web: Research and Applications, 792–796.

Fiorelli, M., Lorenzetti, T., Pazienza, M.T. and Stellato, A. 2017. Assessing VocBench Custom Forms in Supporting Editing of Lemon Datasets, In International Conference on Language, Data and Knowledge, Springer, Cham, 237–252.

Johnson, T. M., Solbrig, H. R., Armbrust, D. C., and Chute, C. G. 2005. Lexgrid editor: Terminology authoring for the lexical grid. In AMIA 2005, American Medical Informatics Association Annual Symposium, Washington, DC, USA, October 22-26, 2005.

Kenter, T., Erjavec, T., Dulmin, M. ž., and Fišer, D. 2012. Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '12, Stroudsburg, PA, USA, Association for Computational Linguistics, 1–6.

Khan, F., Díaz-Vera, J. E., Monachini, M. 2016. Representing polysemy and diachronic lexico-semantic data on the semantic web. In Proceedings of the Second International Workshop on Semantic Web for Scientific Heritage (SW4SH), May 30, 2016, Heraklion, Greece, 37–46.

Piccini, S., Giovannetti, E., Ruimy, N., and Bellandi A. 2016. Le lexique électronique de la terminologie de Ferdinand de Saussure: une première. In Actes du XXVIIe Congrès International de Linguistique et de Philologie Romanes (CILPR 2013), Nancy, 15-20 July, 2013, 255–267.

Piccini, S., Bellandi, A., and Benotto, G. 2016. Formalizing and querying a diachronic termino-ontological resource: the clavius case study. In Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, July 11, 2016, Krakow, Poland, Number 126, 38–41.

Piccini, S. 2016. CLAVIUS: verso la modellazione di una risorsa termino-ontologica diacronica del dominio matematico-astronomico del XVII secolo. Atti del XXVI Convegno internazionale Ass.I.Term "Terminologia e organizzazione della conoscenza nella conservazione della memoria digitale".

Ringersma, J. and Kemps-Snijders, M. 2007. Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme, & R. van Son (Eds.), Proceedings of Interspeech 2007, Baixas, France, 65–68.

Roche, C. 2008. Le terme et le concept : fondements d'une ontoterminologie. TOTh 2007 : Terminologie et Ontologie : Théories et Applications, Jun 2007, Annecy, France, 1-22.

Saussure, F. de 2002. Écrits de linguistique générale. S. Bouquet and R. Engler (eds.). Paris, Gallimard.

Schandl, T. and Blumauer, A. 2010. PoolParty: SKOS Thesaurus Management Utilizing Linked Data, Berlin, Heidelberg: Springer Berlin Heidelberg, 421–425.

Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., Keizer, J., and Pazienza, M. T. 2015. VocBench: A Web Application for Collaborative Development of Multilingual Thesauri. In: Gandon F., Sabou M., Sack H., d'Amato C., Cudré-Mauroux P., Zimmermann A. (eds) The Semantic Web. Latest Advances and New Domains. ESWC 2015. Lecture Notes in Computer Science, vol 9088. Springer, Cham, 38–53.

Szymanski, J. 2009. Cooperative wordnet editor for lexical semantic acquisition. Knowledge Discovery, Knowledge Engineering and Knowledge Management, 128, 187–196.

Weingart, A. and Giovannetti, E. 2016. Extending the lemon model for a dictionary of old occitan medico-botanical terminology. In International Semantic Web Conference, Springer, 408–421.

Revision of the Norwegian dictionaries Bokmålsordboka (BOB) and Nynorskordboka (NOB)

[1]Sturla Berg-Olsen, [2]Peder Gammeltoft, [1]Knut E. Karlsen, [2]Terje Svardal
[1]Språkrådet
[2]University of Bergen
sturla.berg-olsen@sprakradet.no, Peder.Gammeltoft@uib.no, knut.e.karlsen@sprakradet.no, Terje.Svardal@uib.no

The Norwegian Language Council, in cooperation with the University of Bergen, has been granted funding through the government budget to revise the standard dictionaries of Norwegian Bokmål and Norwegian Nynorsk. The web dictionaries, jointly cared for by Språkrådet (the Norwegian Language Council) and the University of Bergen, are used extensively, but few substantial changes have been made since the first editions of the dictionaries were published in 1986.

NOB and BOB were first edited by lexicographers from the Department of Linguistic and Nordic Studies at the University of Oslo, in cooperation with Språkrådet. In 2016, responsibility for NOB and BOB was transferred from the University of Oslo to the University of Bergen, while Språkrådet remains a partner.

The Unit for Digital Documentation at the University of Oslo designed the database format and adapted the dictionaries for web publication (from 1994). This responsibility now rests with the IT department at the University of Bergen.

The dictionary portal for Nynorskordboka and Bokmålsordboka first appeared on the web in 1994. The current interface was published in 2009. The portal allows both joint and separate searches. The dictionary portal ordbok.uib.no can also be reached through dictionaryportal.eu. Since 2009, the two dictionaries have undergone a technical upgrade which has ensured full interlinking between the dictionaries and the Norwegian word bank for Nynorsk and Bokmål, a database with information on orthography and flection in accordance with the official norms. In addition, user-friendly functions such as displaying full paradigms for each lemma, expansion of abbreviations, hyperlinking of cross references between entries and searchable subentries for multi-word expressions have been added.

BOB and NOB are the only freely available online dictionaries that fully display the official norms of the two written varieties of Norwegian, i.e. Bokmål and Nynorsk. For the Language Council, the body responsible for managing and disseminating the official norms of the two varieties, the two dictionaries are of prime importance. The purpose of the revision is to bring them up to the standard necessary for them to continue to be the most important standardization tool for the Norwegian language varieties. Among the concrete tasks to be undertaken in the revision project some of those that will be given priority are:

- *Harmonizing the lemma lists*. The two dictionaries having been edited separately, there are a lot of unmotivated 'holes', i.e. lemmas that are included only in one of them. On the whole, NOB has substantially more lemmas than BOB, and a goal of the revision is to decrease this imbalance.
- *Adding new lemmas and new meanings to existing lemmas*. New material must be added to the dictionaries in order for them to reflect current Norwegian. The selection of new lemmas and meanings must be based on frequency in corpora of modern Bokmål and Nynorsk and on stability over a certain period of time.

- *Revision of definitions with a special focus on 'sensitive words'*. Many of the definitions are written in the 1980s and must be updated. This is particularly important for words which were earlier considered neutral, but are today generally viewed as derogatory.

The revision project starts in 2018 and will end in 2023. Five lexicographers will be employed to carry out the project tasks at the University of Bergen. This project will help to establish Bergen as a strong base for scholarly lexicography that will be able to take on other dictionary projects in due time.

In the paper we will discuss the content of the revision, illustrated with examples and further discuss the challenges and solutions related to the work.

Keywords: revision, monolingual dictionaries, corpus, normative dictionaries

The GA IATE Project: a discursive and interactive partnership for terminology support

Úna Bhreathnach, Christine Herwig, Gearóid Ó Cleircín, Brian Ó Raghallaigh, Hugh Rowland

Dublin City University, European Commission, NUI Galway

una.bhreathnach@dcu.ie, christine.herwig@ec.europa.eu, gearoid.ocleircin@dcu.ie, brian.oraghallaigh@dcu.ie, hughrowland85@gmail.com

The Irish language enjoys a high *de jure* status in the 1937 Constitution as the 'first official language' and as the 'national language' of the state. Notwithstanding this status, the Irish state did not seek to have Irish made an official working language of the European Economic Community (EEC) when it joined in 1973. 34 years later, however, this policy was changed and on the 1st of January 2007 Irish was recognised as an official working language of the European Union (EU). This institutional change in status created a new demand for Irish-language translation and, as a consequence, for Irish-language terminology. In order to cater for this new demand, in 2008, the Irish government, in collaboration with EU institutions, established the *GA IATE project*, the aim of which was to ensure a sufficient supply of Irish language terminology for translation requirements in the EU, through the IATE database. The IATE (Inter-Active Terminology for Europe) is a huge multilingual termbase, launched in 1999, which supports multilingual drafting of EU texts in all 24 official languages of the EU (Fontenelle 2014: 28). IATE has developed in the context of two enlargements of EU membership (2004 and 2007), and the consequent increase in the number of official languages from 11 to 24 (Bhreathnach et al. 2013: 4).

As part of the *GA IATE project*, a considerable amount of Irish-language terminology support for the EU language units has been outsourced to Fiontar & Scoil na Gaeilge (FSG), a school in Dublin City University (DCU) with a long record of terminology research. The project is funded by the Irish government as part of an ongoing effort to support the transition of Irish to the status of a full working language of the EU. The language has been operating under a derogation since 2007, limiting the categories of legislation which must be translated, but this is to be phased out by 2022. The project workflow is managed through a web-based infrastructure developed by FSG, and involves three levels of editorial research, online collaboration with EU translators and validation by the national terminology committee. The selection, extraction and subsequent import of IATE entries is overseen by the Terminology Coordination Unit of the European Commission's Directorate-General for Translation (DGT). In 2012, FSG developed an extranet to facilitate interaction and discussion with EU translators in the Irish-language units of the various institutions. This facility allows the translators to attach comments to draft terminology entries and register their views and suggestions, based on the terms recommended by FSG in the first instance. As the translators are stakeholders in the process, the extranet creates an online partnership of experts who willingly share their expertise to the benefit of the project.

This paper will evaluate the success of a project that has seen c.90,000 IATE entries processed for Irish by the project team over ten years. The specific aims and objectives have evolved over the years from an early emphasis on providing large quantities of terms, to help Irish to 'catch up' with more established EU languages, to a focus in more recent years on adding value to existing Irish entries through the provision of context, definitions and references. It will be shown that the contracting of an

external provider to supply terminology to IATE creates challenges, particularly in the context of an outsider's inevitable lack of understanding of the nuances of EU legislation and of the IATE termbase itself. A more practical issue is the impossibility of synchronising the project database with IATE. This means that IATE entries can occasionally be amended, merged or deleted while the Irish material is being worked on locally in Dublin, a less than ideal scenario. The monthly imports of newly processed Irish entries regularly throw up questions related to this lack of synchronicity.

The methodology employed by FSG in order to provide Irish-language terminology support for the EU institutions will be outlined. In particular, the paper will emphasise the technological solutions developed by FSG in order to facilitate interaction with DGT and the exchange of expertise with the Irish-language translators. It will be argued that the extranet is a productive site of discourse and interaction which allows FSG to complete its work to the highest standards. However, it will also be demonstrated that such a process can have its drawbacks. Extranet discussions of particular term proposals will occasionally illustrate conflicting ideological leanings among translators and terminologists regarding best practice, for example in the case of synonyms or in matters relating to validation.

Keywords: terminology, Irish language, EU translation

References

Bhreathnach, Ú., Cloke, F. & Nic Pháidín, C. (2013) *Terminology for the European Union. The Irish Experience: The GA IATE Project*. Galway: Cló Iar-Chonnacht

Fontenelle, T. (2014) From lexicography to terminology: a cline, not a dichotomy. In Abel, A., Vettori, C. & Ralli, N. (eds) Proceedings of the XVI EURALEX International Congress: The User in Focus, University of Bolzano/Bozen, 2014, pp.25-45.

The adjectival status of past participles in multiword units in Croatian

Goranka Blagus Bartolec
Institute of Croatian Language and Linguistics
gblagus@ihjj.hr

Pastparticiples (ending with *-an*, *-en*, and *-t*) and the group of adjectives referred to in Croatian linguistics as "true"adjectives (descriptive, possessive and material) are similar regarding their syntactic structure. They can be predicates as a complement to the verb *to be*: jabuka je*ubrana*'the apple was*picked*' (pastparticiple)/ jabuka je*svježa*'the apple is*fresh*'(true adjective). Like true adjectives, when a pastparticiple appears as a complement to a noun, it gains attributive features and takes on the grammatical categories of nominal words (gender, number, case), e.g. *nadahnut* govor 'an *inspired* speech' (pastparticiple) / *emotivan*govor 'an *emotional* speech' (descriptive adjective).

Based on the multiword units listed in the *Croatian Collocation Database*/ CCD (http://ihjj.hr/kolokacije/english/) being developed at the Institute for Croatian Language and Linguistics, this paper shall determine the collocation potential of pastparticiples. In short, the CCD contains both fixed multiword units (idioms, multiword terms, multiword proper names, collocations, proverbs) and free multiword combinations/units that are frequent in Croatian. All types of multiword units in Croatian have stable syntagmatic and syntactic structures with past participles as predicates or atributes (eg. *odrezak pečen na žaru*'a grilled steak',*promet je pojačan* 'traffic is intensified',*zrak je onečišćen*'the air is polluted',*zvijezda je rođena*'a star is born', *električno nabijena čestica* 'an electrically charged particle',*stečena prava i dužnosti* 'attained rights and duties'). The CCD isdivided into several columns: entry; part of speech; example of multiword unit; synonym; type of multiword unit (collocations, idioms, multiword terms, multiword proper names, proverbs, free combinations); label (usage and subject-field labels). The firstcolumn (entry) lists the canonical form of each word (infinitive for verbs, nominative case for nouns, pronouns and adjectives) that is a component of a single multiword unit. It is occasionally difficult to determine the canonical form of pastparticiples that are components of different multiword units, i.e. whether they belong to a verb or have adjectival meaning. Based on the analysis of multiword units from CCD that contain pastparticiples, an attempt will be made to determine whether pastparticiples in Croatian dictionaries should only be recorded as morphological (paradigmatic) forms of a verb headword, or, considering their meaning potential in concrete multiword units, as an independent adjective headword. This is important because dictionaries of Croatian regularly list pastparticiples under a verbal headword within the grammar block as a verbal form. A much smaller number of these participles have independent headword status with respect to their expanded adjectival (attributive) use (e.g., *dimljeni losos*'smoked salmon', *zaljubljeni par*lit. 'enamoured couple' / 'couple in love', *zasićena otopina*'saturated solution')

The semantic features of pastparticiples as components of multiword units will be compared with their status in the following contemporary Croatian dictionaries: *Rječnik hrvatskoga jezika*(2000), *Školski rječnik hrvatskoga jezika*(2012), *Veliki rječnik hrvatskoga standardnog jezika*(2015), *Hrvatski jezični portal*, http://hjp.znanje.hr/. A detailed investigation shall be undertaken regarding pastparticiples that appear as a component of multiword units in different contexts of use, but which are not attested in any dictionary as an independent adjectival headword. The intent of this research is to show whether it is possible to determine the degree of adjectivisation of pastparticiples based on

their collocation potential. This approach lies on Ivir's interpretation (1992/1993), which states that the prototypical, core meaning of a word changes depending on the multiword units in which it is used. This research analysed 392 pastparticiples from the CCD. A comparison of examples of multiword units containing past participles in the CCD with the representation of these past participles in Croatian dictionaries makes it possible to outline four criteria by which these verbal forms can be considered independent lexical (adjectival) units:

1. grammatical features (gender, number, case, comparison) that define them as adjectival words (*zakopano blago*'buried treasure', *pomiješani osjećaji*'mixed feelings',*spušteni strop*'lowered ceiling', *zaslužena kazna*lit. 'deserved punishment' / 'comeuppance', *najčuvaniji zatvor*'the highest-security prison', etc.)
2. the use of past participles in terminology of different professions(*sklopljena elektroda*(in chemistry) lit. 'assembled electrode' / 'coupled electrode'), *plaćeni dopust*(in administration) 'paid leave', etc.)
3. the inherent characteristic of some Croatian linguistic styles (journalism, science, legislature, etc.)
4. he homonymous nature of past participles (*smoked*, *minced*, *polluted*, *frozen*, etc.) depending on the meaning and syntactic use: past participles, as bearers of passive verbal action, have a predicate function (*the lake is frozen*), and they must be mentioned in dictionaries under the according verbal headword (*dimiti*'to smoke', *mljeti*'to mince', *onečistiti*'to pollute', *zamrznuti*'to freeze'); they also appear as adjectives (*dimljeni losos*'smoked salmon', *zamrznuta hrana*'frozen food') and deserve the status of independent lexical (adjectival) units in the dictionary.

Keywords: Croatian, past participles

References

Ivir, Vladimir. 1992/1993. Kolokacije i leksičko značenje. *Filologija*, 20–21, 181–189.

Néoveille, An Automatic System for Lexical Units Life-Cycle Tracking

Emmanuel Cartier
University Paris 13 Sorbonne Paris Cité, LIPN – RCLN UMR 7030 CNRS, Labex EFL
emmanuel.cartier@lipn.univ-paris13.fr

This paper details methods, experiments in French and a software prototype designed to track lexical units life-cycle through newspapers monitor corpora. The Néoveille platform combines state-of-the-art processes to detect and track linguistic changes and a web platform for linguists to create and manage their corpora, accept or reject automatically detected neologisms, describe linguistically the validated neologisms and follow their lifecycle on monitor corpora (Cartier, 2016). In this presentation, we will focus on the module dedicated to the life-cycle-tracking system. This task is challenging as it does not imply any creation of a new lexical item, but a *new usage of an already existing lexical item*. We propose to tackle this kind of change through four main parameters :

- the relative frequency change of the lexical units through time : timeline series analysis have a long tradition in Business Analytics and mathematical models have been proposed to detect change points and trends from frequency data; corpus linguistics have also proposed several measures to tackle diachronic change from frequency data (Hilpert and Gries, 2016); we will present the results on several measures and analysis on a French contemporary monitor newspaper corpora;
- change in the combinatorial profile of lexical units: previous approaches on "word sketch" (Kilgariff, 2004) or "behavioral profile" (Gries, 2012) have paved the way to the study of the semantic signature of lexical units through *collocations* and *collostructions*. We generalize these approaches to track combinatorial change at the lexical, lexico-syntactic and syntactic levels through the use of productivity measures applied to language models. We also propose to theoretically ground this approach on diachronic construction grammars, operationalizing so-called constructional change and constructionalization (Traugott and Trousdale, 2013).
- change in the distributional profile of lexical units: the distributional semantic approach (Pantel et al., 2010; Baroni and Lenci, 2010) enables to semantically gather lexical units through similarity of contexts. The distributional semantics approach enables to detect semantic change by explicitating, from one period to another, different similar lexical units (Hamilton, 2016). We will present some results for French and current limitations;
- diastratic and diatopic change: the last parameter enables to track the changes by keeping track of textual genres, domains and geographical metadata attached to documents where occur the lexical units, and in turn the changes in these parameters.

For the above parameters, we will present experiments on a French contemporary corpora spanning 30 years, showing that every parameter is able to track specific changes, and that a combination of parameters enables a more fine-grained caracterization of lexical change. Automatic detection is offering to lexicographers a bunch of tools to track lexical units life-cycles, taking into account linguistic and socio-linguistic parameters. All results will be available on the project website.

Keywords: neology tracking, corpus-based lexicography, semantic change, constructionalization, combinatorial profile, distributional profile, diachronic change

References

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

Cartier, E. (2016), « Neoveille, systčme de repérage et de suivi des néologismes en sept langues », *Neologica* 10, p. 101-131

Gries, S. Th. (2012) Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. In Gonia Jarema, Gary Libben, & Chris Westbury (eds.), *Methodological and analytic frontiers in lexical research*, 57-80. Amsterdam & Philadelphia: John Benjamins. [reprint of 2010h]

Hamilton W. L., Leskovec J., and Jurafsky D. (2016) Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *ACL 2016*.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In *The Cambridge Handbook of English Historical Linguistics*, 36-53.

Harris Z. S. (1988). *Language and Information* . New York: Columbia University Press, ix, 120 pp. [Revised version of the Bampton Lectures given at Columbia University, New York City, in Oct. 1986.]

Kilgariff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004) The Sketch Engine. In: Williams G. and S. Vessier (eds.), *Proceedings of the XI Euralex International Congress,* July 6-10, 2004, Lorient, France, pp. 105-111.

Turney, P. D. & Pantel P. (2010), From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research* 37:141–188.

Traugott, E. C., & Trousdale, G. (2013). *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

Building a new Learner's dictionary for French: establishing the most relevant word lists

Laurent Catach
4H Conseil / Alphalivres
Laurent.catach@bbox.fr

This presentation describes several issues in the dictionary-making process, especially in establishing word lists, in the context of building a new Learner's dictionary of the French language, which is a project conducted by the author and currently in progress. This project aims to propose a brand new Learner's elementary dictionary, digital and print, for the large community of French speakers and learners worldwide as, curiously enough, quite few of modern dictionaries of this type exist for the French language. The project also aims to give access to very affordable printed resources for developing countries, especially for French-speaking countries in Africa, as dictionaries (and books in general) are still very difficult to get and are too expensive.

It is well-known that using limited word lists, usually built on frequency criteria, is a good basis for learning a foreign language, as well as learning one's own language at an elementary level. For French, early works were done in the 60s following Gougenheim's *Dictionnaire fondamental de la langue française* [Gougenheim1958], but are now out-of-date. Frequency lists are also the basis, for example, of the *Listes orthographiques de base du français* [Catach1984], of [Daffles2002], and more recently of D. Picoche's *Vocalire* method [Picoche2015]. We can also cite "official" word lists ([Brunet], [VOB], [MELS]) for primary levels, but unfortunately their differences are quite counter-productive. However, it seems that no recent work has been pursued in this direction using modern computational methods, especially corpora, and on the other hand learner's dictionaries for French have much less been developed by private publishers, compared to English.

In this context, the paper explores the methodology and the complexity of building relevant word lists for this type of dictionaries, using several methods and exploiting all the new possibilities of modern digital lexicography. This type of work, which is crucial in the preparation of a new pedagogical resource, involves studying and combining various frequency lists of words, corpora analysis, taking into account simple words as well as n-grams, compounds or phrases or most frequent inflected forms, examining regionalisms, recommended word lists for schools, thematic vocabularies, words for children, words by levels for learners, new words, headword lists of commercial dictionaries, words of elementary bilingual dictionaries, lexical databases, etc.

All in all, we have used more than 20 different resources and methods in the building of our word lists, and this process proves to be much more sophisticated than simply extracting the most frequent words out of a corpus — though frequency information on words, on several corpora of French, has obviously been exploited here.

We also emphasize, through the example of French, that very large word lists, such as proposed in the Wiktionary or various databases available on the web, are irrelevant for learners and children studying a new language. This kind of data, mainly oriented towards building large computational resources using automatic methods, produce much too large lists of words, which are far beyond the needs of learners, for which progressive methods for learning vocabulary are usually used.

The presentation will also present the current results of this work, which is a set of four different word lists of 3000, 7500, 12000 and 18000 items for the French language. All these word lists are managed in a database with many indexes, metadata and description criteria, so that subsets can easily be extracted to check for consistency, completeness or specific uses, such as obtaining the most important words by domains (e.g. health, animals, sports…), by part-of-speech (verbs, nouns…), by levels, etc.

As we have extensively studied many dictionaries, previous similar work in the field, as well as digital resources available on the web, we believe the work that is presented here is quite original for French. It can also be very useful to people interested in the general dictionary-making process, as it provides a methodology for combining different types of information and criteria on words, in order to produce relevant word lists for pedagogical use.

The presentation will also discuss some of the issues of building relevant word lists as a solid basis for building new other linguistic resources, which can be used in many applications. For example, one important aspect of the project is to contribute to the general making of new and up-to-date relevant educational resources for the French language, so that the project can have a direct impact on education. Also, the building of relevant word lists is crucial as a basis for the next step of the project, which is to build bilingual dictionaries, as most French speakers evolve in a bilingual or multilingual environment. Finally, we believe that these resources, which will be freely available through an open licence, will help publishers, in developing countries, to create new books and material about the French language, which will be very helpful for the development of education.

Keywords: learner's dictionaries, Dictionary-Making Process, French, word lists, frequency, corpora, education, Francophonie, Français langue étrangère (FLE)

References

Brunet = *Liste de fréquence lexicale: 1500 mots les plus fréquents de la langue française, constituée par Étienne Brunet,* publié par le Ministère de l'éducation nationale, disponible sur http://eduscol.education.fr.

Catach, N. (1984). *Les Listes orthographiques de base du français (LOB),* Nathan.

Gouogenheim, G., 1958. *Dictionnaire fondamental de la langue française,* Didier.

MELS = *Liste orthographique à l'usage des enseignantes et des enseignants du primaire,* publié par le ministère de l'Éducation, du Loisir et du Sport (MELS) du Québec, disponible sur www.education.gouv.qc.ca.

Rolland, J.-C., Picoche, J. (2015). *VOCALIRE Les 7500 mots essentiels du lexique français*.

Selva, T., Verlinde, S. Binon, J. (2002). *Le DAFLES, un nouveau dictionnaire électronique pour apprenants du français,* in Braasch, A. and Povlsen, C. (eds), *Proceedings of the tenth international congress Euralex 2002,* Copenhagen, CST. Vol. I, pp. 199-208.

VOB = *Vocabulaire orthographique de base,* publié par le Ministère de l'éducation de Belgique, disponible sur www.enseignons.be.

Lexicographer's Lacunas or How to Deal with Missing Dictionary Forms on the Example of Czech

Lucie Chlumská, Václav Cvrček, Dominika Kováříková, Jiří Milička, Michal Škrabal
Institute of the Czech National Corpus, Charles University
lucie.chlumska@ff.cuni.cz, vaclav.cvrcek@ff.cuni.cz, dominika.kovarikova@ff.cuni.cz, jiri.milicka@ff.cuni.cz, michal.skrabal@ff.cuni.cz

1 Introduction

Every dictionary has to start with a list of headwords or dictionary forms. Selecting items for the list is a complicated issue on its own; however, even after the lexicographer arrives at the final selection of candidates, a new issue arises: how to deal with words that do not have a traditional dictionary form documented in the data (e.g. infinitive in verbs or masculine in adjectives)? Should the lexicographer reconstruct the unattested word form?

In order to answer this question, addressed rather seldom and briefly in the literature (Wolski 1989, Schnorr 1991, Svensén 2009: 105–106; in Czech lexicography: cf. Čermák 1995, Filipec 1995, Kochová & Opavská 2016), this paper aims to describe the scope of this issue, to classify and interpret such headword lacunas from the perspective of Czech lexicography and suggest possible solutions based on large corpus data.

2 Traditional dictionary forms

There has been a strong tradition (not only) in Western lexicography to choose the representative form of a lexeme according to its word class, typically the singular nominative in nouns or the infinitive in verbs, although there is some variation especially in the verb class: e.g. Latin, Greek or Bulgarian dictionaries list the 1st person singular present tense, whereas the Hungarian or Macedonian tradition is to indicate the 3rd person singular. In any case, there will always be lexemes that lack the selected form due to different reasons (with negation being a specific issue, see Kováříková et al. 2012). Leaving aside the possibility to break with the well-established tradition by listing roots instead of whole lexemes (like some classical Arab dictionaries do, e.g. Seidensticker 2008), the lexicographer has to come up with a solution. The first step would be to identify these dictionary form lacunas.

3 Identifying lacunas and their types

When we look into a large representative corpus, such as SYN2015, a hundred-million-word corpus of contemporary written Czech, and extract frequent words (e.g. with more than 20 occurrences) without a representative form, we can get an idea of how likely the lexicographer is to face this issue. In our pilot study, we searched for nouns, adjectives and verbs and got more than 300 nouns with unattested singular nominative, almost 300 verbs with no attested infinitive and almost 1500 adjectives with no documented singular nominative (masculine animate in positive form). Considering that the frequency threshold for a large dictionary tends to be much lower (with the number of unattested forms possibly much higher), it certainly is a phenomenon worth studying.

After a preliminary analysis of the words, we arrived at a tentative **classification** of lacuna types, see Table 1. Table 1 indicates that the typology of lacunas differs greatly in the selected word classes (where 0 does not apply, + applies, ++ strongly applies), which suggests that the optimal solution might also differ depending on the part of speech.

Table 1: Lacuna types and their distribution in word classes

lacuna type		semantic incompatibility	restricted collocability		unknown (lack of data)
			terminology	idioms	
POS	nouns	0	++	++	+
	adjectives	++	+	+	0
	verbs	0	+	+	++

It is obvious that the classification is not a clear-cut one and some words may fall into more than one category. Also, some lacunas could be described as systemic (i.e. the traditional dictionary form cannot exist due to some reason, such as *vdaná* "married [woman]" in masculine), while others may be the result of a lack of data (i.e. they might appear in a larger corpus). There are two complementary **tests** to evaluate whether a word belongs to the former or the latter category:

- to calculate the expected frequency of the missing dictionary form from the frequency of the lemma and the proportion of the representative grammatical category (e.g. nominative for nouns);
- to use a much larger corpus with billions of tokens and verify whether the number of hits is still zero or unexpectedly low.

Based on the result of both methods, we can distinguish between the categories of unattested lacuna forms and formulate our recommendations for lexicographers accordingly.

4 Recommendations for lexicographers and further applications

When we return to our initial question of whether the lexicographer should reconstruct the missing dictionary form, the following suggestions could apply:

- If the word has restricted collocability due to its idiomatic nature or terminological function, the representative word should not be reconstructed as the chances of occurrence in real language are very low, if not outright zero;
- Similarly, if the lexeme shows semantic incompatibility with the desired dictionary form, this should not be reconstructed either;
- If the dictionary form is unattested due to lack of evidence, but has no other serious restrictions, the dictionary form may be reconstructed.

Based on our research, a list of such problematic words could be generated for Czech lexicographers, e.g. in a form of an online tool that would help look up such lacunas.

Keywords: lexicography, morphological paradigm, dictionary form, corpus evidence

Acknowledgements

This study was supported by the programme Progres Q08 "Czech National Corpus" implemented at the Faculty of Arts, Charles University and by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

Čermák, F. (1995). Překladová lexikografie. In F. Čermák & R. Blatná (eds.) *Manuál lexikografie.* Jinočany: H&H, pp. 230-248.

Filipec, J. (1995). Teorie a praxe jednojazyčného slovníku výkladového. In F. Čermák & R. Blatná (eds.) *Manuál lexikografie.* Jinočany: H&H, pp. 14-49.

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

Kováříková, D., Chlumská, L. & Cvrček, V. (2012). What belongs to a dictionary? The example of negation in Czech. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress.* Oslo: University of Oslo, p. 209.

Schnorr, V. (1991). Problems of Lemmatization in the Bilingual Dictionary. In F. J. Hausmann et al. (eds.) *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie, 3. Teilband.* Berlin – New York: Walter de Gruyter, pp. 2813-2817.

Seidensticker, T. (2008). Lexicography: Classical Arabic. In K. Versteegh et al. (eds.) *The Encyclopaedia of Arabic Language and Linguistics Vol. 3.* Leiden – Boston: Brill Academic, pp. 30-37.

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

Wolski, W. (1989). Das Lemma in texttheoretischer Sicht. In F. J. Hausmann et al. (eds.) *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie, 1. Teilband.* Berlin – New York: Walter de Gruyter, pp. 366-371.

Insights into the language of signposts

Anna Dziemianko
Adam Mickiewicz University
danna@ifa.amu.edu.pl

Signposts help users quickly find the right sense in long entries. Unfortunately, little is known about the significance of their linguistic form. Signposts can be synonyms of headwords, short definitions or paraphrases, typical subjects/objects/collocates, hyperonyms, field labels, indications of context, purpose or extralinguistic environment. Given their "'mixed-bag' nature" (DeCesaris 2012: 533), some researchers argue for more linguistic uniformity of signposts, while others – for their linguistic heterogeneity. The current study tests empirically whether linguistically homogeneous or heterogeneous signposts better serve dictionary users. It aims to determine which signposts, homogeneous or heterogeneous, are more beneficial to sense identification, language reception and production as well as immediate and delayed retention of meaning. The paper also investigates whether the usefulness of the type of signposting (homogeneous or heterogeneous) is dependent on the grammatical category of headwords.

To achieve the aims of the study, a main test and a post-test were conducted. The main test was centered around 12 common nouns and verbs, which were used in sentences in their less frequent senses. For each word, a six-sense entry was compiled on the basis of MLDs. Target sense positions were evenly distributed across the entries. To prevent participants from relying on their background knowledge, the words were replaced with much less frequent ones from *Weird and wonderful words*. Two versions of the main test were created, which differed only in signposts. In one version, signposts were homogeneous; in each entry, their grammatical category corresponded to the grammatical category of the headword. In noun entries, signposts were noun phrases, and in verb entries – verb phrases. Such signposts were also functionally homogeneous, as all of them were short definitions. In the other test version, signposts were heterogeneous; their grammatical category differed from the grammatical category of the respective headword. In noun entries, signposts were verb, adjective and prepositional phrases, and in verb entries – noun and prepositional phrases as well as *wh*-clauses. Their linguistic function was also varied. Heterogeneous signposts represented typical subjects/objects/collocates, field/domain labels, purpose and context. The linguistic form and function of heterogeneous signposts were strictly controlled across the entries. In the post-test, the target words were listed. No dictionary entries were included.

150 advanced learners of English (C1 in CEFR) took part in the experiment. Half of them did the test with homogeneous signposts in entries, and the other half took the one with heterogeneous signposting. For each test item, the subjects performed three tasks. First, they indicated the sense (one out of six) in which the target word was used in the supplied sentence (sense identification). Second, they explained in English or in their native language the meaning of the target word (reception). Third, they formulated an original sentence with the test item (production). Immediately after the main test, meaning retention was checked with the help of the post-test, in which the participants explained the target words from memory. The post-test was repeated after one week to check delayed retention. The sequence of items was changed to reduce learning effects.

The results of 2 (*signposts*) x 3 (*task*) x 2 (*POS*) ANOVA (with *signposts* as a between-groups factor, *task* and *POS* as within-group factors), indicate that the type of signposts had an important effect

on test results (F=43.55, p=0.00, partial η^2=0.813). The overall score obtained in the main test with the help of entries with heterogeneous signposts (67.02%) was significantly better than that based on reference to entries with homogeneous signposts (55.95%). The effect was dependent on *task*. A multivariate test for repeated measures revealed a statistically significant interaction between *task* and *signposts* (Wilks' λ=0.18, F=19.96, p=0.00, partial η^2=0.571). Entries with heterogeneous signposts were significantly more useful for sense identification (p=0.01) and reception (p=0.01, Tukey HSD). In production, the results obtained after reference to entries with homogeneous and heterogeneous signposts were comparable (p=0.84, Tukey HSD). The influence of signpost type on the scores for any task was not dependent on the part of speech. The main effect of POS and the interactions involving POS were not statistically significant.

Retention was significantly better when the subjects had consulted entries with homogeneous signposts (20.77%) than heterogeneous ones (8.43%) (F=40.63, p=0.00, partial η^2=0.802; 2 (*signposts*) x 2 (*task*) x 2 (*POS*) ANOVA). The statistically significant and positive effect of signpost homogeneity on retention was noted in immediate (p=0.04) and delayed post-tests (p=0.04, Tukey HSD), and was not dependent on the grammatical category of headwords.

The study reveals that heterogeneous signposts significantly support sense identification and reception, but the form of signposts has no effect on language production. However, homogeneous signposts stimulate both immediate and delayed meaning retention. The full version of the paper explores possible reasons for the observed effects of signpost types.

Keywords: signposts, language, learners' dictionaries

References

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

Bogaards, P. (1998). Scanning Long Entries in Learners' Dictionaries. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin, S. Theissen (eds.), *EURALEX'98 Proceedings*. Liège: University of Liège, pp. 555-563.

Cambridge Advanced Learners' Dictionary. (Fourth edition.) (CALD4) Accessed at: http://dictionary.cambridge.org/dictionary/learner-english/ [25/03/2018].

De Cock, S. & Granger, S. (2004). High Frequency Words: The Bête Noire of Lexicographers and Learners Alike. A Close Look at the Verb 'Make' in Five Monolingual Learners Dictionaries of English. In G. Williams, S. Vessier (eds.), *Proceedings of the 11th EURALEX International Congress*. Lorient: Université de Bretagne-Sud, pp. 233-243.

DeCesaris, J. (2012). On the Nature of Signposts. In R. Vatvedt Fjeld, J. M. Torjusen (eds.), *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 532-540.

Dziemianko, A.(2016). An insight into the visual presentation of signposts in English learners' dictionaries online. In *International Journal of Lexicography* 29(4), pp. 490-524.

Fabiszewski-Jaworski, M. (2011). Spontaneous Defining by Native Speakers of English. In K. Akasu, S. Uchida (eds.), *Asialex2011 Proceedings Lexicography: Theoretical and Practical Perspectives*. Kyoto: Asian Association for Lexicography, pp. 102-109.

Lew, R. (2010). Users Take Shortcuts: Navigating Dictionary Entries. In A. Dykstra, T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress*. Ljouwert: Afûk, pp. 1121-1132.

Lew, R. & Pajkowska, J. (2007). The Effect of Signposts on Access Speed and Lookup Task Success in Long and Short Entries. In *Horizontes de Lingüística Aplicada* 6(2), pp. 235-252.

Longman Dictionary of Contemporary English. (Fifth edition.) (LDOCE5) Accessed at: http://www.ldoceonline.com/ [25/03/2018].

Macmillan English Dictionary Online. (MEDO) Accessed at: http://www.macmillandictionary.com/ [25/03/2018].

Nesi, H. & Haill, R. (2002). A Study of Dictionary Use by International Students at a British University. In *International Journal of Lexicography* 15(4), pp. 277-305.

Nesi, H. & Tan, K. H. (2011). The Effect of Menus and Signposting on the Speed and Accuracy of Sense Selection. In *International Journal of Lexicography* 24(1), pp. 79-96.

Nuccorini, S. (1994). On Dictionary Misuse. In M. Willy, W. Meijs (eds.), *EURALEX'94: Papers Submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam*. Amsterdam: Vrije Universiteit, pp. 586-597.

Oxford Advanced Learner's Dictionary of Current English. (Ninth edition.) (OALD9) Accessed at: http://www.oxfordlearnersdictionaries.com/ [25/03/2018].

Ptasznik, B. (2015). *Signposts and Menus in Monolingual Dictionaries for Learners of English.* Poznań: Wydawnictwo UAM.

Ptasznik, B. & Lew, R. (2014). Do Menus Provide Added Value to Signposts in Print Monolingual Dictionary Entries? An Application of Linear Mixed-Effects Modelling in Dictionary User Research. In *International Journal of Lexicography* 27(3), pp. 241-258.

Scholfield, P. (1999). Dictionary Use in Reception. In *International Journal of Lexicography* 12(1), pp. 13-34.

The Corpus of Contemporary American English. Brigham Young University. (COCA) http://corpus.byu.edu/coca/ [25/03/2018].

Tono, Y. (1984). *On the Dictionary User's Reference Skills*. B.Ed. Dissertation, Tokyo: Gakugei University.

Tono, Y. (1997). Guide Word or Signpost? An Experimental Study on the Effect of Meaning Access Indexes in EFL Learners' Dictionaries. In *English Studies* 28, pp. 55-77.

Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Tübingen: Niemeyer.

Tono, Y. (2011). Application of Eye-tracking in EFL Learners' Dictionary Look-up Process Research. In *International Journal of Lexicography* 24(1), pp. 124-153.

Urata, K., Shimizu, A., Matsuyama, M & Nakao, K. (1999). An Analysis of *Longman Dictionary of Contemporary English*, Third Edition. In *Lexicon* 29, pp. 66-95.

Weird and Wonderful Words. Accessed at: https://en.oxforddictionaries.com/explore/weird-and-wonderful-words [25/03/2018].

Contrastive Collocation Analysis – a Comparison of Association Measures across Three Different Languages Using Dependency-Parsed Corpora

Stefan Evert[1], Thomas Proisl[1], Peter Uhrig[1], Maria Khokhlova[2]
[1] University of Erlangen-Nuremberg
[2] St. Petersburg State University
stefan.evert@fau.de, thomas.proisl@fau.de, peter.uhrig@fau.de, m.khokhlova@spbu.ru

Our analysis focusses on association measures for noun-adjective combinations of dependency-related co-occurrences. In the study we limit ourselves to binary collocations, i.e. pairs of two words, for English, German, and Russian. All three languages belong to the same language family (Indo-European); English and German share a longer common ancestry, being both Germanic languages. Typically different levels of attention are paid to them (most often English data tend to be more analyzed) The languages also differ both in their syntactic and their morphological nature. The Russian language is often quoted as a highly inflecting language demonstrating both synthetic and analytical features. Thus to the best of our knowledge the present work is the first study of this kind.

In our study we would like to answer the following question: what differences do we find between the languages concerning noun-adjective collocations extracted from syntactic dependencies? For this task we evaluate lists of relational collocation candidates extracted with the help of association measures and analyze whether the association measures perform the same across the different languages with respect to precision and recall. We also calculate the statistical correlations between the measures to investigate whether they are the same in the different languages.

From a historical point of view, German and English are more similar to each other than each of them is to Russian, so one could expect a similar behaviour when it comes to their collocational patterning. On the other hand, German and Russian are typologically somewhat similar in that they both have a (more or less) elaborate case system and tend to form compounds as single orthographic words (for Russian this is only true to a certain degree). We would thus expect to see bigger differences between German/Russian on the one hand and English on the other hand with regard to noun-adjective collocations.

We aim to use comparable gold standards that include examples extracted from lexicographic works for all three languages. The information about collocations can be presented in different resources (for examples, explanatory or specialized dictionaries, thesauri, wordnets, databases etc.). For example, we use Oxford Collocations Dictionary for Students of English (OCD2) as a gold standard for English. We confine ourselves to noun-adjective collocations as this type of phrases is well-defined in various lexicographic resources. We decided to use large web corpora that comprise billions of tokens and provide a high coverage of the gold standards, in particular DECOW16A, ENCOW16A (Schäfer & Bildhauer, 2012; Schäfer, 2015) and Araneum Russicum II Maximum (Benko, Zakharov, 2016). The corpora were annotated with state-of-the-art dependency parsers that enabled us to extract specific syntactic relations (such as verb+subject and verb+object). Following the approach of Evert & Krenn (2005), we rank the candidates using a wide range of statistical association measures (those evaluated by (Evert et al., 2017) on a smaller English gold standard) and evaluate collocation identification quality in terms of precision-recall graphs; average precision up to 50% recall (AP50) is used

to make quantitative comparisons. We evaluate both a global ranking of all candidates as in (Bartsch & Evert, 2014) and a separate ranking for each node as in (Uhrig & Proisl, 2012).

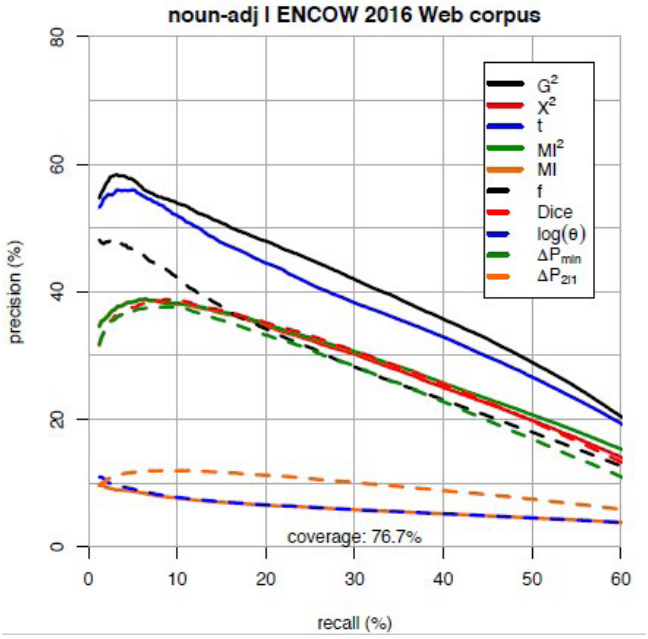


Fig. 1: Precision-recall graphs based on global ranking of all candidates for English verb/object collocations

The experiments on English noun-adjective collocations showed the following results (figure 1). We made a global ranking of all candidates with frequency threshold $f \geq 5$ comparing them to the gold standard from OCD2. The best overall measure is log-likelihood that outperforms the second best measure t-score. The coverage of 76.7% shows that the majority of all noun-adjective collocations from the OCD2 gold standard occur at least 5 times in the large Web corpus. The further experiments suggest that for English collocations log-likelihood proves to be the best measure for all types except adjective-verb collocations. For adjective-verb collocations MI4 gives better results, while subject-verb collocations appear to be particularly hard to identify. Our earlier experiments on Russian data have shown that t-score and Dice proved to extract the largest number of collocations that overlap with the data found in dictionaries (Khokhlova, 2017).

Keywords: collocation, evaluation, contrastive linguistics

References

OCD2 = *Oxford Collocations Dictionary for Students of English*, 2nd edition (2009). Edited by Colin MacIntosh. Oxford: Oxford University Press.

Bartsch, S., Evert, S. (2014). "Towards a Firthian notion of collocation." *OPAL – Online publizierte Arbeiten zur Linguistik* 2/2014: 48–61. Accessed at: http://pub.ids-mannheim.de/laufend/opal/pdf/opal2014-2.pdf [30/03/3018].

Benko, V., Zakharov, V. (2016). "Very Large Russian Corpora: New Opportunities and New Challenges". *Komp-juternaja lingvistika i intellektual'nyje technologii: Po materialam mezdunarodnoj konferencii «Dialog», 15 (22)*. Moskva: Rossijskij gosudarstvennyj gumanitarnyj universitet, pp. 79-93. Accessed at: http://www.dialog-21.ru/media/3383/benkovzakharovvp.pdf [30/03/3018].

Evert, S., Krenn, B.. (2005). "Using small random samples for the manual evaluation of statistical association measures". *Computer Speech and Language*, 19(4), pp. 45–466. https://doi.org/10.1016/j.csl.2005.02.005

Evert, S., Uhrig P., Bartsch S., Proisl, T. (2017). "E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification." In: *Proceedings of eLex 2017 – Electronic lexicography in the 21st century: Lexicography from Scratch*. Leiden: Lexical Computing, pp. 531–549. Accessed at: https://elex.link/elex2017/wp-content/uploads/2017/09/paper32.pdf [30/03/3018].

Khokhlova, M. (2017). On The Differences between Association Measures for Automatic Collocation Extraction: Evaluation against Dictionaries. In *SGEM International Multidisciplinary Scientific Conference on Social Sciences and Arts 2017*. Sofia. V. 2, pp. 887–892.

Schäfer, R. (2015). Processing and Querying Large Web Corpora with the COW14 Architecture. In: Bański, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, Andreas Witt (eds.) *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Mannheim: IDS Publication Server, 28–34. Accessed at: https://ids-pub.bsz-bw.de/files/3826/Schaefer_Processing_and_querying_large_web_corpora_2015.pdf [30/03/3018].

Schäfer, R., Bildhauer, F. (2012). "Building Large Corpora from the Web Using a New Efficient Tool Chain." In: Calzolari, N., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard B., Mariani J., Moreno A., Odijk, J., Piperidis, S. (eds.). (2012). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association, 486–493. Accessed at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/834_Paper.pdf [30/03/3018].

Uhrig, P., Proisl, T. (2012). "Less hay, more needles – using dependency-annotated corpora to provide lexicographers with more accurate lists of collocation candidates." *Lexicographica* 28, pp. 141–180. https://doi.org/10.1515/lexi.2012-0009.

Shades of word meanings in a Croatian dictionary based on literary citations

Ivana Filipović Petrović
Croatian Academy of Sciences
ivana.filipov@gmail.com

In the dictionary making process, one of the major tasks of lexicographers which is commonly stressed is identifying the distinct meanings of word forms, a process referred to as word sense disambiguation. However, the question of how lexicographers determine different, usually numbered, meanings of lexical items often stays unanswered. It is often said that lexicographers are somewhat shy of explaining their working methods on splitting and lumping senses or even that they are unaware of what they are doing, relying on their intuition which cannot be explained (Stock 1984: 131). Moreover, as Ayto (1988: 49) points out, metaphors make lexicographers especially nervous because they do not know what to do with them. The era of corpus evidence in lexicography, i.e. data becoming available in the form of machine-readable corpora, has prompted some questions and doubts about the very existence of word meanings and, consequently, the need for sense disambiguation in dictionary-making. The question of whether word meanings really exist has been asked by many great lexicographers since the 1980s as a result of studying corpus evidence (Kilgarriff 1997; Nida 1997; Hanks 2008). The answer is that they do exist, but only in context. According to Hanks (2008: 133), outside the context of a meaning event, words have meaning potentials, rather than just meaning. The meaning potential of a word comprises a number of components which may be activated cognitively by other words in the context in which it is used. It has been argued that if word senses do not exist, there is no use trying to disambiguate them. For decades, linguists and lexicographers have tried to prove this point with the example of the word *bank,* showing that in human communication, i.e. in actual usage, the misunderstanding whether the conversation is about a financial institution or an area of land along the side of a river hardly ever arises (Stock 1984; Cruse 2004; Atkins and Rundell 2008). Corpus-based studies (e.g. Hanks 2008) have shown that contextual clues disambiguate. However, in many uses neither of the two meanings of *bank* is fully activated and examples invoke one or the other of the main senses only to some extent. Therefore, the issue at hand is vagueness rather than ambiguity. All this presents a major challenge for lexicography, given that the very nature of dictionaries, with their linear organization, forces words to be considered as isolates (Moon 2008: 313). Perfectly aware of the fact that distinctiveness between senses of a lexical item is a flexible and context-based phenomenon (Geeraerts 2001: 2), lexicographers try to find ways to reconcile language in use and its reflection in the dictionary. The aim of this paper is to illustrate the process of identifying different word meanings and the lexicographic treatment of polysemous words in a descriptive dictionary of Croatian which is based on manually collected examples of literary works. Examples of usage were compiled by lexicographer Julije Benešić (1883-1957) from the literary works of Croatian writers who were active between the mid-19th and mid-20th centuries. Julije Benešić compiled the dictionary in the mid-20th century, i.e. in the pre-corpus era, when the lexicographer's intuition was the predominant criterion for what a good dictionary example is. Benešić's dictionary remained unpublished until 2008, when extensive work began to complete it. Given that the selection of examples limited to literary works strongly influences word sense disambiguation, we will show the steps and techniques in defining polysemous entries in Benešić's dictionary using the adjective *šaren* ('colorful'). Benešić chose 11 examples reflecting the use of *šaren*, which include the following expressions: *the old lady*

was wearing colorful clothing; they are putting colorful flags on the mast; misery is colorful as a bloody carnival; they are a very colorful group of people; his mother is a colorful snake and she has vicious plans with him. The key issue we are dealing with in our work is how many different meanings should be included based on examples of use. In order to provide a consistent and reliable lexicographic treatment of polysemous words, we employ the Pragglejazz group procedure for identifying metaphorically used words in discourse (Steen 2007; 2010). In addition, we will show the benefits of using this procedure in Benešić's dictionary in particular and in lexicography in general. This is especially important for improving and facilitating sense disambiguation in dictionary making.

Keywords: word sense disambiguation, lexicographic treatment of polysemous words, Croatian dictionary based on literary citations, Pragglejazz group procedure

References

Atkins, Sue B. and Rundell, Michael. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Ayto, John. 1988. Fig. leaves. Metaphor in dictionaries. In Snell-Hornby, M. (ed.). *ZüriLEX '86 Proceedings.* Tübingen: Francke. 49–54.

Cruse, D. A. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Second edition. Oxford: Oxford University Press.

Hanks, Patrick. 2008. Do words meanings exist? In Fontenelle, T. (ed.). *Practical Lexicography: A Reader.* Oxford: Oxford University Press.

Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities*, Kluwer Academic Publishers, 31 (2): 91–113.

Moon, Rosamund. 2008. Dictionaries and collocation. In Granger, Sylviane (ed.). *Phraseology: An interdisciplinary perspective.* Amsterdam: John Benjamins Publishing. 31–336.

Nida, Eugene A. 1997. The molecular level of lexical semantics. *International Journal of Lexicography*, 10 (4): 265–274.

Steen, Gerard et. al. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor & Symbol* 22 (1): 1-39.

Steen, Gerard et. al. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU.* Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Stock, Penelope. 1984. Polysemy. In Hartmann, R.R.K. (ed.), *LEXeter'83 Proceedings.* Tubingen: Niemeyer, 131–140.

An explicit and integrated intervention programme for training paper dictionary use in Greek primary school pupils

Zoe Gavriilidou
Democritus University of Thrace
zoegab@otenet.gr

A recent trend in dictionary research resulted in the birth of pedagogical lexicography and the subsequent shift of interest from the dictionary as a product to the dictionary use and users' reference skills or strategies. In response to this development, many claims have been made about the teachability of users' reference skills (Wingate, 2004; Lew & Callas, 2008) or strategies (Gqvriilidou, 2002; 2012; 2013). Dictionary use training may improve specific skills (lemmatization, look up, etc.) that lead to successful dictionary use, increase users' awareness about the variety of instances that different types of dictionaries could be used and the characteristics of all different dictionary types, and enhance overall strategic dictionary use profile of users.

Although there have been cases that dictionary use during specific tasks such as reading, writing or vocabulary learning has been questioned (Leffa, 1992; Aust et al. 1993; Taylor & Chan, 1994; Koga, 1995; Tang, 1997; Nesi 1999; Hill & Laufer, 2003), there is a consensus in relevant research that using dictionaries in language learning can be beneficial for the students (Tomaszczyk, 1979; Scholfield, 1982; Summers, 1988; Tono, 1988; Luppescu & Day, 1993; Knight, 1994; Fraser, 1999a; Γαβριηλίδου, 2002; Wingate, 2002, Nesi & Hail 2002, Chen, 2011). Based on this research, the revised *National Curriculum for Greek language teaching* in primary and lower secondary schools (Pedagogical Institute 2011) placed particular emphasis on dictionary use. Furthermore, acknowledging the positive effect of dictionary use in reading comprehension, writing, vocabulary learning or oral communication, new school dictionaries have been elaborated, and dictionary use exercises have been included in the new Textbooks for language teaching in Greece. However, only one Textbook, the one for Second Primary entitled *Ταξίδι στον κόσμο της γλώσσας* 'A trip in Language's world' (Gavriilidou, Beze & Sfyroera 2006) incorporated a whole unit of fourteen pages (plus another eight pages in the exercise book) which includes an intervention programme that aims to familiarize students at an early stage (second Elementary, children aged eight years old) with the conventions of their school dictionary and systematically train them in dictionary use.

The purpose of this paper is twofold:

a) to offer tentative guidelines for teachers who wish to provide dictionary users with systematic training on how to successfully use a dictionary, by presenting the characteristics of the above mentioned innovative intervention programme, and discussing its benefits to dictionary use,
b) provide a valid research methodology to researchers who wish to study the effect of training programmes to dictionary use.

To do so, first, we focus on the revised curriculum of Greek Language Teaching in Greek Primary and lower secondary schools and present briefly the school dictionaries released recently in Greece. Second, we comment on literature concerning methodological issues and the characteristics of the programme and more specifically, the significance of the intervention's duration, the importance of a justified choice of tasks for practicing dictionary use strategies, the measurement of dictionary strategy use before and after the instruction and finally the implementation of *explicit* and *integrated* strategic dictionary use training. Then, we present how the design of the specific intervention programme

responds to the cultivation of reference skills that enhance dictionary use found in previous literature. Finally, we present a research protocol for studying the effect of an intervention training programme to dictionary use.

This study contributes to the discussion about both teachability of strategic dictionary use and the effective forms of dictionary use training with integrative language skills. Bearing in mind the call for additional and more rigorous teacher-led intervention programmes for training dictionary use in different learning contexts around the world, we show how an *explicit* and *integrated* intervention programme can result in changes of users' habits during dictionary use. The evaluation criteria of such a programme could be the following: improvement in the dictionary use performance, strategy maintenance upon time and finally expansion of dictionary use in various cases.

Keywords: dictionary use strategies, pedagogical lexicography, dictionary use reference skills, strategic dictionary teaching

References

Aust, R., M. J., Kelley & W. Roby (1993). The use of hyper-reference and conventional dictionaries. *Educational Technology Research and Development*, 41 (4), 63-73.

Chen, Y. (2011). Dictionary Use and vocabulary learning in the context of reading. *International Journal of Lexicography*. December, 1-32.

Fraser, C. A. (1999a/c). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, 21, 225–241.

Gavriilidou, Z., M. Sfyroera & L. Beze (2006) (IN GREEK). *Ταξίδι στον Κόσμο της Γλώσσας* [A trip to language's world]*, Βιβλίο Γλώσσας Β' Δημοτικού*, Βιβλίο Μαθητή, ΟΕΔΒ, Ανάδοχος: Ελληνικά Γράμματα. Διαθέσιμο στο http://ebooks.edu.gr/modules/ebook/show.php/DSDIM-B105/691/4564,20745/ (Ανακτήθηκε στις 10/1/2017).

Gavriilidou, Z. (2002). (IN GREEK) Η διερεύνηση των λόγων χρήσης λεξικού ως προϋπόθεση για τη διδασκαλία στρατηγικής χρήσης του λεξικού στην τάξη [Investigation of dictionary use situations as a presrequisite for strategic dictionary teaching]. Στο P. Kambaki (Επιμ.), *Η διδασκαλία της νέας ελληνικής ως μητρικής γλώσσας (Teaching Greek as L1)*, (σσ. 45-60). Komotini.

Gavriilidou, Z. (2012). (IN GREEK)Παιδαγωγική Λεξικογραφία και στρατηγικές χρήσης λεξικού [pedagogical lexicography and dictionary use strategies]. Στο Καμπάκη-Βουγιουκλή, Π. & Μ. Δημάση, *Πρακτικά Συνεδρίου Ζητήματα Διδακτικής της Γλώσσας,* Εκδόσεις Κυριακίδη, 14-23.

Gavriilidou, Z. (2013). Development and validation of the strategy Inventory for Dictionary Use (S.I.D.U). *International Journal of Lexicography, 22*(2), pp. 135-154.

Hill, M. & B. Laufer (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition, *IRAL* 41 (2003), 87–106.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, 78, 286–298.

Koga, Y. (1995). The effectiveness of using an electronic dictionary in second language reading. *Bulletin of the Liberal Arts of Hiroshima University* Part 2, 44, 239-244.

Leffa, V. (1992). Making foreign language texts comprehensible for beginners: An experiment with an electronic glossary. *System*, 20(1), 63-73.

Lew, R., & Galas, K. (2008). Can dictionary use be taught? The effectiveness of lexicographic training for primary school level Polish learners of English. In E. a. Bernal (Ed.), *Proceedings of the XIII EURALEX International Congress* (pp. 1273-1285). Barcelona: Universitat Pompeu Fabra.

Luppescu S. & R. R. Day (1993). Reading dictionaries and vocabulary learning. *Language Learning,* 43(2), 263-287.

Nesi, H. (1999/1999a). A User's Guide to Electronic Dictionaries for Language Learners. *International Journal of Lexicography* 12(1), 55-66.

Nesi, H. & R. Haill (2002). A study of dictionary use by international students at a British University. *International Journal of Lexicography*, 15 (4), 277-305.

Scholfield, P. (1982). Using the English Dictionary for Comprehension. *Tesol Quarterly* 16(2), 185-194.

Summers, D. (1988). *The role of dictionaries in language learning* in: Ronald Carter-Michael McCarthy (eds.), 111-125.

Tang, G. (1997). Pocket electronic dictionaries for second language learning: help or hindrance. *TESL Canada Journal* 15 (1), 39–57.

Taylor, A. & A. Chan (1994). Pocket Electronic Dictionaries and their Use, In Willy Martin et al. (eds), *Proceedings of the 6th Euralex International Congress.* Amsterdam: Vrije Universiteit, 598–605.

Tomaszczyk, J. (1979). Dictionaries: users and uses. *Glottodidactica, 12*, σσ. 103 - 119.

Tono, Y. (1988). Can a Dictionary Help One Read Better? On the Relationship between E.F.L. Learners' Dictionary Reference Skills and Reading Comprehension. James, G. (Ed). 1988. *Lexicographers and Their Works*. Exeter Linguistic Studies 14: 192-200.

Wingate, U. (2004). Dictionary use - the need to teach strategies. *LANGUAGE LEARNING JOURNAL, 29*, pp. 5 - 11.

eTranslation TermBank: stimulating the collection of terminological resources for automated translation

Tatjana Gornostaja, Albina Auksoriūtė, Simon Dahlberg, Rickard Domeij, Marie van Dorrestein, Katja Hallberg, Lina Henriksen, Jelena Kallas, Simon Krek, Andis Lagzdiņš, Kelly Lilles, Asta Mitkevičienė, Sussi Olsen, Bolette Sandford Pedersen, Eglė Pesliakaitė, Claus Povlsen, Andraž Repar, Roberts Rozis, Gabriele Sauberer, Ágústa Thorbergsdóttir, Andrejs Vasiļjevs, Artūrs Vasiļevskis, Mari Vaus, Jolanta Zabarskaitė

Tilde, University of Copenhagen, Árni Magnússon Institute for Icelandic Studies, Institute of the Estonian Language, Jožef Stefan Institute, International Network for Terminology, Institute for Language and Folklore, Swedish Centre for Terminology, Institute of the Lithuanian language

tatjana.gornostaja@tilde.com

The abstract briefly describes the project "eTranslation Termbank" to be presented as a poster. The poster overviews the scope, goal, outcomes, expected impact of the project, and the project consortium. The eTranslation TermBank project provides terminological resources to the Connecting Europe Facility (CEF) eTranslation service[1] to improve the quality and coverage of automated translation. Under a terminological resource within the project, we define a set of monolingual, bilingual, or multilingual terminological metadata and data in a machine-readable form. The project aims to identify, collect, process, and provide new terminological resources for the CEF eTranslation service. Terminological resources in the three sector-specific domains covered by the project: health, legal (business legislation), and customer protection, which are not available via the ELRC-SHARE language resource repository[2] at the time, are considered new and are appropriate for the three sector-specific CEF Digital Service Infrastructures: eHealth[3], e-Justice/Business Registers Interconnection System[4], and Online Dispute Resolution[5]. The CEF eTranslation services are available for the following main target groups: European and national public administrations interested in integrating automated translation in their digital public services; for translators in European and public administrations interested in finding out how they can use existing and future CEF eTranslation services as a tool in their day-to-day work of producing high quality translations; other non-translating staff in European and national public administrations interested in finding out how the CEF eTranslation service can facilitate their day-to-day work and cross-border information exchange. The eTranslation Termbank project will deliver the following outcomes: 150 terminological resources for official languages of the European Union plus Norwegian and Icelandic; "Terminology for Europe" network for sustainable cooperation with the CEF eTranslation service;

1 See at https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

2 The European Language Resource Coordination repository (ELRC-SHARE) is intended for documenting, storing, browsing, and accessing language resources that are considered useful for feeding the Connecting Europe Facility Automated Translation platform, more information at https://elrc-share.eu/info.

3 See at https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/2017/05/30/eHealth and https://ec.europa.eu/cefdigital/wiki/display/CEFDSIS/eHealth+2.0.

4 See at https://beta.e-justice.europa.eu/?action=home&plang=en and https://beta.e-justice.europa.eu/106/EN/business_registers_in_eu_countries.

5 See at https://ec.europa.eu/consumers/odr/main/?event=main.home2.show&lng=EN.

pipeline of terminological resources flow to the CEF eTranslation service; directory of entities dealing with terminology on national levels; directory of existing terminological resources in the three sector-specific domains; methodology for preparation and application of terminological resources for the CEF eTranslation service.

Keywords: terminology, terminography, language resource, terminological resource, terminology management, translation, automated translation, digital public service

Towards the Improvement of the Treatment of Dialectal Headwords in the *Unabridged Dictionary of Standard Korean*

Song-I Han, Hae-Yun Jung
Kyungpook National University
flmh1416@hanmail.net, haeyun.jung.22@gmail.com

According to a survey conducted by 'Statistics Korea', the number of cross-national marriages in South Korea reached 151,608 by 2015. As a result, the need for dialect education has increased since dialects constitute strong identity markers in Korea. This study proposes to improve the treatment of dialectal words in the main Korean language dictionary, namely the *Unabridged Dictionary of Standard Korean* (hereafter, UDSK) so as to reflect daily practical language use. Currently, the examples provided by the UDSK are mainly extracted from literary works, which does not reflect the everyday language. Therefore, we propose to make use of a corpus that reflects the daily life language to meet the needs of a wider range of users. In order to achieve this, we focus on food-related vocabulary which we extracted from the Vocabulary Section of the annual 'Dialect Survey' carried out by the National Institute of the Korean Language. We propose a method for selecting and describing dialectal headwords in order to supplement the UDSK and fairly represent the richness of Korean dialects. The extraction of the dialectal forms and the study of distribution patterns were performed using the Korean Dialect Search Programme, which is part of the Korean Language Information Search Programme provided by the National Institute of the Korean Language.

The current preface of the UDSK explains that the macrostructure covers "standard words, North Korean words, dialectal words, archaic words, as well as frequent non-standard words". When considering dialectal headwords, it seems nonetheless that they are not fairly and consistently represented. Moreover, it does not mention on what criteria the dialectal headwords have been included to the macrostructure. If we look up a dialectal form, it is conventionally described as '(region name) dialect of [standard form]'. The following examples are typical descriptions of widely distributed dialectal words: *cey*3 is described as '(Gangwon, Gyeongsang, Jeolla, Chungbuk, Hamkyong) dialect of *kye* (chaff)' and *nasingi* as '(Gyeongnam, Chungbuk) dialect of *nayngi* (shepherd's purse)'. The standard form *nayngi* (capsella bursa-pastoris, commonly known as 'shepherd's purse') has 178 dialectal equivalents to but only a few of them are included in the UDSK. The forms *sinnayi, nasikeyi, hangkakkwu*, and many other dialectal variants of *nayngi* are not registered as such. In the case of *hangkakkwu* in particular, the UDSK did include it in its macrostructure, but as the dialectal form of thistle in Jeolla. However, *hangkakkwu* is also used as a dialectal form of *nayngi* in Hadong. The lack of consistency and accuracy of the UDSK treatment of dialectal words may create misunderstanding and confusion in learners of Korean, especially non-Korean wives for whom, according to Korean customs, cooking with their mothers-in-law is central to their daily life.

In the case of paper dictionaries, including all the dialectal forms would have been an impossible task. The UDSK has been digitalized and put online since 1999; however, these shortcomings have not been addressed. Our proposed solution is to create a dialectal network of related words wherein all the dialectal forms are searchable and linked to their standard forms and other variants if any. In 2017, an open dictionary called *Urimalsaem* has been made accessible to the general public. *Urimalsaem* includes for each headword a 'lexical map' in their microstructure, as seen in Figure 1.

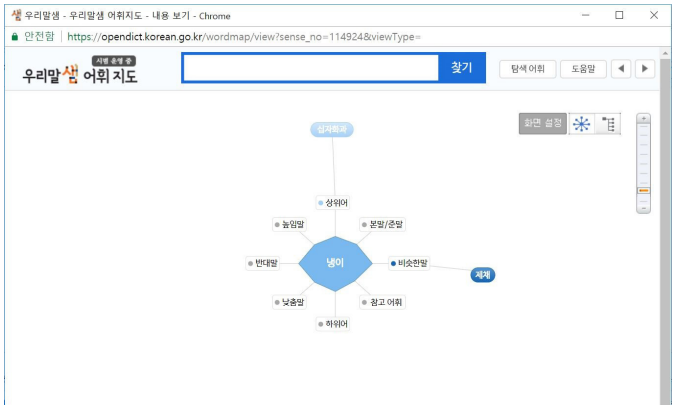


Figure 1. Lexical map for *nayngi* in *Urimalsaem*.

This lexical map consists of the headword at the core and its various lexical relations, namely 'hyperonym, hyponym, abbreviated form, similar words, antonym, honorific form, plain form, other related words', which are clickable and verifiable when they exist. Nonetheless, there is no link to 'dialectal forms'. In the case of *nayngi*, only the hyperonym and a similar word (name of a similar plant) are linked to it. In fact, *Urimalsaem* has been compiled on the basis of the UDSK's macrostructure and may present similar shortcomings, although it is being updated independently. As the National Institute of the Korean Language is also implemented this lexical map system on the online UDSK, we suggest including the standard/dialectal relation into the lexical map. In addition to the lexical map, we also propose to include in the microstructure of dialectal forms and their related standard forms a map of dialectal usage that visualises the actual spatial uses of the dialectal forms (Figure 2) and should be placed in the 'pragmatic information'.

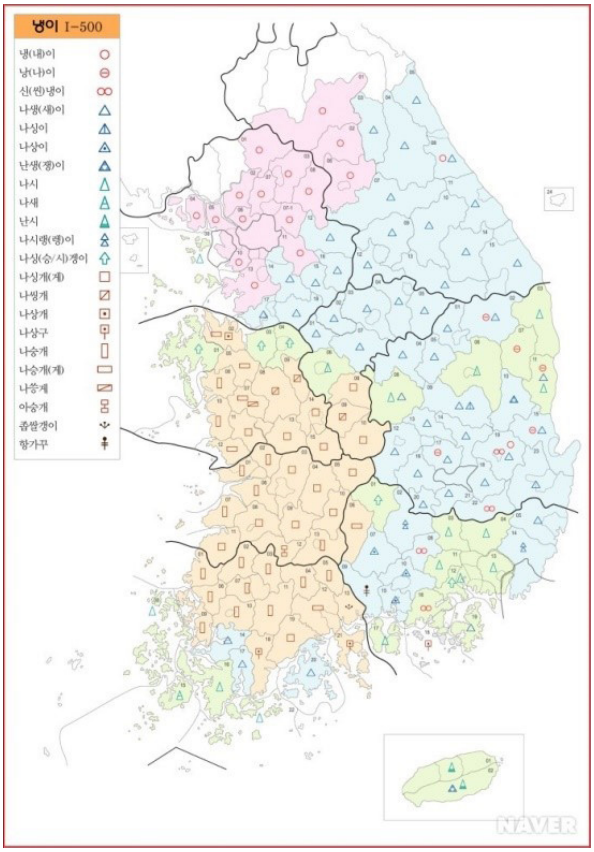


Figure 2. Dialectal map of *nayngi*.

While programmes exist to create such maps, they are for now only used for academic research purposes. But we believe that the implementation of such material in the dictionary will not only help learners of Korean, but also contribute to reflect and preserve linguistic diversity, thereby constituting an important milestone in the development of lexicography.

Keywords: dialect, dialectal variants, dialectal map, lexical map, headword, microstructure, food vocabulary

References

National Institute of the Korean Language, www.korean.go.kr
Statistics Korea, http://www.index.go.kr/main.do
Unabridged Dictionary of Standard Korean, stdweb2.korean.go.kr
Urimalsaem, https://opendict.korean.go.kr

Does phraseology determine meaning?

Patrick Hanks
Research Institute for Information and Language Processing
University of Wolverhampton
Wolverhampton, WV1 1LY, UK
patrick.w.hanks@gmail.com

Lexicography is currently in the doldrums, due in part to the collapse of traditional business models for funding lexicographical research, which used to be dependent on predicted sales of printed books (i.e. dictionaries). If lexicography is to have a future, it needs to show its usefulness in new domains, notably language teaching and computational linguistics. To do this successfully, it will have to develop radical new approaches to accounting for words and meanings. This paper investigates one such area of potential usefulness, namely the relationship between phraseology and meaning.

Generative linguists such as Beth Levin (1993, 2005) have long argued, or rather assumed, that each word has a meaning and that this meaning determines its possible uses. Lexicographers, on the other hand, examine usage in order to find out what a word means. In recent years, examination of usage has been greatly facilitated by the advent of large electronic corpora. The prediction of pre-corpus systemic functional linguists such as Firth (1957) and Sinclair (1966) that word usage is highly patterned have been borne out by innumerable detailed studies of usage. Do such studies support the conclusion that meaning determines possible phraseology? Or do they instead support the conclusion that phraseology determines meaning? The relationship between meaning and usage raises theoretical questions that are of great concern to lexicographers and empirical linguists alike. There is clearly a relationship between meaning and patterns of usage, but the directionality is uncertain. What determines what?

The bulk of the paper will be taken up (insofar as time allows) with discussion of some theoretical questions and implications raised by findings of the procedure called corpus pattern analysis (Hanks 2004, 2013). Examples include:

- The role of domain in determining possible meaning: e.g., is a lawyer *filing* a lawsuit doing the same thing as an admin assistant *filing* papers in a filing cabinet? And what about a journalist *filing* a story or a pilot *filing* a flight plan?
- The role of subargumental clues in determining (or selecting) meaning. Consider the ambiguity of he took his place; contrast the unambiguous statement *She took his place*. What about other collocations of take + place?
- When is ellipsis possible? If you can say "He fired" and mean "He fired a bullet from a gun", why can't you say "The university fired" and mean "The university fired some employees from their jobs"?
- What is the relationship between verbs and nouns in making meanings?

A preliminary discussion of the relationship between nouns and verbs in making meanings can be found in Hanks (2012).

Among the tentative conclusions of the paper are that meaning does not determine usage, but nor does usage determine meaning. Instead, the relationship is symbiotic. A person who engages in linguistic behaviour chooses to make active use of patterns of usage (and sometimes to exploit them in unusual

ways), in order to realize some sort of communicative and/or social intention. Central to this enterprise is reliance on (or exploitation of) stereotypes of usage patterns in order to realize stereotypical (and sometimes, original newly created) meanings. It would be useful to compare the stereotype theory of Hilary Putnam (1975) with the prototype theory of Eleanor Rosch (1978). Are they compatible or, perhaps, even identical?

To understand in detail how meaning works, it will be necessary to map prototypical meanings onto prototypical patterns of usage. A product of such a mapping would be lexicographical: an inventory of stereotypical patterns of usage, each of which would be associated with a meaning (presuppositions and entailments) – or a translation into a corresponding idiomatic pattern in a foreign language. Such an inventory would serve as a basis for pattern matching, of the kind proposed by Yorick Wilks as long ago as 1973. Wilks's work offers useful guidance for interpreting texts, in the form of the aphorism "Best match wins". Wilks's problem in 1973 was that no inventory of patterns existed for any language. Sadly, this is still the case. But the difference is now we have sufficient evidence, in the form of large corpora, to enable people to build such inventories, for many languages. It is up to lexicographers to build them, and up to rich software houses (or foundations, or research funding agencies) to find them.

Keywords: meaning, phraseology, corpus pattern analysis, lexicography of the future

References

Firth, J. R. (1957): *Papers in Linguistics 1934–1951*. Oxford University Press.

Hanks, Patrick (2004): 'Corpus pattern analysis'. In Geoffrey Williams and Sandra Vessier (eds.), *Euralex Proceedings*. Université de Bretagne-Sud, Lorient, France.

Hanks, Patrick (2012): 'How people use words to make meanings: Semantic types meet valencies." In Alex Boulton and James Thomas (eds), *Input, Process and Product: Developments in Teaching and Language Corpora*. Masaryk University Press.

Hanks, Patrick (2013): *Lexical Analysis: Norms and Exploitations*. MIT Press.

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

Levin, Beth, and M. Rappaport Hovav (2005). *Argument Realization*. Cambridge University Press,

Putnam, Hilary (1975): 'Is semantics possible?' and 'The meaning of "meaning"'. In *Mind, Language, and Reality: Philosophical Papers.* Cambridge University Press.

Rosch, Eleanor, and Barbara Lloyd (eds: 1978): *Cognition and Categorization.* Lawrence Erlbaum Associates.

Sinclair, John M. (1966): 'Beginning the study of lexis'. In C. E. Bazell et al. (eds.), *In Memory of J. R. Firth*. Longman.

Wilks, Yorick (1973): 'Preference semantics'. In Edward Keenan (ed.), *The Formal Semantics of Natural Language.* Cambridge University Press.

Building a sign language corpus – Problems and challenges The Danish Sign Language Corpus and Dictionary

Jette Hedegaard Kristoffersen, Thomas Troelsgård
UCC
JEHK@ucc.dk, ttro@ucc.dk

Corpora of spoken and written languages often consist of a large amount of digitised text, e.g. books, magazines, or newspapers. Sign languages have no standard written representation, at least none that is easily written and read by laymen, and hence – like it is the case for spoken language corpora – sign language corpora consist of (video) recordings supplied with transcriptions. But whereas the transcription of spoken language to a considerable extent can be performed through rendering uttered words through words of a written language, transcription of signed languages is further complicated by the lack of a standard written language. Furthermore, specialised video tools are needed for performing the transcription and utilising the corpus. Mainly for these reasons, building a sign language corpus of an adequate size and accuracy is a cumbersome process, and larger corpora have emerged only in the last decade, e.g. for the sign languages of Australia (Johnston, 2008), the Netherlands *(Corpus NGT*), Sweden (*Swedish Sign Language Corpus Project*), Great Britain (*The British Sign Language Corpus*), Poland (*PJM (PSL) corpus*), and Germany (*DGS-Corpus*).

When building a sign language corpus, the segmentation of the sign video into intervals that each holds a sign is an essential task. Differences in the major phonological categories (handshape, orientation, place of articulation, movement) of two adjacent signs result in transitional movements that can make it difficult to pinpoint where one sign ends and the next one begins, just as you rarely find a sign token in natural signing that perfectly resembles the citation form of its sign lemma. Furthermore, just as in spoken language, assimilation is a common feature of sign language, where e.g. a handshape can change under influence of the preceding or following handshape. Thus, a thorough description of the segmentation principles is a prerequisite for achieving an adequately consistent transcription. For this purpose we described a set of rules based on the "hold-movement-hold" model as described by Scott Liddell and Robert Johnson (Johnson & Liddell, 2011), and inspired by the principles used at the German Sign Language Corpus Project (Hanke et. al., 2012).

Another major challenge is to secure an unambiguous lemmatisation. Some corpus projects, e.g. the German Sign Language Corpus, include a detailed formal description of the sign form, e.g. in HamNoSys, a phonetic transcription system for sign languages, developed at the University of Hamburg (Hanke, 2004). Other projects represent a sign solely through a gloss – typically a word from the surrounding spoken language, chosen as a mnemonic because it captures (one of) the core meaning(s) of the sign. For the Danish Sign Language Corpus Project we chose to use only glosses because of limited resources. Like in many other sign language corpus projects, we provide glosses with affixes that signify additional information of the actual token, e.g. variation, modification, meaning in the context etc. In order to secure a high degree of consistency in the glossing, we started by building a controlled vocabulary consisting of the 2.200 lemmas of the Danish Sign Language Dictionary (*Ordbog over Dansk Tegnsprog*, cf. Kristoffersen and Troelsgård, 2012) and of about 5.000 additional signs from our collection of signs not yet described in the dictionary. Additionally, we supplied Danish equivalents for all signs in the vocabulary – partly from the dictionary and the sign collection, partly through linking to synonyms via the Danish wordnet (*DanNet*) – in order to facilitate the best

possibilities of finding the sought-for sign. Both features – building a controlled vocabulary of types, and linking of types to equivalents – are native parts of the tool we have chosen to use for our work, iLex (cf. Hanke and Storz, 2008), a system that is developed at the University of Hamburg, and which has kindly been placed at our disposal by the developers. Before starting the actual lemmatisation process, we associated a studio recording of each sign (from the dictionary and the sign collection) to the corresponding type. During the lemmatisation process, new signs are added to the vocabulary if no existing sign type matches the token in question, and the type is then associated with the token (which will be shown as video evidence for the new sign).

The current stage of the Danish Sign Language Corpus Project aims to perform a basic transcription of 154 monologues. Table 1 shows the current project progress. The time consumption is about 1:40 for segmentation, and 1:100 for lemmatisation, excluding error-correction, proofreading, breaks etc. Figure 1 and 2 show examples of the two steps of the transcription.

Table 1. Transcription progress of the first group of recordings, 154 monologues.

	Recordings	Hours	Percent
Goal	154	14:30	100,0
Segmented	84	07:56	54,8
Lemmatised	13	01:06	7,6

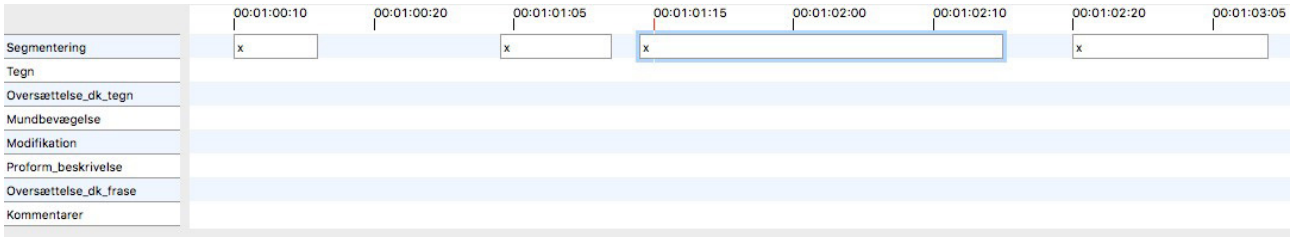


Figure 1. Transcription, step 1, segmentation. Each segment is defined by its timecodes, and tagged with a placeholder "x".

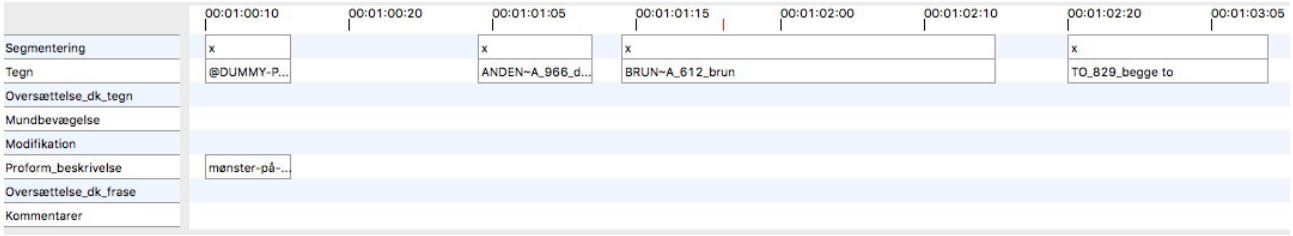


Figure 2. Transcription, step 2, lemmatisation. Each sign is assigned to a sign type, or to a dummy type for later analysis.

With the emergence of carefully transcribed corpora, sign language lexicographers gain prerequisites similar to lexicographers working with written/spoken languages. Although sign language corpora are considerably smaller than corpora of written/spoken languages, sign language dictionaries will then start to catch up with written language dictionaries with respect to quality, size, and range of information types.

Keywords: sign language corpus, sign language dictionary, corpus transcription, Danish sign language

References

Auslan Corpus. Accessed at: http://www.auslan.org.au/about/corpus [28/11/2017]

Corpus NGT. Accessed at http://www.ru.nl/corpusngtuk [28/11/2017]

DanNet [The Danish wordnet]. Accessed at: http://wordnet.dk [28/11/2017]

DGS-Corpus [German Sign Language Corpus]. Accessed at: http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html [28/11/2017]

Hanke, T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In: O. Streiter, C. Vettori (eds.) From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication. *Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon*. Paris: ELRA, pp. 1-6.

Hanke, T. et al. (2012). Where Does a Sign Start and End? Segmentation of Continuous Signing. In Crasborn, O. Efthimiou, E. et al (eds): *LREC 2012 Workshop Proceedings. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon.* Paris: ELRA, pp. 69-74

Hanke,T. & Storz, J. (2008). iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In Crasborn, O. Efthimiou, E. et al (eds): *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: ELRA, pp. 64-67.

Johnson, R.E. & Liddell, S.K. (2011). A Segmental Framework for Representing Signs Phonetically. In *Sign Language Studies* 11(3), pp. 408-463.

Johnston, T. (2008). Creating a Corpus of Auslan within an Australian National Corpus. In Haugh et al. (eds.) *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*. Macquaire University: Cascadilla Proceedings Project, pp 87–95.

Kristoffersen, J. H. & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language dictionary. In S. Granger & M. Paquot (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-318.

Ordbog over Dansk Tegnsprog [The Danish Sign Language Dictionary].Accessed at http://www.tegnsprog.dk [28/11/2017]

PJM (PSL) corpus [Polish Sign Language Corpus]. Accessed at: http://www.plm.uw.edu.pl/en/node/241 [28/11/2017]

Swedish Sign Language Corpus Project. Accessed at : http://www.ling.su.se/english/research/research-projects/sign-language/swedish-sign-language-corpus-project-1.59270 [28/11/2017]

The British Sign Language Corpus. Accessed at: http://www.bslcorpusproject.org [28/11/2017]

Osservatorio degli italianismi nel mondo (OIM) – a digital resource for surveying Italian loanwords in the world's languages

Matthias Heinz, Lucilla Pizzoli
Universität Salzburg, Università degli Studi Internazionali di Roma
matthias.heinz@sbg.ac.at, lucilla.pizzoli@unint.eu

The paper presents a large-scale digital humanities resource, the *Osservatorio degli italianismi nel mondo* (OIM), which aims at an extensive documentation of the significant lexical impact of Italian as a donor language worldwide in an ongoing international project hosted by the Accademia della Crusca, Florence (cf. Heinz/Gärtig 2014, Heinz 2017, Pizzoli 2017). Like only very few other languages Italian has a long-standing history as a culturally rich and diverse source for borrowings; besides contact-induced borrowings due to geographical proximity, Italianisms are especially frequent in many culturally relevant domains (e.g. musical terminology) in the languages of Europe and beyond.

The lexicographical data become accessible through an online database integrating language contact dictionaries of Italianisms in a growing number of languages. The core of the database draws on the lexicographical corpus gathered for the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT), whose collections of Italian loanwords in French, English and German are currently available online, while the release of data for Spanish, Portuguese, Catalan, Polish and Hungarian (with Maltese closely following up) on a new digital platform is imminent. For the latter and further (many of which Non-European) languages data are made available from a preceding survey project on Italianisms (cf. Serianni 2017); at the same time the revision and refinement of existing core data is underway.

From a metalexicographical point of view the information programme of this resource contains some innovations. Its microstructure aims not only at describing lexicogrammatical and semantic evolutions from both a diachronic and synchronic perspective but also at classifying each lemma according to type of borrowing and at evaluating the actual vitality of a loan within a given language. The OIM search engine allows for the application of filters in order to determine the lexical strength of Italian as a source language in a given language variety (overall and according to lexical sectors). Information about loan vitality assumes a central importance for measuring the penetration of Italian into different areas and the persistence of the influence of Italian culture in a diachronic perspective. In order to measure uniformly the current spread of loans a framework for assessing loan vitality is being developed, using corpora of texts in different discourse traditions (both literary and non literary texts). By assigning a progressive number to each certified entry, a "density index" can be obtained: this tool can easily reveal the impact of the Italian language in the host language.

The paper will highlight the major challenges, guidelines and solutions for implementing the macro- and microstructure in the process of preparing the source material for the database, while also pointing out further steps towards incorporating an ever growing body of target languages in this collaborative research endeavour.

Keywords: language contact, loanword dictionary, Italian, digital lexicography, metalexicography

References

DIFIT = Stammerjohann, Harro et al. (2008): *Dizionario di italianismi in francese, inglese, tedesco*. Firenze: Accademia della Crusca.

OIM = *Osservatorio degli italianismi nel mondo* (http://www.italianismi.org), progetto dell'Accademia della Crusca, coord. Luca Serianni, Matthias Heinz.

Heinz, Matthias (2017): "Dal DIFIT all'OIM: sfide lessicografiche e prospettive di implementazione", in: id. (a cura di): *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti*. Atti dell'incontro OIM (Firenze, 20 giugno 2014), Firenze: Accademia della Crusca, 21-38.

Heinz, Matthias/Gärtig, Anne-Kathrin (2014): "What a multilingual loanword dictionary can be used for: searching the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT)", in: *Proceedings of the XVI EURALEX Congress: The User in Focus*. Bolzano/Bozen 15-19 July 2014, eds. Andrea Abel, Chiara Vettori, Natascia Ralli, 1099-1107.

Pizzoli, Lucilla (2017): "Per un dizionario degli italianismi nel mondo: rilancio di un progetto". *Testi e linguaggi* 11, 171-182.

Serianni, Luca (2017): "L'italiano nel mondo. Intenti e propositi di un progetto editoriale sugli italianismi", in: Matthias Heinz (a cura di): *Osservatorio degli italianismi nel mondo: punti di partenza e nuovi orizzonti*. Atti dell'incontro OIM (Firenze, 20 giugno 2014), Firenze: Accademia della Crusca, 39-54.

Dynamically generated content in digital dictionaries: use cases

Holger Hvelplund
IDM
hvelplund@idm.fr

More than ever – due to changes in Google algorithm and the growth in traffic to social platforms – website traffic is generated by content. Users now typically land directly on content pages and less and less on general pages. It is putting the focus on three rather recent key quality digital publishing dimensions: (1) increase traffic by increasing amount of authoritative curated content on existing topics; (2) expand shares of traffic by adding authoritative, curated content on new topics; and (3) speed – i.e. how quickly you can significantly grow one or both of the first dimensions.

Addressing the three dimensions with traditional dictionary content production workflows and at a marginal cost is not possible. In this paper we present an alternative solution based on: (1) extracting content from constantly updated monolingual and parallel corpora; (2) enhancing the content in close collaboration with loyal end-users.

The use cases we present illustrate: (1) how dictionary entries are automatically and constantly updated with 'good' example sentences from constantly updated monolingual and parallel corpora; (2) how to make clear distinction between editorially curated content and dynamically generated content; (3) how end users can suggest headwords for new entries (neologisms), suggest definitions or translations for entries extracted from word lists and provide feedback on dynamically extracted example sentences; (5) how lexicographers can promote dynamically generated content to editorially curated content; and (6) how content specialists can create and update filters used for corpus extraction of 'good' example sentences.

Keywords: digital publishing, dynamically generated content, the dictionary-making process, user-generated content, lexicography and corpus linguistics, neologisms, crowd sourcing, SEO, CMS

References

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3). 209-226.

Jakubíček, M., A. Kilgarri , A., Kovář, V., Rychlý, P., Suchomel, V. (2013). The TenTen Corpus Family. Proc. Int. Conf. on Corpus Linguistics, Lancaster, UK.

Holger Hvelplund, H., Kilgarriff, A., Lannoy, V., White, P. (2013). Augmenting online dictionary entries with corpus data for Search Engine Optimisation. *Proceedings of the eLex 2013 conference, Tallinn, Estonia.*

Kilgarriff, A., Suchomel, V. (2013). Web Spam. *Proc. 8th Web as Corpus Workshop* (WAC-8), Lancaster, UK.

Lannoy, V. (2010). The IDM Free Online Platform for Dictionary Publishers. *Proc. Euralex*, Leeuwarden, Netherlands.

Practical Post-Editing Lexicography with Lexonomy and Sketch Engine

Milos Jakubicek, Michal Měchura, Vojtech Kovar, Pavel Rychly
Lexical Computing, Natural Language Processing Centre, Masaryk University, Faculty of Informatics, Masaryk University
milos.jakubicek@sketchengine.co.uk, valselob@gmail.com, xkovar3@fi.muni.cz, pary@fi.muni.cz

In this paper we present an implemented solution for post-editing dictionaries within the Lexonomy dictionary writing system interconnected with the Sketch Engine corpus management system. We follow up on the work on automatic dictionary drafting ("One-Click Dictionary") and assume a dictionary draft to be post-edited, or an existing (edited) dictionary to be extended are in place. We cover focus on features that enable lexicographers to obtain relevant corpus evidence or corpus-driven analysis in a user-friendly way from within the environment of the dictionary writing system and thus speed up their workflow. We exemplify the usage scenario in an ongoing project on Danish-English-Korean bilingual dictionary.

Introduction

This paper focuses on ongoing trends in automatizing the lexicographic processes with the help of advanced natural language processing techniques applied on top of large text corpora. By now corpora are at hands of lexicographers for over two decades, representing sources of empirical evidence. The influences were mutually beneficial – lexicography has significantly contributed to corpus development as well as to advances in corpus linguistics. While at the beginning corpora were primarily used for manual inspection of data by lexicographers, many advances in natural language processing allow for extended automation of the lexicographic process where lexical information is first automatically obtained from corpora and later post-edited by lexicographers.

We first briefly describe the two systems where the presented techniques are implemented – Sketch Engine corpus management system and Lexonomy dictionary writing system – and then show how they are used for creating initial dictionary drafts automatically and/or post-editing of existing dictionaries and dictionary drafts.

Sketch Engine

Sketch Engine (Kilgarriff et al., 2014) is a leading corpus management system primarily designed for lexicographic purposes. It is web-based and provides access for several hundreds of corpora for (as of December 2017) over 90 languages. It has been used for lexicographic projects for dozens of languages[1] and implements several analytic functions useful for dictionary drafting. These include single- and multiword retrieval, automatic extraction of good dictionary examples (GDEX), automatic extraction of definitions (GDEF) or collocational analysis of words' behaviour known as "word sketches" which serves as a backbone for obtaining a distributional thesaurus or word sense clustering.

1 For an overview please see https://www.sketchengine.co.uk/bibliography-of-sketch-engine

Sketch Engine has a separate corpus building part which allows users to build corpora from their own texts, or from texts automatically retrieved from the web according to their specifications taking the form of seed words which are passed to a search engine to find relevant online results (the so called WebBootcat approach). User corpora can be subject to automatic terminology extraction in order to retrieve domain-specific lexical items suitable for headword list generation.

Lexonomy

Lexonomy (Měchura, 2017) is a simple web-based dictionary writing system integrating a dictionary publishing platform. It allows its users to devise custom dictionary structures (XML templates) using a simple graphical editor and subsequently import entries or manually create and edit new ones. Lexonomy is interconnected with Sketch Engine using two interaction modes: the push model and the pull model.

In the push model, Sketch Engine pushes an initial dictionary draft fully automatically into a new dictionary in Lexonomy using its API access. In the pull model, users of Lexonomy retrieve different entry parts such as corpus examples from Sketch Engine (again, using its API access) without leaving the Lexonomy interface.

One-Click Dictionary

The push model was exploited to implement a One-Click Dictionary feature in Sketch Engine (see Jakubíček et al., 2017) which exports a dictionary draft based on a corpus selected by the user. Depending on the annotation available for the respective corpus, the export entails following features:

- headword list generation
- collocation extraction
- word sense clustering
- example sentences
- definitions
- thesaurus

While some of these functions (e.g. collocation extraction) achieve very high accuracy, the entry still needs to be manually post-edited. In fact, efficient implementation of the post-editing is key to the success of this approach: post-editing must never be more time consuming for the lexicographer than writing the entry from scratch.

Post-Editing in Lexonomy

The post-editing features in Lexonomy facilitate the pull model for interacting with Sketch Engine. They are usable for dictionaries that were initially drafted automatically from Sketch Engine as well as any other ones present in Lexonomy, regardless whether they were imported or manually created. They include the following operations:

- evaluation of lemma coverage against a corpus and retrieval of additional headwords
- retrieval of examples sentences using GDEX
- retrieval of definitions using GDEF

- retrieval of collocations
- word sense clustering
- retrieval of thesaurus items

Each of these steps contains an interface where editors can easily accept or reject the data as retrieved from Sketch Engine; or post-edit it (including lumping and splitting of word senses).

Use case: Danish-English-Korean

The post-editing workflow as described is at the moment exploited in a commercial project for creating a Danish to English and Korean bilingual dictionary. In the final paper we will provide feedback from editors and report on overall success of this approach within the project.

Conclusions

In this paper we show a practical implementation of the post-editing approach based on Sketch Engine and Lexonomy. The main goal of this strategy is to foster corpus use and ease the access to corpus evidence while advancing the automation of the dictionary creation process and speed up editors' work. We discuss the post-editing workflow of particular entry parts as well as experience from an existing project on Danish-English-Korean bilingual dictionary.

Keywords: corpora, post-editing, Lexonomy, Sketch Engine

SWRL your lexicon: adding inflectional rules to a LOD

Fahad Khan[1], Andrea Bellandi[1], Francesca Frontini[2], Monica Monachini[1]
Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa, Italy
PRAXILING UMR 5267 Univ Paul Valéry Montpellier 3 & CNRS - Montpellier, France
name.surname@ilc.cnr.it, francesca.frontini@univ.montp3.fr

Over the past few years the publication of lexical resources as Linked Data (LD) has taken on ever greater significance within the field of computational lexicography. So far the efforts of the community have been largely directed towards the definition of standards[1] and the conversion of single resources (see McCrae et al 2012, Khan et al 2016), but with less of a focus on the technical possibilities afforded by this new mode of publishing lexical data. However, the fact is that the Semantic Web gives us access to a whole ecosystem of standards, languages, and technologies. In this paper we will look at one of these languages, the Semantic Web Rule Language[2] (SWRL) and explore whether it might potentially play a useful role in the publication of lexical resources.

SWRL provides an extension of the Web Ontology Language (OWL) with Horn-like clauses that allows users to overcome some of OWL's expressive limitations. Even though there is a long tradition of using rule languages in computational linguistics, previous work on the use of SWRL in this domain seems to be thin on the ground (although see Wilcock 2007). And so one of the main aims of the work presented here is a better understanding of the value of including SWRL rules in an LD lexicon. Our case study concerns the morphological layer of the wide-coverage Italian lexicon Parole Simple Clips[3] (PSC-M), which provides inflectional information and morphological analysis of thousands of Italian words. What is important for our purposes here is that in addition to listing all the possible inflected forms for each word, PSC-M also contains inflectional rules which can be applied to derive these forms from the lemma. In Figure 1(a), we can see one such rule represented in LMF (Francopoulo ed. 2013).

```
<TransformSet>
  <Process>
    <feat att="operator" val="remove"/>
    <feat att="string" val="4"/>
  </Process>
  <Process>
    <feat att="operator" val="add"/>
    <feat att="string" val="IAMO"/>
  </Process>
  <GrammaticalFeatures>
    <feat att="morphofeat" val="P1IP"/>
  </GrammaticalFeatures>
</TransformSet>
```

(a)

```
hasVerbClass(?x, Class300) ^
hasStem1(?x, ?y) ^
swrlb:stringConcat(?z, ?y, "IAMO") -> hasP1IP(?x, ?z)
```

(b)

Figure 1: A rule deriving the first person plural of the present indicative for verbs such as "mangiare". (a) LMF version (b) SWRL version. In the SWRL version the premise of each rule is composed of 3 atoms: the first identifies the inflectional class of an entry, the second the stem of the entry, and the last concatenates the correct suffix for the inflected form with the correct stem.

1 See the Ontolex Lemon final specifications (ONTOLEX - W3C. 2016 <https://www.w3.org/community/ontolex/>)

2 https://www.w3.org/Submission/SWRL/

3 AA. VV., 2016, PAROLE-SIMPLE-CLIPS, Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR, (http://hdl.handle.net/20.500.11752/ILC-88)

Each rule applies to a subset of the verbs in a lexicon, those which form an inflectional class with the same behaviour; the same is true in Italian of other the parts of speech. While LMF allows such rules to be represented and associated with lexical entries it would be necessary to write specialised software to make them operational. On the other hand with SWRL rules we can use commonly available Semantic Web technologies to make patterns such as Figure 1(b) actionable. So that it is possible to enter a new word in the dataset, associate it to an inflectional class and be able to retrieve all its inflected forms.

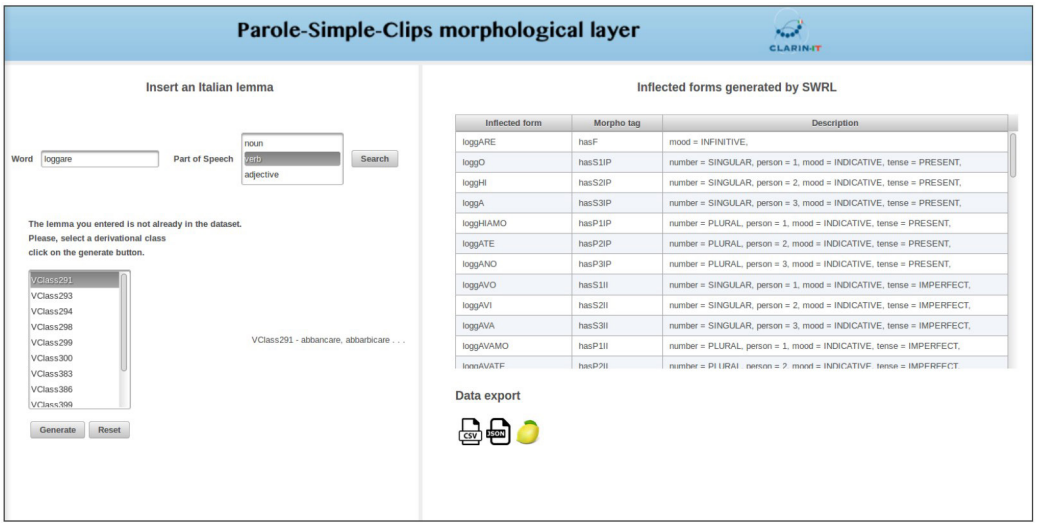


Figure 2: Inflected forms generator. The user enters the lemma “loggare” (to log in) which is not already in the dataset.

In order to demonstrate both the usefulness of the rule-based part of the lexicon as well as to make it easier for users to work with, we have developed an interface[4] for querying the dataset by entering the lemma forms of words. If the lemma exists in the dataset then the interface will show all of its inflected forms. If it doesn’t however, then thanks to the fact that our morphological patterns are represented by SWRL rules, it is easy to add new lemmas to the dataset if they belong to one of the pre-existing morphological classes (see Figure 2). After inserting a lemma and its POS, the system queries a sparql endpoint and a list of relevant inflectional rules is produced[5]. For each rule example lemmas are provided in order allow the user to identify the correct inflectional class. On hitting enter, the SWRL rule generates the full inflectional paradigm on the fly.

We believe that SWRL could play an important role in overcoming some of the limitations of previous RDF-based lexicons, and so we would like to start a discussion on the implementation of rules in LD lexicons for the representation of morphology and other lexical features. In the final work we shall provide detailed information on the conversion of the PSC-M morphological rules into SWRL and on the validation that we carried out using the full inflected forms contained in the original PSC DB, as well as a number of technical details that may be of use to anyone wishing to implement SWRL rules as part of a LD dataset.[6]

Keywords: SWRL, inflectional morphology, Italian

4 A demo of the application of rules to neologism is available at http://lari-lsj.ilc.cnr.it/pscMorphoRules.

5 Currently the system shows an unordered list of all classes to the user, but in the future we are planning to order the list in terms of relevance and remove classes which cannot be candidates.

6 We have made a version of the lexicon available containing both the rules used to generate the lexicon and the properties that result as an RDF dump at http://lari-datasets.ilc.cnr.it/pscMorph#; a SPARQL endpoint is available at http://lari-datasets.ilc.cnr.it/pscMorph/queryForm.html.

Acknowledgments

This work in part was funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

References

Del Gratta, R., Frontini, F., Khan, F., Monachini, M. 2015. Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web Journal* 6(4). 387–392.

Francopoulo, Gil. 2013. *LMF Lexical Markup Framework*. John Wiley & Sons.

Khan, F., Bellandi, A., Frontini, F., Monachini, M. 2017. Using SWRL Rules to Model Noun Behaviour in Italian. *LDK 2017: Language, Data, and Knowledge*, 134–142. (Lecture Notes in Computer Science). Springer, Cham.

Khan, F., Frontini, F.. 2014. Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, 224–228. Pisa University Press.

Khan, F., Frontini, F., Boschetti, F., Monachini, M. 2016. Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon. *Digital Humanities 2016: Conference Abstracts*, 593–596. Kraków: Jagiellonian University & Pedagogical University.

McCrae, J., Montiel-Ponsoda, E., Cimiano, P. 2012. Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*, 25–34. Springer.

ONTOLEX - W3C. 2016. Final Model Specification - Ontology-Lexica Community Group. https://www.w3.org/community/ontolex/wiki/Final_Model_Specification (8 June, 2017).

Ruimy, N., Corazzieri, O., Gola, E., Spanu, A., Calzolari, N., Zampolli, A.. 1998. LE-PAROLE Project: The Italian Syntactic Lexicon. *EURALEX '98*.

Wilcock, G. 2007. An OWL Ontology for HPSG. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 169–172. (ACL '07). Stroudsburg, PA, USA: Association for Computational Linguistics.Linguistics, 1–6.

The Good, the Bad and the Noisy? An Analysis of Inter-Annotator Agreement on Collocation Candidates in Different Grammatical Relations

Iztok Kosem, Polona Gantar, Simon Krek, Jaka Čibej, Špela Arhar Holdt
Faculty of Arts, University of Ljubljana
"Jožef Stefan" Institute
Faculty of Computer and Information Science, University of Ljubljana
iztok.kosem@ff.uni-lj.si, apolonija.gantar@guest.arnes.si, simon.krek@guest.arnes.si, jaka.cibej@ff.uni-lj.si, Spela.ArharHoldt@ff.uni-lj.si

Recent trends in lexicography have focussed on automating certain aspects of language description, especially those related to collocations and examples (e.g. Rundell and Kilgarriff 2011). Automatic extraction of collocations poses a number of challenges. As cut-off parameters must be set in order to keep the data size manageable, the selection of a statistic measure for ranking the colllocation candidates is key. While logDice (also used in the Sketch Engine tool), which we used in our research, is predominantly used for lexicographic and language teaching purposes, Gantar et al. (2016, 2015) have observed that logDice often misses, or attributes very low ranking to, certain important collocates, which is why they started combining logDice and raw frequency rankings when extracting and analysing collocates. Adding grammatical parameters, i.e. grouping collocations by grammatical relations, improves the results; however, this brings other potential issues such as the reliability of capturing different grammatical relations via POS tags, the approach used by Word Sketch.

We have set out to address these issues in the KOLOS (Collocations in Slovene) research project (2017-2021), with the aim to evaluate and improve the current method by testing new approaches (e.g. distributional semantics) and testing collocation extraction on parsed corpus data. In this paper we present the results of the initial evaluation of approximately 8,400 collocation candidates in 227 different grammatical relations (an average of approximately 37 collocations per relation) on a randomly selected set of lemmas from the Slovene Lexical Database (Gantar and Krek 2011). The evaluation was conducted by six annotators (linguists) in the Pybossa platform. The question asked was "Is the following phrase a collocation?", and the options offered were YES, NO, and I DON'T KNOW. We were expecting that the annotators would each have their own understanding what a collocation is, so some minimal guidelines were provided; still, this was also one of the aims of the task, namely to determine to what extent can different linguists agree on what a collocation is and what it is not. Each collocation was evaluated by three annotators; the average inter-annotator agreement (Cohen's kappa) was 0.39, while the average percentage of same answers was 0.65.

The findings were especially interesting when analysed by grammatical relation. Firstly, the relations differed in terms of collocational productivity, i.e. the percentage of good collocation candidates (a collocation candidate was considered good if it was annotated with YES by at least two annotators). Only 58 out of 227 relations had 50% or more good collocation candidates; on the other hand, nearly half of the relations had 25% or less good collocation candidates. Among the most frequently noticed reasons for such low overall collocational productivity are tagging errors, which are closely connected to the morphological complexity of the Slovene language (e.g. homonymous forms for different word classes or different case forms). Another problem are incorrect collocations because they only capture a part of an established phrase, e.g. *primer* could be identified as as noun in nominative or

accusative, although it can be part of a frequent phrase *na primer* 'for example'. Furthermore, because the lemma set was randomly selected, certain relations were limited to one particular lemma, which means it is difficult to make any generalisations.

Secondly, there were significant differences in terms of annotator agreement: in many relations, the annotator agreement was very high, and there were no cases in which the annotators completely disagreed (all three answers were different). However, for some grammatical relations, the level of inter-annotator agreement was extremely low, with as many as 20-50% of candidates within the relation being evaluated with three different responses (e.g. verb + preposition *v* + noun in the accusative case). Discussions with the annotators revealed that although some guidelines were provided, there were clear differences in their understanding of the notion of collocation; while some understood it very broadly and were mainly attempting to confirm or reject statistically-based output, other were already evaluating collocation candidates in terms of their relevance for a dictionary of collocations. Moreover, the annotators also complained that they were missing additional options as collocation candidates sometimes seemed fine, but needed a minor change in presentation, for example the noun in the collocation needed to be in plural.

Based on these findings, we decided to conduct a second stage of the analysis, for which we prepared a more carefully selected set of 333 lemmas, which were heterogeneous in terms of word class and its subcategories (e.g. plural nouns, countable nouns, transitive vs. intransitive verbs etc.), corpus frequency, level of polysemy, etc. 17,613 collocation candidates in 142 grammatical relations are being evaluated by seven annotators at the time of writing. Additional answer options and more detailed instructions have been provided to the annotators. The results of this second evaluation, which will also serve as a training data for a distributional semantic task, will also be presented at the conference.

Keywords: collocation, evaluation, annotator agreement, Slovene

References

Gantar, P., Krek, S. (2011): Slovene lexical database. D. Majchráková and R. Garabík (eds.): *Natural language processing, multilinguality*. Brno: Tribun EU. 72–80.

Gantar, P., Gorjanc, V., Kosem, I., Krek, S. (2015): Going semi-automatic and crowdsourced: collocation dictionary of Slovene. In Kosem, I. (ed.). *Electronic lexicography in the 21st century: linking lexical data in the digital age*. *eLex 2015, book of abstracts.* Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton: Lexical Computing. 2015, p. 37.

Gantar, P., Kosem, I., and Krek, S. (2016): Discovering Automated Lexicography: The Case of the Slovene Lexical Database. *International Journal of Lexicography*, 29 (2): 200–225.

Rundell, M., Kilgarriff, A., (2011): Automating the creation of dictionaries: where will it all end? Meunier F., De Cock S., Gilquin G. and Paquot M. (eds.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Benjamins, pp. 257–281.

Innovative Usage of Graphical Illustration in Lexicography. Writing Definitions

Ewa Kozioł-Chrzanowska, Piotr Żmigrodzki
The Institute of the Polish Language at the Polish Academy of Sciences
ewa.koziol06@gmail.com, pzmigr@gmail.com

This presentation aims to demonstrate how the Google Images (GI) search engine can facilitate the definition writing process (compare Verdiani 2016). The graphic illustration, gathered by the GI search engine, can be used by lexicographers in the process of writing dictionary entries. This tool allows users to search images and combine them with the given words and phrases. Furthermore, it provides naive, non-encyclopedic point of view (crucial for the contemporary general metalexicography, e.g. Mikołajczak-Matyja 1998), as well as simply a significant amount of information available in an instant, independently of time and place.

Firstly, we would like to focus on the types of definienda to show the language units for which the analysis of graphical illustration is the most useful: rarely used, specialized vocabulary, especially if they name similar objects, the units with the restricted usage (particular situations, groups of users, historical periods, dialects etc.). The GI enables lexicographers to describe these language units properly and, e.g., see the detailed differences between the objects "wystawka" ("a veranda") and "wykusz" ("a bay window"). Such details are basically unrecognizable by the verbal corpus.

Secondly, we would like to concentrate on the types of definientia to indicate how the GI can be used in the different verbal explanatory modes (traditional vs. cognitive explications etc.). The tool will help to choose an appropriate genus proximum (in classic definitions – to avoid mistakes described by Bańko 2001), examples related closely to the meaning of the given word (in ostensive definitions), it also provides useful information for preparing non-encyclopedic and cognitive definitions.

The GI used as a facility for gathering lexicographical information has also a lot of disadvantages. We would like to point them out by showing the exampes in which GI provided linguistically irreleavant or even wrong information. Despite that the tool may be a valuable (in some cases – indispensable) addition to traditional text sources.

Keywords: multimodal lexicography, writing definitions, defining, illustrations

References

Bańko M., 2001, Z pogranicza leksykografii i językoznawstwa: studia o słowniku jednojęzycznym. Warszawa: Wydział Polonistyki UW.

Mikołajczak-Matyja N., 1998, Definiowanie pojęć przez przeciętnych użytkowników języka i przez leksykografów. Sorus: Poznań.

Ostermann C., 2012, Cognitive lexicography of emotion terms, in: Proceedings of the 15th EURALEX International Congress. 7-11 August 2012, R. Vatvedt Fjeld, J. M. Torjusen (eds.). University of Oslo: Oslo, pp. 493-501.

Scholze-Stubenrecht W., 2013, The World Wide Web as a resource for lexicography, in: Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume, eds. R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand. de Gruyter Mouton: Berlin-Boston, pp. 1365-1375.

Storjohann P., 2017, Cognitive Features in a Corpus-based Dictionary of Commonly Confused Words, in: Electronic lexicography in the 21st Century. Proceedings of eLex 2017 Conference, eds. I. Kosem et al., Brno, pp. 138-154.

Verdiani S., 2016, Definizioni illustrate. La selezione di un repertorio di immagini per migliorare la trasparenza semantica delle voci del dizionario di apprendimento Tedesco Junior, in. Proceedings of the XVII Euralex International Congress. Lexicography and Linguistic Diversity, ed. T. Margalitadze, G. Meladze, Ivane Javakhishvili Tbilisi State University: Tbilisi, pp. 146-155.

Korean Expressions of Mitigation in Product Reviews

Minju Lee, Hyeonah Kang, WeonSeok Shin, Kil Im Nam
Kyungpook Natioanl University
lmj0355@hanmail.net, hyuna_72@naver.com, gx3820@hanmail.net, nki@knu.ac.kr

This study examines the types and functions of mitigating expressions in Korean product reviews. According to Hyland (2004), mitigation has a meta-discursive function which reveals the point of view of a text or the author's attitude and thus, can be seen as an interactional strategy between writer and reader. In that sense, mitigating expressions are considered 'semantic units' that provide insights into the writer's subjective view on a particular object or phenomenon. With the development of information and communication technologies, the boundary between writer and reader tends to fade away and their mutual interactions to become more important. In such a context, the study of mitigating expressions helps to cast light on subjectivity, that is, on the language on one's sentiments, evaluations, and/or attitudes, by analysing and classifying the relevant expressions, thereby providing crucial data for the description of subjective expressions in dictionary entries. This study makes use of a 'Product Reviews Corpus' that consists of subjective texts and daily life language, and aims to classify mitigating expressions and examine their functions as basic data for the lexicographic description of lexical bundles.

For this study, we built a corpus consisting of product reviews that have been web-crawled from online shopping websites and which focus on 'knits/sweaters, shirts/blouses, jeans, formal dresses, dresses'. This corpus comprises 340,000 texts, amounting to 4,100,000 *ecel* (Korean word unit), and has been morphologically annotated. This 'Product Reviews Corpus' is a small-scale corpus, but it is characterised by a theme centred on the evaluation of clothing products and the emergence of a new form of discourse through the recurrence of various expressions. These expressions can be used to reveal the immediate thoughts of the speaker, but also to dissipate the hearer's feeling of uneasiness at the speaker's utterance. It is precisely this latter function that Hyland (1998) has described as mitigation. Mitigating expressions indicate the degree of certainty of a statement and enable the writer to cautiously make claims and effectively communicate with the reader. In that sense, they are not formulaic expressions that occur only in a restricted context, but are instead strategic expressions centred on speaker/hearer effective communication.

For this study, 1,000 cases were randomly selected from the corpus and the results of our analysis are as follows. First, if we consider the totality of the cases, the proportion of mitigating expressions amounts to 45.8%, compared to that of straightforward expressions which is 54.2%; that is to say, mitigation expressions are used in near half of the reviews. Despite the fact that the review is a subjective text wherein the author can freely express their feelings, mitigation as a way to attenuate their sentiments is also greatly used by consideration for the reader. In addition, it could also be seen as a politeness strategy to avoid face-threatening behaviours.

Second, the classification of mitigating expressions that appear in product reviews includes types such as assumptions, minimizers, supposition/hope, and passive expressions. The forms and ratios for each type are shown in the following table.

Table 1: Classification of Korean expressions of mitigation in Product Reviews.

Type	Lemma			Frequency (%)
Minimizers	Lexical/grammatical expressions	Adverb and adverbial expression	kunyang, , kulekcelek, kuce, yakkan, cokum/com, uyoylo	347(50.2)
		Conjunctive adverb	kulena, kulayto, kulentey/kuntey, taman , haciman	
		Adjectival verb	kulehtha	
		Modal verb ending(prefinal/final)	-(n)tey, -taman, -(u)na, -ciman	
Assumptions	Lexical/grammatical expressions	Noun	tus, cengto	249(36.0)
		Auxiliary adjectival verb	tussiphta, tushata, manhata	
		Modal verb ending(prefinal/final)	-keyss-, -l kes, -kes kathta	
	Lexical bundle	Kam-i iss-ta, phyen-i-ta		
Reported speech	Lexical/grammatical expressions	Quotative particle	(la)ko	31(4.4)
Supposition/ hope	Lexical/grammatical expressions	Adverb and adverbial expression	iltan, hoksina	24(3.4)
		Auxiliary adjectival verb	-a/e siphta, -ko siphta	
		Modal verb ending(prefinal/final)	-l theyntey, -(u)myen	
Passive expressions	Lexical/grammatical expressions	Auxiliary verb	-(key) toyta	18(2.6)
Others				14(2.6)
Total				682(100)

The above table shows that the lexical and grammatical expressions extracted are diverse, but the types are rather limited. This suggests that only specific types are used in reviews, as in academic texts and other genres. The mitigating expressions appearing in the Product Reviews Corpus correspond to the writer-centred mitigating expressions described by Hyland (1998). These are used when the writer wants to state facts about a product or his feelings about it. Their functions are precisely to decrease the intensity of these statements and anticipate the reactions of readers who think otherwise.

Third, the mitigating expressions found in the Product Review Corpus can be used to supplement the treatment of lexical bundles in Korean dictionaries. Rundell (1999: 37) suggested that the dictionary should provide not only a word's syntactic behaviour, but also its semantic features, contextual effects, collocational preferences and selectional restrictions. For example, in Korean dictionaries, the lemmas '*kes* (thing, fact)' and '*kathta* (is same)' are described in their respective entries, but the high-frequency lexical bundle '*-kes kath-* (it seems that)' is only briefly described under the headword '*kathta*' as 'expression of assumption or uncertain conclusion' with no examples or further information on usage. However, this lexical bundle is highly frequent in subjective texts (see Table 1). Therefore, the dictionary description of '*-kes kath-*' should include not only collocational information but also pragmatic information as it can be used as a mitigation marker in a context of judgement

or evaluation. Thus, the mitigating expressions found in the Product Review Corpus can serve as valuable basic data for the improvement of lexicographic treatment of collocation description and pragmatic information.

Keywords: mitigating expressions, mitigation, sentiment analysis, corpus methods, dictionary

References

Hyland, K. (1998). Hedging in scientific research articles, Amsterdam/Philadelphia: John Benjamins.

Hyland, K., Tse, P. (2004). Metadiscourse in Academic Writing : A Reappraisal, Applied Linguistics 25/2, Oxford University Press, 156-177.

Rundell, M. (1999). Dictionary Use in Production, International Journal of Lexicography, 12(1), 35-53.

Usage Notes – Cinderella in the Dictionary-Making Process

Hana Mžourková
Department of Language Cultivation of the Czech Language Institute of the Czech Academy of Sciences
mzourkova@ujc.cas.cz

Lexicography in the first decades of the 21st century has shown different trends in creating digital dictionaries. Digital dictionaries, formerly created as information databases, are increasingly being designed and developed as informational tools, able to take advantage of all the possibilities available within the digital environment, including search engines and search engine filters (Bergenholtz, 2011; Tarp, 2011).

Contemporary lexicography places a great emphasis on the user's perspective and behavior within the digital environment (Lew & de Schryver, 2014). Many studies consider dictionary users within the digital realm and focus on user satisfaction; however, matters of dictionary microstructure in relation to dictionary users are equally important.

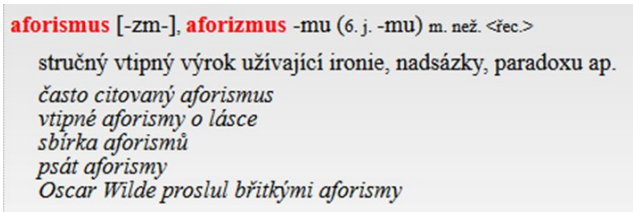
The first part of this paper focuses on the crucial element of headword structure – a usage note – which belongs to the wide concept of usage information. This concept is broad and includes all lexicographic information targeted at the user such as stylistic and field-of-interest labels, example sentences, and usage notes (Atkins, 1992/93). In this paper we focus only on usage notes advising the user on relations between parts of the headword (for example between orthographical doublets, inflectional doublets or headword's meanings) and their communication status. The information in usage notes is hardly replaceable with a stylistic label. While stylistic labels distinguish the level of formality of headwords, usage notes differentiate whether the headword is marked or unmarked. Usage notes thus help users to distinguish between all sorts of language elements with the same communication validity, but subsume different pragmatic relations.

We see the usage notes as a key tool helping the dictionary user to better understand the headword's information and use it correctly in communication. Not only do usage notes contain data needed by the user, but they are also essential for clear and economical structure of the dictionary headword. The digital dictionary environment offers a wide range of options, but at the same time carries a risk that too much information will overwhelm the reader and may lead to misinterpretations of the data found (Bergenholtz, 2011). Usage notes can partially substitute various headwords' units (some example sentences, additional paradigms and others) and contribute to the economy of headwords. So far, few studies exist examining usage notes (Whitcut, 1985; Atkins & Rundell; 2009). In this paper we consider how to best take advantage of this neglected topic.

The second part of our paper focuses on the benefits of usage notes in the dictionary-making process. In Czech Language Institute two digital databases will be merged into one, consistent with modern methods of language data presentation. To create an informative and user-friendly online environment, the conflict of the different conceptions of these two databases must be dealt with before synthesizing the data. The first database, Internet Language Reference Book (ILRB), accessible since 2008 at: http://prirucka.ujc.cas.cz/en includes usage notes. The second, Academic Dictionary of Contemporary Czech (ADCC), a new monolingual dictionary of Czech language, accessible since 2017 at: http://www.slovnikcestiny.cz/web/uvod.php, currently being developed as a dictionary for general users (Béjoint, 2016), does not include usage notes.

We will illustrate different approaches to the dictionary user and the application of usage notes in ILRB and ADCC on a headword *aforismus*, which has its orthographical doublet *aforizmus*. Both

doublets are orthographically correct, but the second one is marked by users as rarely used in written communication. Still, the ADCC doesn't take the pragmatic difference between doublets into account and therefore the user might understand the doublets *aforismus – aforismus* incorrectly as neutral communication equivalents (Figure 1).



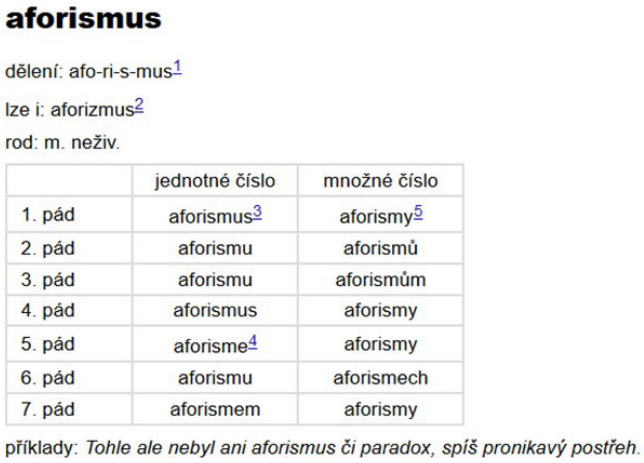
aforismus [-zm-], **aforizmus** -mu (6. j. -mu) m. než. <řec.>

stručný vtipný výrok užívající ironie, nadsázky, paradoxu ap.

často citovaný aforismus
vtipné aforismy o lásce
sbírka aforismů
psát aforismy
Oscar Wilde proslul břitkými aforismy

Figure 1: Classifying the headword *aforismus – aforizmus* in ADCC

The ILRB works with usage notes, using linked interpretative passages in superscripts (Figure 2).



aforismus

dělení: afo-ri-s-mus[1]

lze i: aforizmus[2]

rod: m. neživ.

	jednotné číslo	množné číslo
1. pád	aforismus[3]	aforismy[5]
2. pád	aforismu	aforismů
3. pád	aforismu	aforismům
4. pád	aforismus	aforismy
5. pád	aforisme[4]	aforismy
6. pád	aforismu	aforismech
7. pád	aforismem	aforismy

příklady: *Tohle ale nebyl ani aforismus či paradox, spíš pronikavý postřeh.*

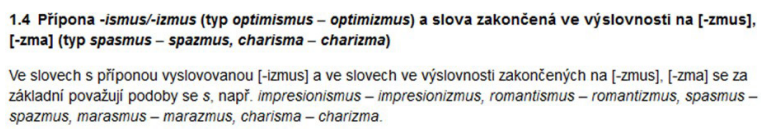
Figure 2: Classifying the headword *aforismus – aforismus* in ILRB

However, the ILRB's way to access relevant information (the answer to the question: Shall I use *aforismus* or *aforizmus* in written communication?), may be too complicated for the user. Firstly, the user must find the applicable link in the blue superscript (2), then open it, and again, search in the text to find the relevant information (Figure 3 and 4).

Pravopis a výslovnost přejatých slov se *s – z* (Skrýt)

1 Dublety
1.1 Základní pravopisná podoba je se *z* (typ *muzeum – museum*)
1.2 Dublety jsou stylově rovnocenné (typ *kurz – kurs*)
1.3 Základní pravopisná podoba je se *s* (typ *diskuse – diskuze*)
1.4 Přípona *-ismus/-izmus* (typ *optimismus – optimizmus*) a slova zakončená ve výslovnosti na [-zmus], [-zma] (typ *spasmus – spazmus, charisma – charizma*)
2 Jediná možnost psaní je se *s* (typ *designovat, konsekuce*)
3 Poznámka o shodě stylového příznaku

Figure 3: Linked list of usage notes (superscript No 2)



1.4 Přípona *-ismus/-izmus* (typ *optimismus – optimizmus*) a slova zakončená ve výslovnosti na [-zmus], [-zma] (typ *spasmus – spazmus, charisma – charizma*)

Ve slovech s příponou vyslovovanou [-izmus] a ve slovech ve výslovnosti zakončených na [-zmus], [-zma] se za základní považují podoby se *s*, např. *impresionismus – impresionizmus, romantismus – romantizmus, spasmus – spazmus, marasmus – marazmus, charisma – charizma*.

Figure 4: Linked usage note 1.4 for orthographical doublets

To address this problem, we recommend using pop-up dialogues containing usage notes (based on interpretative passages from ILRB) and locate them in suitable places within the headword (Figure 5).

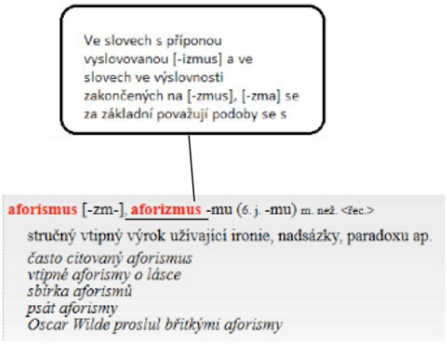


Figure 5: Displaying merged data from ADCC and ILRB

In an extended version of the paper we also consider the typology, model location and metalanguage of usage notes. All of these features should be used during the dictionary-making process to achieve a user friendly digital dictionary.

Keywords: digital dictionary, dictionary user, headword, markedness, usage note, user-friendliness

References

Atkins, B. T. S. (1992/93). Theoretical lexicography and its relation to Dictionary-making. *Journal of the Dictionary Society of North America*, 14: 4–43.

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

Béjoint, H. (2016). Dictionaries for general Users: History and development; Current Issues. In P. Durkin, editor, *The Oxford Handbook of Lexicography*, pp. 7–24.

Bergenholtz, H. and Bergenholtz, I. (2011). A Dictionary is a Tool, a Good Dictionary is a Monofunctional Tool. In A. Fuertes-Olivera and H. Bergenholtz, editors, *e-Lexicography. The Internet, Digital Initiatives and Lexicography*, pp. 187–207. Continuum, New York.

Fuertes-Olivera, A. and Bergenholtz, H. (2011). Introduction: The Construction of Internet Dictionaries. In A. Fuertes-Olivera and H. Bergenholtz, editors, *e-Lexicography. The Internet, Digital Initiatives and Lexicography*, pp. 1–16. Continuum, New York.

Lew, R. and de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *Internetional Journal of Lexicography*. 27 (4): 341–359.

Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. In A. Fuertes-Olivera and H. Bergenholtz, editors, *e-Lexicography. The Internet, Digital Initiatives and Lexicography*, pp. 54–70. Continuum, New York.

Whitcut, J. (1985). Usage Notes in Dictionaries: The Needs of the Learner and the Native Speaker. In R. Ilson, editor, *Dictionaries, Lexicography and Language Learning*, pp. 75–80, Pergamon Press, Oxford..

Semi-automating the Reading Programme for a Historical Dictionary Project

Tim van Niekerk[1], Johannes Schäfer[2], Heike Stadler[2] and Ulrich Heid[2]
DSAE[1], Rhodes University; IwiSt[2], University of Hildesheim
t.vanniekerk@ru.ac.za, {johannes.schaefer, stadle, heidul}@uni-hildesheim.de

We report on a major enabling step towards the revision of the scholarly reference work *A Dictionary of South African English on Historical Principles* (*DSAE*, Silva *et al.* 1996), namely the semi-automatic generation of a digitally-sourced lexical database on which new and updated dictionary entries will be based; as well as the addition, in parallel, of a new corpus of South African English (SAE) to the project. Drawing on online data sources and an extensive list of known SAE word forms, we have developed a software toolchain to gather, encode, annotate and collate textual sources, producing: (i) a 3.1-billion part-of-speech-annotated corpus of South African English; (ii) a lexical database of illustrative quotations for about 20,000 known SAE word forms, available for selection at the entry-revision stage; and (iii) lists of potential variants and inclusion candidates. These steps replace, where recent electronic sources are concerned, the mechanical aspects of quotation gathering, normally undertaken manually through a reading programme requiring years of teamwork to acquire sufficient coverage (cf. Hicks, 2010).

1 Need for quotations

A Dictionary of South African English on Historical Principles is a diachronic variety dictionary, first published in 1996 as a single-volume *OED*-like print dictionary, and available online at http://dsae.co.za since 2014. It is based heavily on bibliographically-referenced quotations: much of the *DSAE*'s 25-year compilation process involved collecting approximately 300,000 index card citations as evidence for entries and their sense-divisions. Just as the latest *OED* revision dedicated "a vast amount of well-directed energy" towards gathering new quotations (Brewer: 241), so the *DSAE* revision requires increased data holdings of post-1995 citations. The current project deploys semi-automation to boost these resources, dramatically reducing the labour involved.

2 Data sources

We draw on two sources of data: the first is a newspaper corpus of approximately 100 million tokens, created for quotation-gathering purposes by the authors from online sources dating from 2015–2017. The second is a generic corpus of about 3 billion words, generated from 2011–2014 .za web domain sources by the NLP Group of the Computer Science Department at Leipzig University, as part of its CURL (Crawling Under-resourced Languages) project (see Quasthoff *et al.* 2015). The Leipzig dataset offers, by design, limited bibliographical and contextual information, but its sentence-segmented structure nevertheless lends itself to quotation mining. Together these corpora facilitate research into current SAE on a new scale: in 2009 it was reported that "there is no large corpus to represent South African English" (Pienaar & De Klerk 2009: 356) and, apart from proprietary, unfinished, or very small special-purpose corpora of under 1 million words, no others were available to the project prior

to the current collaboration. All data are POS-tagged (using the Treetagger) including a lexicon-based lemmatization and loaded into the IMS Open Corpus Workbench (CWB) to enable efficient further querying.

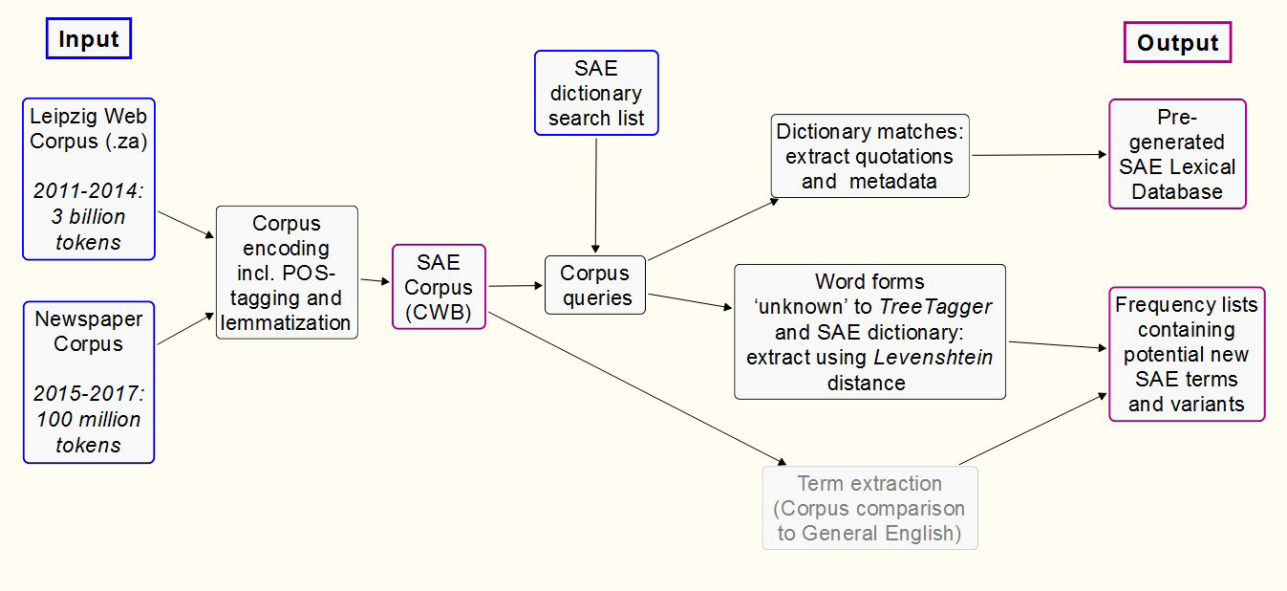


Figure 1: Toolchain workflow generating SAE Corpus, Lexical Database and lists of potential new words from 2011–2017 sources.

3 From manual to semi-automated quotation gathering

The traditional manual citation collection process required five stages: (i) identify, (ii) capture, (iii) proofread, (iv) annotate, and (v) review captured quotations. The new toolchain automates the mechanical aspects of the reading, capture and, to some extent, the annotation processes (stages i, ii and iv) against the Newspaper and Leipzig corpora. Proofreading (stage iii) largely falls away. Importantly, the toolchain extracts quotations for documented word forms and variant spellings, linking them to their associated dictionary entries. The former 5-stage manual workflow is reduced to the final stage (v), possibly with further additional annotation (stage iv) being desirable. With this toolchain we extracted quotations from our SAE corpus for a DSAE subset of 7,025 headwords with 21,768 variants. We manage to find at least one variant of approximately 85% of these headwords. Overall, around half of the variants of the DSAE headwords are represented in our corpus with an average frequency of approximately 100.

4 Towards semi-automated discovery of new terms

During the corpus-encoding process, we also exploit lacunae in the TreeTagger's English lexicon, which is not oriented specifically towards South African English; hence those terms it flags as 'unknown' sometimes highlight new South Africanisms. The toolchain thus prompts a new strategy in data collection: lexicographers can focus on identifying single instances of previously-undocumented South Africanisms, and supply these to the toolchain for future matching against corpus data in an iterative workflow. Thereafter the task of the lexicographer is reduced to more specialised work requiring human intelligence, namely the evaluation and selection of quotation evidence, moving the project from the data collection stage much closer to the dictionary drafting process.

Two further types of processing allow semi-automatic discovery of (1) variant spelling forms and (2) entirely new SAE headword candidates. After additional filtering of 'unknown' items to remove proper nouns and known words, we identify potential unrecorded variants using a combination of Levenshtein distance and corpus frequency. Corpus frequency is indicated in the resulting output and is especially useful to the editor in evaluating multiple possible orthographical candidates at a glance. The variant detection process also sometimes produces entirely new word forms. Future work will extend this approach by using term extraction techniques to help find undocumented new word forms in comparison with General English.

Keywords: dictionary workflows, lexical databases, historical lexicography, language varieties, South African English, reading programmes, corpora

References

Brewer, C. (2007). *Treasure-house of the Language: the Living* OED. London and New Haven: Yale University Press.

Hicks, S. (2010). Firming up the foundations: Reflections on verifying the quotations in a historical dictionary, with reference to *A Dictionary of South African English on Historical Principles.* In *Lexikos* (20), pp. 248–271. Accessed at http://lexikos.journals.ac.za/pub/article/view/142 [03/12/2017].

Pienaar, L. & De Klerk, V. (2009) Towards a corpus of South African English: Corralling the sub-varieties. In *Lexikos* (19), pp. 353–71. Accessed at http://lexikos.journals.ac.za/pub/article/view/444 [03/12/2017].

Quasthoff, U., Goldhahn, D. and Eckart, T. (2015) Building large resources for text mining: The Leipzig Corpora Collection. In: *Text Mining: From Ontology Learning to Automated Text Processing Applications*, Biemann, C. and Mehler, A. (Eds.), Berlin: Springer.

Silva, P., Dore, W., Mantzel, D., Muller, C. & Wright, M. (1996). *A Dictionary of South African English on Historical Principles.* Cape Town: Oxford University Press. Online at http://www.dsae.co.za.

Towards a historical Anglo-Norman Dictionary

Heather Pagan
Anglo-Norman Dictionary
hap@aber.ac.uk

Over the last two decades, the Anglo-Norman Dictionary (www.anglo-norman.net) has been under revision, a process driven initially by the extensive expansion of its corpus. In parallel with the editorial revision, the dictionary has been transformed into an entirely online entity, with major technical developments over the last decade allowing considerable expansion of the semantic presentation, including hyperlinks to entries in cognate dictionary entries and tools for searching by semantic category.

Throughout the revision process, the primary organisation of the dictionary has remained a semantic one, and indeed, the online introductory material for the dictionary warns the reader that the provided citations for a given lemma should not be used as indications of earliest (or latest) attestation as they have been chosen for their semantic, rather than historical, value. The Reader's Guide to the second edition of the *Anglo-Norman Dictionary* (ed. W. Rothwell, D. Trotter et al., London: MHRA, 2005) warns the reader against over-interpretation of the citations given:

"Readers are nevertheless reminded that AND is not a historical or etymological dictionary. No systematic attempt has been made to supply a chronological account of vocabulary or of semantic developments; an attestation which occupies first place in an entry may well not be the chronologically oldest attestation, which ill not always be included at all; and words or meanings may in fact have survived in later use than the quotations in the Dictionary could suggest. The Dictionary's entries are *semantically*, not historically structured, and whilst the one might in theory coincide with the other, this will not always be so. Moreover, the range of attestations available to the editors does not of course always allows a complete historical account even had it been our intention to supply it. *Caveat lector*: absence of evidence is not evidence of absence." (Trotter, 2005, xxvi)

Nevertheless, a number of historical dictionaries and thesaurus projects, including, for example, the Oxford English Dictionary, the Dictionnaire Etymologique de l'Ancien Français, and the Bilingual Thesaurus of Everyday Life in Medieval Britain, rely on the data provided by the AND as evidence of earliest / latest use of a particular word, or as evidence of language contact. Given the demand for this information from dictionaries of Medieval French and Latin as well as Middle English, the editors of the Anglo-Norman Dictionary have embarked on a programme to transform the AND entries from entries solely driven by a semantic presentation, into a historical dictionary, or at least, one which will provide information on the earliest attestation of a sense, dates of individual citations as well as texts and a reorganisation of all given citations. These initial steps will allow researchers to go beyond the semantics of an Anglo-Norman word, facilitating an analysis of the transformation of the language between the eleventh and sixteenth centuries.

This paper would like to examine the process by which this transformation will occur, highlighting some of the difficulties in dating Anglo-Norman citations as well as the technical challenge of retroactively adding citations to completed entries. The paper will also examine some of the initial results of this process, showing how this research will help clarify the complicated contact relationship between Anglo-Norman and continental varieties of medieval French as well as between Anglo-Norman and Middle English and medieval Latin.

Keywords: Anglo-Norman, Middle English, historical dictionary

Multi-word Lexical Units in General Dictionaries of Slavic Languages

Andrej Perdih, Nina Ledinek
Fran Ramovš Institute of the Slovenian Language ZRC SAZU
andrej.perdih@zrc-sazu.si, ledinekn@gmail.com

Multi-word lexical units (MLUs) represent a large part of the lexicon of a language. Because of the complexity of their use several questions arise regarding their inclusion in general dictionaries, such as how to identify them during the dictionary compilation process, where in the microstructure or macrostructure of a dictionary are they to be positioned, how to treat different types of MLUs, how to deal with border cases MLUs etc.

The structure, typology and the dictionary macro- and microstructure positioning of MLUs were examined in several Slavic dictionaries: the Slovenian *Slovar slovenskega knjižnega jezika / Dictionary of the Slovenian Standard Language* (1970–1991[1], 2014[2]) the Croatian *Veliki rječnik hrvatskoga jezika* / Great Dictionary of the Croatian Language (1998[3]), the Slovak *Slovník súčasného slovenského jazyka* / *Dictionary of Contemporary Slovak* (2006–) the Polish *Wielki słownik języka polskiego* / *Great Dictionary of Polish* (2007–) and the Russian *Tolkovyj slovar' russkogo jazyka s vklyucheniem svedenij o proishozhdenii slov* / Explanatory Dictionary of Russian Language Including Explanations of Word Origin (2008).

Based on theoretical grounds and on the analysis of the aforementioned dictionaries the article deals with several questions regarding MLUs in dictionaries. Firstly, the question of MLUs being used as headwords (and subheadwords) is discussed, because there are lexical units that can be treated either as single-words (written with a whitespace) or MLUs, and because of the status of some MLUs as individual semantic units in the language system. The question of nesting of MLUs is also discussed, since in dictionaries various MLU types can be presented in a single nest, while presenting other MLU types in other parts of the dictionary microstructure. Where can MLUs also be presented in dictionaries? The obvious possibility is existence of several nests, and besides functioning as headwords often the MLUs are listed among collocates regardless of whether a sense definition is added to a MLU or not.

The findings are compared to solutions presented in the new dictionary of Slovenian language, part of which has already been published (eSSKJ: *Slovar slovenskega knjižnega jezika 2016–2017*). In eSSKJ, the MLUs are not treated as headwords except for the types that can be interpreted also as single-words, such as structures of verbs with reflexive morphemes, loan words written as multi-words according to the original orthographical rules, and clusters that are not clear examples in Slovenian orthography. The majority of MLUs are treated as separate special type entries in the dictionary database. Two types of MLUs are treated in the database: phraseological and non- phraseological. Upon exporting from the database for the published version, however, MLUs are included in the entries of their relevant constituents (*presti kot mačka* is shown in the entries *presti* and *mačka*), where two "nests" are available: phraseological nest (contains phraseological and also paremiological MLUs) and non-phraseological nest (contains terminological and non-terminological MLUs). For these MLUs sense, syntactic and stylistic information are provided. The MLUs with no semantic shift compared to their constituents, are included among the collocations, as there is no other semantic or other information to be presented to the dictionary users.

Boundaries among various MLU types are not always clear. For example, it is often unclear whether a certain MLU should be treated as a terminological MLU with a meaning on its own even after a detailed discussion with specialists. Language specific issues also arise: in standard Slovenian language masculine forms of classifying adjectives differ from non-classifying adjectives by the addition of an *-i* ending. Followed by a substantive they form a MLU (which is not so often true for non-classifying adjectives). However, the classifying and non-classifying adjective forms may coincide due to phonetic development in non-standard varieties of Slovenian, therefore the use in written (and spoken) standard language is often erroneous, which causes lexicographers problems in certain cases, where any of these forms can be (or actually are) reasonably used – unfortunately, it is sometimes impossible to deduct the real meaning of the text, therefore in eSSKJ these types are represented as collocations with both adjective forms (*plišast/plišasti kužek*) as both are possible in real language.

Keywords: multi-word lexical units, lexicography, metalexicography

A Semantic Web Approach to Modelling and Building a Bilingual Chinese-Italian Termino-ontological Resource

Silvia Piccini, Andrea Bellandi, Emiliano Giovannetti
Istituto di Linguistica Computazionale "A.Zampolli" (ILC-CNR) Pisa, Italy
name.surname@ilc.cnr.it

This paper introduces a bilingual Chinese-Italian onto-terminological resource[1], devoted to modelling the Chinese terminology of Matteo Ricci's World Map (1602), together with the Italian translation by Pasquale D'Elia (D'Elia 1938). The Map (Figure 1) was created in collaboration with the Chinese mathematician and astronomer Li Zizhao, and is entitled 輿 萬 國 全 圖 圖 *Kunyu Wanguo Quantu* (literally "Map of the Ten Thousand Countries of the Earth").

Its publication in China was significant as it was the first map to show the Americas, and to represent the world as a sphere. Its large number of cartouches provide information about the geography, history and customs of the world at that time as well as cosmological and cosmographic data. The map had a revolutionary impact from a linguistic standpoint as well: a large number of neologisms were introduced by Ricci, many of which have survived until today.

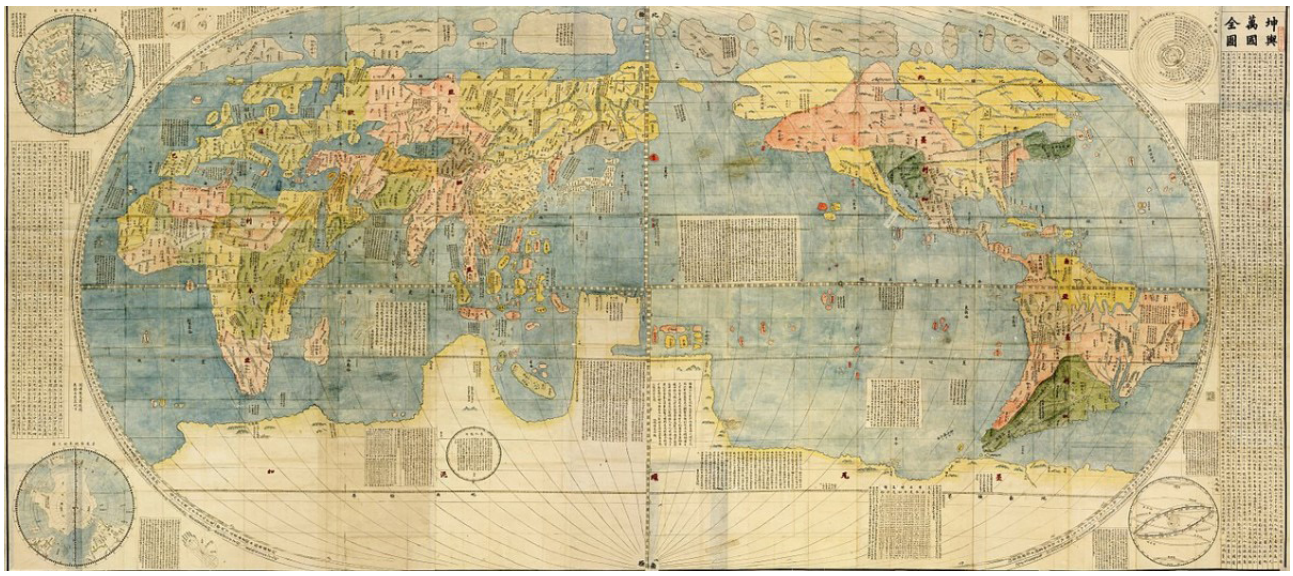


Figure 1: An anonymous color edition of the Ricci's World Map *Kunyu Wanguo Quantu* (1602)

In our resource the conceptual and the terminological levels are separated although intimately linked, in accordance with recently developed paradigms (Roche 2012) and methodologies (Després & Szulman 2008). These levels have been described using two key Semantic Web technologies, i.e. the Web Ontology Language (OWL) and the Resource Description Framework (RDF). In the ontological level concepts and properties have received a structured and formal representation using OWL-DL. Instead the morphological and semantic description of the terms composing the terminological level are based on the lexical model *lemon.*[2] This latter allows for the publication of lexicons in RDF and

1 The resource can be consulted and downloaded from: http://lexo-dev.ilc.cnr.it:8080/TMLexicon. It currently includes 81 Chinese lexical entries and 78 Italian lexical entries, for a total of 368 elements between entries, forms, senses and ontological classes.

2 http://lemon-model.net/lemon#

constitutes a standard *de facto*, as it uses W3C and ISO standards, such as LMF (Lexical Markup Framework) (Francopoulo et al. 2006), LexInfo (Cimiano et al. 2011) – this latter aligned with ISOCat - and LIR (Linguistic Information Repository) (Montiel-Ponsoda et al. 2008).

Our resource is well-placed to contribute to Lexicography research, since it mirrors the recent focus on publishing and sharing lexical resources in an open interconnected Web of Data (McCrae et al. 2011; McCrae et al. 2014). In addition it can contribute to boosting the surprisingly scarce presence of Chinese in the Linguistic Linked Data cloud (Lee & Hsieh 2015; Fang et al. 2016).

Following the *lemon* model, each Italian and Chinese lexical entry is instance of the class *Lexical Entry*, and through the use of the relation *Entry* they are associated to *Italian Lexicon* and *Chinese Lexicon*, instances of *Lexicon*. Thus, the terms can be grouped on the basis of their associated language, indicated in the model by the ISO code 639.

The relationships *canonicalForm* and *otherForm* link the lexical entries to all their forms occurring in the text. The morpho-syntactic properties of the lexical forms are described in detail (POS, gender, tense, etc.). Each lexical entry is associated with one or more senses as in the case of polysemous words. The lexical sense, instance of the *lemon* class *Lexical Sense*, is defined by a set of lexico-semantic relations expressing the paradigmatic relations among terms (hypernym, synonym, approximateSynonym, etc.). In addition, the lexical sense is linked through the relation *reference* to a concept of an ontology, where the conceptualization of the world prevailing in China in the XVII century as well as the conceptualization of D'Elia in XX century are formally represented. Each Chinese term is provided with the definitions in French and English taken respectively from (Covreur 1890; Mathews 1943).

Modelling the seventeenth-century Chinese language constituted a springboard to tackle lexicographic issues concerning the adoption of models designed mainly for Western languages, so as to create lexicons in languages which are typologically different from the so-called "standard average European". More appropriate linguistic categories were introduced to model specific features of the Chinese language, such as: two classes, *Nominoid verb* (名 動詞) and *Nominoid adjective* (名 形 詞), which denote verbs and adjectives sharing characteristics with nouns (研究 "search / to search", 天 "sky / celestial"); two sub-properties of the lemon *representation*, *Pinyin Transliteration* 拼音 and *Zhuyin Fuhao Transliteration* 注音 符號; the class *Prosodic Property*, which subsumes the class *Tone* instantiated by the four tone characteristics of classical Chinese; the property *radical*, representing the semantic units used to classify the graphemes already in the Shuowen jiezi by Xu Shen; the class *Localizer* expressing the relative position of objects (中 "middle" , 東 "east", etc.).

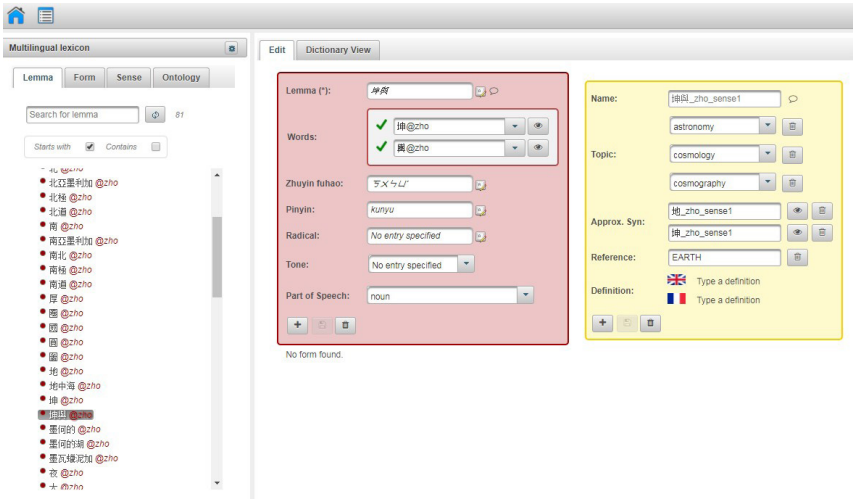


Figure 2: The lexical entry 坤與 *kunyu* "earth" in the editor LexO

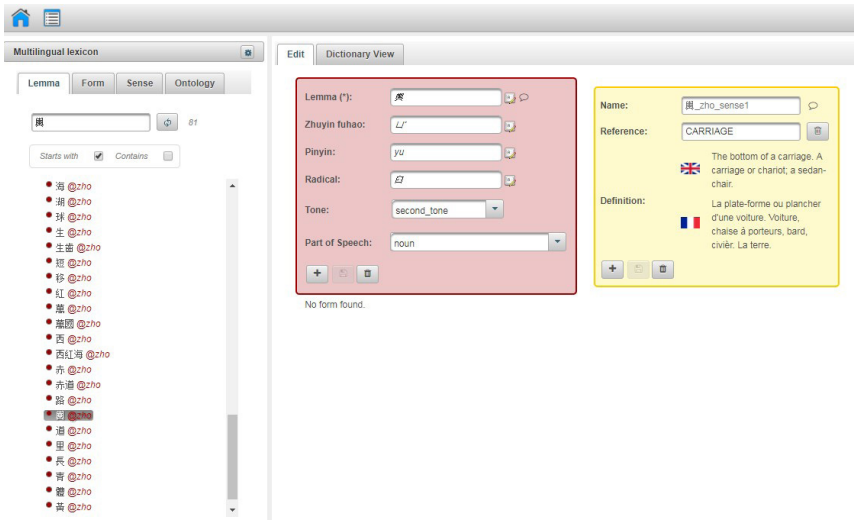


Figure 3: The lexical entry 坤 *kun* “earth” in the editor LexO

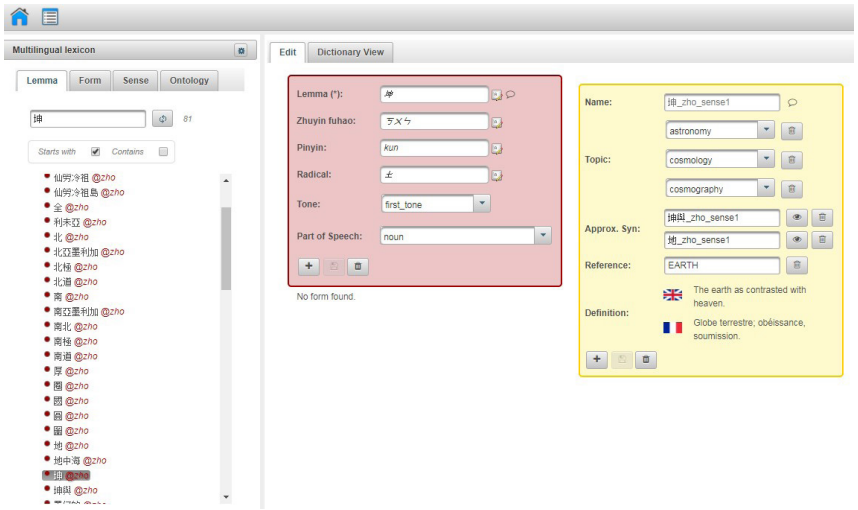


Figure 4: The lexical entry 與 *yu* “carriage” in the editor LexO

In Figure 2. the term *kunyu* (composed of *kun* “earth” Figure 3. and *yu* “carriage” Figure 4.) is described. It reflects the ancient cosmological theory *gaitian* according to which the earth is flat and assimilated to a cart. Nonetheless this very same term is used by Ricci to indicate the earth as spherical. This conceptual shift is lost in the translated word »terra«, but re-emerges explicitly when comparing the formalisation of the concepts EARTH_zho and EARTH_ita (Figure 5).

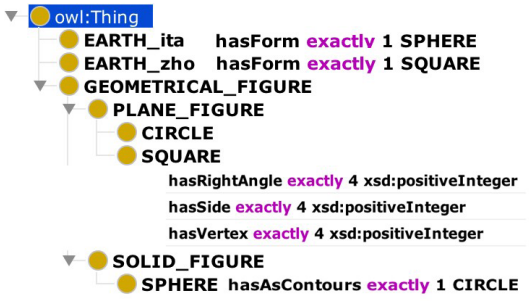


Figure 5: The formalisation of the concepts EARTH_zho and EARTH_ita, visualised in Protégé

In this way it is possible to gain insight into the lexical and conceptual work of Ricci, in translating new concepts using the original Chinese words, and how they are interpreted in Italian. This resource makes it possible to fully understand the revolutionary impact of this historical artefact by explicating the complex relationship between language, culture and thought, and could offer many other interesting possibilities of investigating lexical and conceptual data.

Keywords: lexicography, computational terminology, onto-terminological resource, lemon, semantic web, linguistic linked data, classical Chinese

References

Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M. (2011). "Lexinfo: A declarative model for the lexicon-ontology interface". *Web Semantics: Science, Services and Agents on the World Wide Web* 9 (1), 29-51.

Couvreur, S. J. (1890). *Dictionnaire classique de la langue chinoise*. Ho-kien-fou, impr. de la Mission catholique.

D'Elia, P. (1938). *Il mappamondo cinese del P. M. R., conservato presso la Biblioteca Vaticana, commentato, tradotto e annotat*o, Città del Vaticano, XXVI-275.

Desprès, S., Szulman, S. (2008). "Réseau terminologique versus Ontologie". *Actes de la deuxième conférence TOTh*, Annecy, 5 et 6 juin 2008, 17-34.

Fang, Z., Wang, H., Gracia, J., Bosque-Gil, J. and Ruan, T. 2016. "Zhishi.lemon: On Publishing Zhishi.me as Linguistic Linked Open Data". In Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.), *The Semantic Web – ISWC 2016: 15th International Semantic Web*, Springer, 47-55.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C. (2006). "Lexical markup framework (LMF)". *Proceedings of the Fifth International Conference on Language Resources and Evaluation* LREC 2006, Genova, 233-236.

C.-Y. Lee and S.-K. Hsieh. Linguistic linked data in chinese: The case of chinese wordnet. ACL-IJCNLP 2015, page 70, 2015.

Mathews, R.H. (1943). *Mathews' Chinese-English Dictionary*. China: China Inland Mission Press.

McCrae, J., Spohr, D., Cimiano Ph. (2011). "Linking Lexical Resources and Ontologies on the Semantic Web with Lemon". In Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.), *Proceedings of the 8th extended Semantic Web conference* (ESWC-11), 245–259. Heidelberg: Springer.

McCrae, J.P., Fellbaum, C., Cimiano, P. (2014). "Publishing and linking WordNet using RDF and lemon". In *Proceedings of the Third Workshop on Linked Data in Linguistics*.

Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W. (2008). "Modelling multilinguality in ontologies". In *Proceedings of the 21st International Conference on Computational Linguistics* (COLING).

Roche, C. (2012). "Ontoterminology: How to unify terminology and ontology into a single paradigm". *Proceedings of Eight International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA). Istanbul. Turkey, 2626-2630.

Russian Academic Dictionaries of the 20th and Early 21st Centuries: the Possibility of an Integral Description in a Dictionary Database

Elizaveta Puritskaya
Institute for Linguistic Studies (ILI), Russian Academy of Sciences
purichi@list.ru

The report examines the possibilities and issues of the integrated description of the data available in the Russian language academic dictionaries (the 20th – early 21st centuries) by means of the Electronic Dictionary Database (EDD). The lexical system of the Russian language is constantly changing. Particularly significant changes occur in the late 20th and early 21st centuries. The global cultural, social, and economic transformations defining the open character of modern society lead to numerous borrowings and rapid obsolescence of many words and expressions; the democratization of language intensifies the "vertical mobility" of stylistically low vocabulary. These processes require timely reflection in the dictionaries and topical normative assessment.

In the information age the role of dictionaries in society is also changing, as are the scenarios for their use: dictionaries migrate from bookshelves to the Internet and mobile devices. Digital technologies open up new opportunities in the creation of dictionaries: a departure from the fixed structure of the dictionary and the dictionary entry, various types of lexicographic content, rapid addition and modification of lexicographic information, openness of the project to the public. However, the currently available Russian digital lexicographical projects are a simple electronic copy of one or more dictionaries.

The most convenient for the user today is not a classical dictionary, even an electronic one, but a dictionary database providing the widest lexical range, a detailed description of the word and inclusion of a broad illustrative material.

Such an EDD is being developed at the Institute for Linguistic Studies (ILI) of the Russian Academy of Sciences on the basis of the "Great Academic Dictionary of the Russian Language" (published since 2001, Volumes 1 to 23 up to date). In such a database not only new and actual phenomena of the Russian language can be quickly reflected: a retrospective description of words is also important. In the Russian academic lexicography the conventional boundaries of the modern Russian literary language are drawn "from Pushkin to our days". Therefore, the database primarily contains the dictionaries created according to this principle, that is, the dictionaries of the 20th century (starting with the "Dictionary of the Russian Language" edited by D. Ushakov, in 4 volumes, 1934-1940). The report examines the main structural elements that allow to create an integral description of a lexical unit combining the materials of these dictionaries in such a way that the users of the database were able to compare the interpretations of words, normative and stylistic labels and various forms in which the word is used in speech.

One of the structural elements of the large academic dictionaries is the reference part containing: 1) information on the first fixation of the word in Russian dictionaries since the 18th century and in the dictionaries of the Old Russian language; 2) information about changes in the forms, pronunciation, and writing of the word from the beginning of the 19th century until now; 3) information on the etymology of foreign words; 4) obsolete and other grammatical forms that do not correspond to the modern norm.

However, the reference information about the word is presented very briefly due to large volume and complex structure of the dictionary entry in the classical dictionary.

The database contains information on the semantic characteristics of words in the dictionaries of different time (e.g., in the form of a comparative table); in the reference part it expands information about other forms of spelling and pronunciation of the word (distinct from the normative ones) with references to sources and usage; it provides information on the normative and stylistic labels a word was marked in academic dictionaries in the 20th century. In the case where the actual status of the word is neutral, information on the history of the label is given with a special icon. The work on combining information about a word from different sources in one dictionary entry in the database, in addition to technical difficulties, has been associated with a number of substantive problems. To combine information about the semantic characteristic of words in different dictionaries is a difficult task primarily because of the differences between these dictionaries. In the Russian academic tradition, dictionaries are divided into three types: large (more than 10 volumes), medium (4 volumes) and small (single-volume). The volume of the dictionary determines the uneven depth the semantic development of a word can take: a word can be presented as a polysemantic unit, its meanings can be combined, some of the meanings can be omitted, developed as a nuance of meaning, etc. Differences in the interpretations and normative and stylistic characteristics of words in different dictionaries are also determined by different approaches of the authors to normality, as well as by changing the status of the word itself in the literary language system. In addition, differences in the interpretation of words are also determined by the inevitable subjectivity of lexicographer's work. The comparison of the semantic characteristics of the word and the formulation of normative and stylistic labels in different dictionaries enables to trace changes in its semantics and stylistic status and also to see how the interpreted word was perceived by lexicographers in different epochs.

Keywords: academic lexicography, dictionary, dictionary database

Semi-automatic extraction of processes affecting beaches from a specialized corpus

Juan Rojas-Garcia, Riza Batista-Navarro, Pamela Faber
University of Granada, University of Manchester, University of Granada
juanrojas@ugr.es, riza.batista@manchester.ac.uk, pfaber@ugr.es

EcoLexicon (http://ecolexicon.ugr.es) is an electronic, multilingual, terminological knowledge base (TKB) on environmental sciences. Since most concepts designated by environmental terms are multidimensional and dynamic (Faber, 2011), the flexible design of EcoLexicon permits the contextualization of data so that they are more relevant to specific subdomains, communicative situations, and geographic areas (León-Araúz et al., 2013). However, to facilitate the geographic contextualization of concepts such as those belonging to the semantic category of LANDFORM, it is necessary to know what typical processes – natural or artificial – affect each type of landform according to the research papers published by experts, depending on its geographical location, and how the processes interact each other (wave erosion, accretion, hurricane-induced erosion, etc.).

This paper describes a semi-automatic method for extracting knowledge about processes affecting beaches as a type of landform, from a specialized corpus of English Environmental Science texts. The Standford Named Entity Recognizer (NER) (Finkel et al., 2005) was first applied to an initial corpus on environmental sciences, manually collected for EcoLexicon, consisting of 24 million tokens. The NER automatically labels sequences of words in the corpus which are the proper names of three entities: PERSON, ORGANIZATION, and LOCATION. This research targeted the entities annotated as LOCATION since these include the names of beaches (Sound Beach, Barcovan Beach, Black River Beach, etc.).

For all the named beaches recognized in the corpus, their respective geographic coordinates, i.e. longitude and latitude, were automatically retrieved from *Google Maps* API and then automatically visualized on top of a static map, with rectangles that further informed about the occurrence frequency of each named beach in the corpus – the darker the rectangle, the larger the frequency of occurrence (see Figure 1). This type of visualization accounted for the representativeness of the corpus in reference to the location of beaches and the number of times that they were mentioned. Moreover, a hierarchical clustering technique was deployed in order to group the named beaches, based on their latitude and longitude. This allowed us to automatically annotate each beach with the geographical area (Florida, California, Spain, The Netherlands, etc.) it belongs to.

For each beach, the documents in which it appeared were extracted from the initial corpus. Then, additional research papers for each beach were automatically retrieved from Scopus Database, by means of the *Crossref REST API* (Chamberlain, 2016), in such a way that the named beach was mentioned in the title of the documents, and they belonged to the discipline of Environmental Sciences. The compiled corpus consisted of 10 million tokens.

Subsequently, the corpus was uploaded to the term extractor *TermoStat* (Drouin, 2003). The search was set to multiword terms (MWT), and the candidate MWTs list outputted was inspected for manually selecting the appropriate MWTs, which were added to the list of MWTs recorded in EcoLexicon database. Furthermore, MWTs referring to environmental processes were extracted from the Environment Ontology ENVO (Buttigieg et al., 2013; Buttigieg et al., 2016). After lemmatizing the corpus,

all the MWTs previously collected were automatically matched in the corpus and underscored in the programming language R (R Core Team, 2017).

Then, a document-term matrix of co-occurrences was obtained (named beaches in the rows, and term in the columns), where only the terms evoking processes were manually selected and then transformed into binary variables (presence vs. absence). Finally, the clustering technique ROCK for categorical variables (Guha et al., 2000) was adopted to group the named beaches, based on the processes that affect them, as reflected in the corpus data. In addition, an association rules machine learning method was also employed to discover relations between the processes in the form of "if X and Y, then Z".

The preliminary results show that there is a slight association between the geographical areas of the named beaches and the processes mentioned by researchers affecting them. Furthermore, a set of interesting association patterns reveals connections between the processes, such as "if *sea level rise* and *storm*, then *beach erosion*", "if *river discharge* and *water elevation*, then *surface current*", or "if *cliff erosion*, then *beach sediment*". Once these experimental results were validated by Coastal Engineering experts, the knowledge extracted with this method facilitates the geographical contextualization of EcoLexicon with regard to beaches, in the sense that a specific named beach can be linked to its more highly associated processes dealt with in the corpus data.

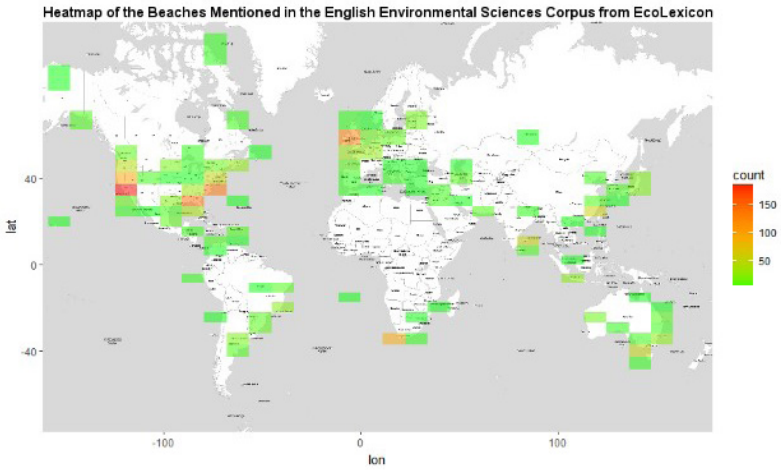


Figure 1: Heatmap of the beaches mentioned in the English Environmental Sciences corpus from Ecolexicon Database.

Keywords: terminological knowledge base, geographical contextualization, beach, conceptual information extraction, text mining

References

Buttigieg, P. L.; Morrison, N.; Smith, B.; Mungal, C. J., and Lewis, S. E. (2013). The environment ontology: contextualising biological and biomedical entities, *Journal of Biomedical Semantics*, 4:43.

Buttigieg, P. L.; Pafilis, E.; Lewis, S. E.; Schildhauer, M. P.; Walls, R. L., and Mungal, C. J. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation, *Journal of Biomedical Semantics*, 7:57.

Chamberlain, S. (2016). *crminer: Fetch 'Scholary' Full Text from 'Crossref'*. R package version 0.1.5.9110. URL: https://github.com/ropensci/crminer

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage, *Terminology*, 9 (1): 99-115.

Faber, P. (2011). The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-action Interface, *Terminology*, 17 (1): 9-29.

Finkel, J. R.; Grenager, T., and Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan: ACL 2005, 363-370.

Guha, S.; Rastogi, R., and Shim, K. (2000). ROCK: A Robust Clustering Algorithm for Categorical Attributes, *Information Science*, 25 (5): 345-366.

León-Araúz, P.; Reimerink, A., and Faber, P. (2013). Multidimensional and Multimodal Information in EcoLexicon. In A. Przepiórkowski, M. Piasecki, K. Jassem, and P. Fuglewicz (eds.), *Computational Linguistics*. Berlin, Heidelberg: Springer, Studies in Computational Intelligence, 458: 143-161.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Corpus linguistics and lexicography: exploring and extending their synergy to word formation description via the example of Modern Greek

Paraskevi Savvidou
National and Kapodistrian University of Athens
psavvidou@phil.uoa.gr

The present paper explores the contribution of corpus linguistics in the lexicographic description of word formation. Firstly, it attempts to demonstrate that word formation description is rather a neglected area of the synergy between corpus linguistics and lexicography as well as to propose an explanation of this 'gap'. Secondly, it proposes a preliminary methodology for applying corpora more adequately in this particular area of (meta)lexicography, based on the results from a study on Modern Greek word formation.

The starting point is the observation that corpus linguistics has not been applied in the lexicographic description of word formation in a way parallel to other aspects of (meta)lexicography. The first studies on corpus-based lexicography seem to rather neglect the word formation part, i.e. the potential contribution of corpus linguistics to revising the entries of the affixes or compound parts as individual lemmas of a dictionary, as well as the description of the constructed words. Moreover, the relevant studies are very few to date, in comparison with research on other aspects of a exicographic description (e.g. phraseology). In order to explore the reasons for this lack of interest, we attempt a historical overview of the three involved fields, namely corpus linguistics, lexicography and word formation morphology. We argue that the extent and the ways of applying corpora in word formation study (aiming both its theoretical and lexicographic description) seem to be restricted by a latent distinction between the formation and the use level, which is associated with the relevant dichotomies between grammar and lexis, as well as between form/structure and semantics. Morphology as a field that traditionally was giving emphasis on the first parts of these pairs was excluded from the primary fields of interest of corpus linguistics, as the later demonstrates the inseparability of the two parts of the above distinctions, by giving primacy to the second one. In addition, the application of corpus approach to this field was partial, as emphasis was given on aspects of word formation that concern more the 'grammar' part, as productivity, without attempting to associate it with meaning, and on corpus methods that exploit mainly the quantitative aspect of corpus linguistics and not the combination of qualitative and quantitative analysis.

Given the fact that corpus linguistics was unfolded as a perspective that goes beyond theoretical assumptions and distinctions that do not come from data analysis, we argue that the extent and the ways of applying corpora in the (meta)lexicography of word formation can be seen as a consequence of what Sinclair (2004) used to call restrictions of the pre-computer age. Therefore, the ways to overcome the limitations lie in the corpus linguistics theory itself and the application of all of its principles and techniques in the lexicographic description of word formation. In order to propose a preliminary methodology towards this direction, we draw evidence from a study on modern greek. We analyze an indicative number of affixes and compound parts in the data of the Corpus of Greek Texts (CGT). CGT consists in approximately 30,000,000 words from a wide range of genres of written and oral texts (see Goutsos 2010). The elements nder examination are the following: *psilo- (ψιλο-), theo- (θεο-), -istik(os) (-ίστικ(ος)), - iatik(os) (-ιάτικ(ος)), megalo- (-εγάλο-), mega- (-έγα-)* and *-iar(is)*

(-ιάρ(ης)). These elements were selected in order to include morphemes from various morphological categories (prefixes, suffixes, compound parts), with different kind of meanings (descriptive, evaluative) and from various types of vocabulary (general, scientific).

We argue that the presupposition of a methodology for applying corpora in the lexicography of word formation is a corpus-based methodology for studying affixes, compound parts and morphological rules, that begins from their meaning(s) identification and is extended to all other aspects of morpheme behavior for each meaning; these aspects include combinatoriality by defining the characteristics of bases that the morpheme tends to select (semantic characteristics, including connotations and semantic prosody, grammatical categories, register), as well as productivity and frequency both in the total corpus as well as in each individual text type. Moreover, the characteristics of the constructed words, such as grammatical categories etc., will be defined too. We present the ways that individual corpus methods can be applied in the study of all the above characteristics and how the results can provide evidence to the lexicographic description of word formation, contributing both to the macrostructure (i.e. selection of lemmas) as well as the microstructure (i.e. entry information) of a dictionary. The proposed methodology is based on the results of the comparison between our findings about the behavior of the morphemes under examination and the corresponding descriptions of Modern Greek dictionaries. In summary, we argue that the contribution of corpus linguistics in this field consists in the possibility of the combination of qualitative and quantitative analysis in the study of all the aspects of morpheme behavior as well as in the integrated analysis of formation and use level. By the adoption of these two methodological principles, corpus linguistics could revise the (meta) lexicography of word formation in a way parallel to that in which it changed our view lexicography in general, with emphasis on phraseology.

Keywords: corpus linguistics, lexicography, word formation, modern greek

Towards a new type of dictionary for Swahili

Gilles-Maurice de Schryver
Centre for Bantu Studies, Ghent University
gillesmaurice.deschryver@UGent.be

1 Problem statement

The Bantu language **Swahili** is ***the* lingua franca of East Africa**, **spoken by up to 100 million** first- and second-language speakers, especially in Tanzania and Kenya, but also in the neighbouring countries to their west and south (Mohamed 2009: iv-v). It is one of the most well-known African languages, and yet, the existing lexicographic output is the result of a century-and-a-half-old *craft* rather than a modern science (Benson 1964). In the present paper a theoretical framework is developed for modern Swahili lexicography.

2 Issues in existing swahili lexicography

What is **common to all the existing dictionaries for Swahili** is that their compilation was a fully **manual process, based on introspection**. The main strategies employed for the construction of the macrostructure were either (i) random, (ii) rule-oriented, or

(iii) enter-them-all approaches. In the random approach "words are simply added whenever they happen to cross the compiler's way", in the rule-oriented approach "a set of rules/guidelines presented in the dictionary's front matter must be followed whenever a word cannot be looked up directly" (so the assumption is that everything is covered 'in theory'), and in the enter-them-all approach "the compilers are obsessed to include all conceivable nominal and verbal derivations [working] through a modular paradigm in order to pursue such a comprehensiveness" (de Schryver & Prinsloo 2001: 219-25). **Modern dictionaries are corpus-based**, not only for the world's major languages (Hanks 2012), but also for the Bantu languages (de Schryver & Prinsloo 2000a, b), with the best aiming to be corpus-driven (de Schryver 2010). Quite surprisingly, corpora of Swahili *have* been used in lexicography, but mainly to evaluate existing dictionaries (Hurskainen 2004, De Pauw *et al.* 2009), rather than for their compilation.

3 Open questions in need of answers

the main idea is to compile a new type of dictionary, one to be **grown as a work in progress**, but to ensure a sound product at all times. This entails that attention needs to go to a number of facets: (i) **Born digital**. In this day and age, the lexicographic tool must first live in a digital environment, with only an optional transfer to paper at a later stage, not the other way round as remains all too common (eLex 2017). Is this theoretically sound for Swahili? (ii) **Define the target user groups**. Too many dictionaries are compiled without a clear picture of who the users will be. Serious thought must be given to this aspect, as it dictates several dictionary compilation decisions (Tono 2009: 39 ff.). Ideally, the project's theoretical base could cater for both native speakers and learners moving between Swahili and English, but also for speakers of Swahili who wish to remain within a Swahili environment. Is this

feasible? (iii) **Aim for a semi-bilingual reference work**. At face value, opting for a semi-bilingual approach seems like a good idea (Lew 2004). Such a work has characteristics of both a bilingual Swahili-English-Swahili dictionary, and a monolingual Swahili dictionary. But is this hunch also corroborated with actual dictionary use? (iv) **Know the users' lemmatisation needs**. No Bantu dictionary is purely word-based nor purely stem-based; all put the lexical items from the various word classes on a sliding cline between these extremes (de Schryver 2008: 86-87). With the exception of Johnson (1939), there seems to be a broad consensus on how to lemmatise the lexicon in Swahili. Tradition doesn't necessarily correspond to today's (digital) user needs, so one should dare question current practice. (v) **Let a corpus drive the compilation**. With the various preceding 'moving targets' (cf. ii to iv) as compilation proceeds (de Schryver 2005), can a corpus truly guide the macro-, medio- and microstructural compilation? (vi) **Use an existing dictionary writing system (DWS)**. Good off-the-shelf lexicographic software exists (Abel 2012), but can such packages handle all of the above? (vii) **Aim for structured dictionary compilation**. A DWS imposes a rather fixed structure; but is it flexible enough to deal with on-the-fly adaptations of the type envisaged for this Swahili dictionary project? (viii) **Study dictionary use**. In a digital environment, it is possible to unobtrusively study dictionary-use behaviour, while optionally allowing direct feedback in addition (de Schryver & Joffe 2004). In doing so, will one be in a position to check whether the target user groups are as expected (cf. ii)?, will one be able to fine-tune the exact type of dictionary type to work with (cf. iii)?, will one have the means to adapt the lemmatisation strategies (cf. iv)?, will one be able to judge whether the use of a corpus is the right approach (cf. v)?, will one be able to translate the feedback into feasible changes to the DTD or XML schemas (cf. vi)?, and will one end up with a unified overall structure (cf. vii)? The answers to these questions form the base for the sought theoretical framework.

Keywords: Swahili, digital dictionary, semi-bilingual, corpus-driven, user-friendly

References

Abel, Andrea. 2012. Dictionary Writing Systems and Beyond. In: Granger, Sylviane & Magali Paquot (eds). *Electronic Lexicography*: 83–106. Oxford: Oxford University Press.

Benson, T. G. 1964. A Century of Bantu Lexicography. *African Language Studies* 5: 64–91.

De Pauw, Guy, Gilles-Maurice de Schryver & Peter W. Wagacha. 2009. A corpus-based survey of four electronic Swahili–English bilingual dictionaries. *Lexikos* 19: 340–52.

de Schryver, Gilles-Maurice. 2005. Concurrent over- and under-treatment in dictionaries – The *Woordeboek van die Afrikaanse Taal*

as a case in point. *International Journal of Lexicography* 18(1): 47–75.

de Schryver, Gilles-Maurice. 2008. A new way to lemmatize adjectives in a user-friendly Zulu–English dictionary. *Lexikos* 18: 63– 91.

de Schryver, Gilles-Maurice. 2010. Revolutionizing Bantu lexicography – A Zulu case study. *Lexikos* 20: 161–201.

de Schryver, Gilles-Maurice & David Joffe. 2004. On How Electronic Dictionaries are Really Used. In: Williams, Geoffrey & Sandra Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*: 187–96. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2000a. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 1: The macrostructure. *South African Journal of African Languages* 20(4): 291–309.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2000b. Electronic corpora as a basis for the compilation of African-language dictionaries, Part 2: The microstructure. *South African Journal of African Languages* 20(4): 310–30.

de Schryver, Gilles-Maurice & D.J. Prinsloo. 2001. Towards a sound lemmatisation strategy for the Bantu verb through the use of frequency-based tail slots – with special reference to Cilubà, Sepedi and Kiswahili. In: Mdee, James S. & Hermas J.M. Mwansoko (eds). *Makala ya kongamano la kimataifa Kiswahili 2000. Proceedings*: 216–42, 372. Dar es Salaam: TUKI, Chuo Kikuu cha Dar es Salaam.

eLex. 2017. *Electronic Lexicography in the 21st Century*. Available online at: https://elex.link/elex2017/.

Hanks, Patrick. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398–436.

Hurskainen, Arvi. 2004. Computational testing of five Swahili dictionaries. In: Karlsson, Fred (ed.). *Proceedings of the 20th Scandinavian Conference of Linguistics, Helsinki, January 7–9, 2004* (Department of General Linguistics Publications 36). Helsinki: University of Helsinki.

Johnson, Frederick. 1939. *A Standard Swahili-English Dictionary (Founded on Madan's Swahili-English dictionary)*. Nairobi: Oxford University Press.

Lew, Robert. 2004. *Which Dictionary for Whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English*. Poznań: Motivex.

Mohamed, Amir A. 2009. *Kiswahili for Foreigners* [3rd revised edition]. Zanzibar: Goodluck Publishers.

Tono, Yukio. 2009. Pocket Electronic Dictionaries in Japan: User Perspectives. In: Bergenholtz, Henning, Sandro Nielsen & Sven Tarp (eds). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow* (Linguistic Insights): 33–67. Bern: Peter Lang.

The Category of Modality and Its Reflection in Lexicology

Nino Sharahsenidze
Ivane Javakhishvili Tbilisi State University
e-mail: nino.sharashenidze.tsu.ge

Modal semantics is present on every level of language. However, only the forms that permanently express the modal semantics in a language are considered markers of the modal category. The use of these forms always adds modal semantics to a phrase or discourse. Such forms exist in every language, and their reflection in lexicography is an important issue. As for the Georgian language, the analysis of the category of modality is still underway. Therefore, it is an urgent task to study and analyze the forms expressing modality. Another interesting issue is related to the reflection of these forms in lexicography. The paper analyzes the data of earlier and contemporary dictionaries. The paper focuses on the following questions: how modal semantics must be reflected in dictionaries and which form of lexicographic reflection is more perfect? In this regard, the paper outlines the development perspectives of the national corpus of the Georgian language (GNC; http://clarino.uib.no/gnc/page), the Dialectological Corpus of the Georgian Language (GDC) and the Georgian Scientific Metalanguage Corpus (SMLC) (http://www.corpora.co/#/) from the viewpoint of semantic analysis, namely, the expression of modal semantics.

In the Georgian language, the category of modality is expressed by modal forms and modal elements. In the grammar of contemporary Georgian language, these are considered as particles. Among others, these forms include *unda* (must), *šeiʒleba* (may), *ikneb* (maybe), *egeb* (might), *lamis* (nearly), *neṭav* (I wish), *mainc* (still), *albat* (probably), and *titkos* (it seems). The use of the indicative or subjunctive mood of the verb accompanied by the above-mentioned forms creates modal semantics. In the Georgian language, the category of mood is included in the complete paradigm of the verb and forms several screeves. The subjunctive forms are chiefly used in subordinate clauses alongside certain modal forms.

The modal forms are a result of grammaticalization. They chiefly consist of particles obtained from verbs. They are words devoid of lexical meaning which have turned into particles. Their function has been mentioned in „The Dictionary of Georgian Morphemes and Modal Elements“. However, this dictionary is meant for specific usage and the information provided in this dictionary is not found in the Explanatory Dictionary of the Georgian Language. For instance, the modal form „უნდა” (must). The Explanatory Dictionary points out that this lexical unit is a verb used in several phrases, as well as „an auxiliary word used in the subjunctive mood“. The given paper analyzes the dictionary data of several modal forms and proves that the information provided by the dictionaries is insufficient for the lexicographic reflection of the category of modality. Furthermore, the above-mentioned dictionaries provide an incomplete description of the semantics of modal forms.

The information provided by the dictionary should more or less fully reflect the meaning of the word. On the example of one of the analyzed forms, we argue that the dictionary should contain the following information:

unda – (must)

1. Its verbal meaning and related constructions: 1. He/she wants, 2. It is necessary.
2. The meaning of the given particle in contexts denoting modality, its usage with different verb forms (the so-called screeve forms: present subjunctive, subjunctive two, resultative two) and the following related meanings: obligation, wish, assumption, logical possibility.

3. The semantics of constructions with *unda* („must") and the negative particle: *visa car unda* (=whoever), *vin car unda* (whoever), *sadac ar unda* (wherever);
4. The meaning of constructions with "must" and interrogative pronouns: *amis gketebas ra unda* (=it's easy), *amas ak ra unda* (he/she/it is out of place here), *ra unda* (=not).
5. The meaning of phrases with the form **unda (must)**: *ra ginda sulo da gilo* (everything), *ra tkma unda* (of course).
6. The dictionary should also focus on the conjunctions obtained from the given form: *tund* (even if), *tundac* (even if), *gind* (even if), *gindac* (even if), *tugind* (even if), as they are related both in form and meaning.
7. The dictionary should also note the functional-formal change undergone by the given form. In addition, it should mention the relation between the forms: *una* > *unda* > *unebs* > *hnebavs*

The use of modal forms changes the semantics of the sentence. Thus, a glossary should provide a complete set of information on the functions of the modal form.

Semantic analysis is an important stage of processing of the language corpus. In the process of semantic analysis, this form should be underlined as a marker of modality. For this reason the marker **"mod"** should be introduced in order to distinguish the third person singular form of the verb from the modal element performing the modal function.

Keywords: modality, modal form modal meaning

Better Late than Never: Some Remarks on Usage Notes in Present-Day Czech Dictionaries

Martin Šemelík, Michal Škrabal
Department of Germanic Studies and Institute of the Czech National Corpus, Charles University
martin.semelik@ff.cuni.cz, michal.skrabal@ff.cuni.cz

In this contribution, we comment on the design of usage notes in selected Czech monolingual and, more importantly, translation dictionaries, taking the perspective of pedagogical lexicography. By the term usage notes, we mean here "[a] discursive paragraph providing additional information on a word or phrase, and inserted close to the respective dictionary entry. In general dictionaries or learner's dictionaries, usage notes, sometimes specially marked out on the page in boxed panels, draw the reader's attention to synonymous and related words or phrases, explanations of idiomatic expressions, stylistic or other restrictions on usage" (Hartmann & James 1998: 150, cf. also Whitcut 1985). Utilising various classification criteria, usage notes can be classified into diverse types, such as (a) prescriptive vs. descriptive vs. proscriptive; (b) pronunciation vs. grammatical, etc. (i.e. the level of language being the criterion); (c) internal vs. external (being/not being a part of the dictionary entry) (cf. Šemelík, Bezdíčková & Koptík 2016).

Our contribution is based upon the following assumptions:

- "A dictionary is an artefact, like a dam or a hospital: built to serve a purpose." (Whitcut 1989: 88)
- The essential point behind the making of each and every dictionary is its target users, or more specifically, the dictionary functions of the respective dictionary (cf. Tarp 1995).
- Both monolingual and translation dictionaries can be used as learning tools sui generis.
- Errors of various types are a natural part of learning a language. On the other hand, language accuracy and typicality of usage play a crucial role in the process of language acquisition. In this respect, a dictionary can be helpful in many different ways, the usage notes being one of them.

In lexicographic works of Czech provenience, usage notes are not a completely unknown lexicographic asset; nevertheless, we almost exclusively find them in the products of commercial publishing houses. Most academic dictionaries of both the past and the present day (including the forthcoming academic dictionary of current Czech) seem to neglect them. In the Czech Republic, this fact is symptomatic of the mainstream approach to lexicography as a strictly linguistic discipline that views the dictionary as being, above all, an artefact of linguistic analysis rather than a tool used for solving language problems – even though this, in fact, is what it is. The view that usage notes are only a kind of "ribbon" whose function it is to make the dictionary more "beautiful" when the lexicographers can find some time to insert them into their dictionary must be abandoned. Within the lexicographic process, the target users and their needs – or dictionary functions – play an important role. If the usage notes contribute to the already determined dictionary functions, they are not a mere "ornamentation" but, on the contrary, represent one of the pillars on which the dictionary stands.

Against the background of an analysis of the present practice as it is utilized in selected dictionaries, we attempt some concrete design proposals for the usage notes in the *Large Academic German-Czech Dictionary* and in the *Latvian-Czech Dictionary*. In this respect, many different questions arise, for instance:

- Which concrete phenomena should be addressed in the usage notes?
- Which requirements should the metalanguage used in the usage notes meet?

- Which type of usage notes do dictionary users expect? The prescriptive, descriptive or proscriptive ones?
- Which role do concepts such as standard, norm, usage and codification play in this regard?

In this connection, we posit that a large electronic corpus of L1 texts produced by L2 native speakers is urgently needed in the field of translation lexicography as these would enable a more objective (statistically relevant) analysis of language data, contributing to better quality and higher usefulness of usage notes. As regards monolingual Czech lexicography, similar data are already available in the database of the Language Cultivation Service of the Czech Language Institute that contains tens of thousands inquiries.[3] On the basis of an analysis of these publicly-made inquiries, which has already been conducted (cf. Černá et al. 2002, Pravdová 2012), it is possible to identify those difficult parts of the Czech language that pose problems to native speakers and to enrich the information structure of the respective dictionary entries by means of usage notes (cf. Prošek 2016). As the title of this contribution suggests, the time has come for a change of practice in Czech lexicography in favour of both the needs of target dictionary users and usage notes.

Keywords: usage notes, learner's dictionaries, Czech lexicography

Acknowledgements

This study was supported from the Charles University project Q10 "Language in the shiftings of time, space, and culture" and from the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

Černá, A., Svobodová, I., Šimandl, J. & Uhlířová, L. (2002). *Na co se nás často ptáte*. Praha: Scientia.

Hartmann, R. R. K. & James, G. (1998). *Dictionary of Lexicography*. London – New York: Routledge.

Pravdová, M. (ed.) (2012). *Jsme v češtině doma?* Praha: Academia.

Prošek, M. (2016). Slovník češtiny pojatý jako moderní „learner's dictionary"? In M. Lišková. V. Vodrážková & Z. Děngeová (eds.) *Akademický slovník současné češtiny a software pro jeho tvorbu aneb Slovníky a jejich uživatelé v 21. století. Sborník abstraktů z workshopu.* Praha: Ústav pro jazyk český AV ČR, p. 63.

Šemelík, M., Bezdíčková, A. & Koptík, T. (2016). Verlierer gibt es hier also keinen oder Usage notes in ausgewählten Wörterbüchern. In M. Vachková, M. Šemelík & V. Kloudová (eds.) *Acta Universitatis Carolinae, Philologica 4, Germanistica Pragensia*, pp. 173-196.

Tarp, S. (1995). Wörterbuchfunktionen: Utopische und realistische Vorschläge für die bilinguale Lexikographie. In H. E. Wiegand (ed.) *Studien zur bilingualen Lexikographie mit Deutsch II.* Hindesheim – New York: Olms, pp. 17-62.

Whitcut, J. (1985). Usage notes in Dictionaries: The Needs of the Learner and the Native Speaker. In R. Ilson (ed.) *Dictionaries, Lexicography and Language Learning (ELT Documents 120)*. Oxford: Pergamon Press – British Council, pp. 75-80.

Whitcut, J. (1989). The Dictionary as a Commodity. In F. J. Hausmann et al. (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie (Handbücher zur Sprach- und Kommunikationswissenschaft 5.1).* Berlin – New York: de Gruyter, pp. 88-94.

3 From January to November 2017 only, 5389 phone calls with 7492 queries in total were logged.

And What Does Google Say? – Web Search Results under Scrutiny: From Traditional to Web-Based Lexicography

Mojca Šorli
University of Ljubljana
mojca.sorli@guest.arnes.si

Research into dictionary use has increased dramatically in the last decade (e.g., Dziemanko 2012; Lew 2011a,c; Müller-Spitzer 2012, 2014; Nesi 2012; Welker 2010), yet many questions remain unanswered. While the potential for improving dictionary functionalities has been explored, particularly as a result of digitalisation and the advancement of language technologies, less attention has been focused on newly identifiable relationships, such as that between lexical resources on the Web and so-called "traditional" resources. It is often emphasised that contemporary dictionary users are not sensitive to the authority of professional dictionary publishers, instead prioritising ease of access and of use. However, the transition to predominantly web-based dictionaries and lexical resources has been marked by more than one paradox in a time when lexicography is moving "towards information science" (Tarp 2012; Verlinde et al. 2010). In fact, Sinclair's prediction of lexicography operating "at the intersection of Linguistics and Information Technology" goes way back (1984: 6). We may be inclined to generalise about users' search preferences along the lines of "users consult Google/the Web rather than individual (web-based or electronic) dictionaries", but such assertions may be misleading. There are two things that should be kept in mind here. The first is that a Google search merely aggregates web-based lexical resources (dictionaries, thesauruses, word reference forums, etc.) rather than offering linguistic solutions. The second is that, as a consequence, it offers links to existing (predominantly traditional) lexical resources, notably dictionaries. In the first part of the present article, we examine the first page of Google searches on selected single words and phrases and on this basis seek to establish (better: revise) the role of traditional – that is, edited – dictionaries featured in so-called "Google searches". As a search engine that can be used to organise and select web-based text according to the user's needs, Google is obviously not in itself a (lexical/linguistic) resource, but we perceive it as such because it is now a starting point for enquiries and data searches of all forms. Nevertheless, as pointed out, we will most likely be directed by Google to established (edited) lexicographic works with long publishing traditions (competing for prominence on the leading search engine pages) or to products of collaborative lexicography. Once we have clarified the distinctions (and overlaps) between "web-based" resources and their "traditional" counterparts (be it in paper or CD-ROM and DVD format), on the one hand, and between "traditional" – in the sense "professionally edited" – and "collaborative" resources, on the other, we can finally examine which aspects of the "new" dictionary in terms of content may appeal to the majority of dictionary users/compilers. Existing research into dictionary use suggests that most users appreciate aspects such as data based on actual usage, authentic examples, regular updates of data, accessibility of information, hyperlinks, clear definitions, etc. In the second part of the article, we are dealing with aspects of lexicographic description shared by collaborative lexicography and some directions of professional lexicography. Lexicographers can learn a lot about users' needs and expectations from alternative approaches to language description, such as *Urban Dictionary* or *Wictionary*. The stated aim of *Wictionary*, for example, is "to include not only the definition of a word, but also enough information to really understand it" (http://en.wiktionary.org/wiki/Wiktionary:Main_Page). It was John Sinclair who, decades ago, initiated a long tradition of empirical lexical analysis that provided a scientific

basis for expanding lexicographic description to do just that. However, despite the technological advances, much remains to be done in the way of making dictionary "a device through which the user will observe the living language ... language *through* the dictionary ... the next target of progressive lexicography" (1987: 5). This would involve shifting the focus from word-based to text-based dictionaries. In order to explore what this really means, we will, limiting ourselves to defining practices, analyse the relationship between well-established mainstream monolingual dictionaries and some examples of collaborative lexicography. Typically, collaborative lexicography includes data on circumstances of meaning, giving greater prominence to the evaluative function of words in social interaction, which is guided by social convention. We seek to demonstrate that including this kind of information in communicatively orientated lexical resources is important.

Keywords: web search, Google, dictionary, user habits, meaning description, social convention

References

Dziemanko, A. (2012). On the use(fulness) of paper and electronic dictionaries. In Electronic lexicography (320-341). Oxford: OUP.

Lew, R. (2011a). Studies in Dictionary Use: recent Developments. International Journal of Lexicography, 24(1), 1-4.

Lew, Robert (2011b). User studies: Opportunities and limitations. In K. Akasu & U. Satoru (Eds.), ASIALEX2011 Proceedings Lexicography: Theoretical and practical perspectives (7-16). Kyoto: Asian Association for Lexicography.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2012). Online dictionary use: Key findings from an empirical research project. In S. Granger, M. Paquot (Eds.), *Electronic Lexicography.* Oxford: Oxford University Press, 425-457.

Müller-Spitzer, C. (Ed.) (2014). *Using online dictionaries.* Lexicographica. Series Maior. Berlin: de Gruyter.

Nesi, H. (2012). Alternative e-dictionaries: Uncovering dark practices. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 363-378.

Sinclair, J. M. (1984). Lexicography as an academic subject. In Hartmann, R. R. K. (Ed.), *LEXeter 83 Proceedings*, Lexicographica. Series Maior No. 2. Tubingen: Max Niemeyer Verlag, 3-12.

Sinclair, J. M. (1987). *The Dictionary of the Future*. Collins English Dictionary Annual Lecture. University of Strathclyde, 6 May 1987.

Tarp, S. (2012). Do we need a (new) theory of lexicography? *Lexikos* 22, 321-332.

Verlinde, S., Leroyer, P. & Binon, J. (2010). Search and You Will Find. From Stand-alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats. *International Journal of Lexicography* 23.1. 1–17.

Welker, H. A. (2010). *Dictionary use: a general survey of empirical studies*. Brazília: Eigenverlag.

My first Brazilian Sign Language Dictionary

Janice Gonçalves Temoteo Marques, Antonielle Cantarelli Martins, Walkiria Raphael

University of Campinas (Unicamp), Universidade Federal de Pelotas – UFPel, University of São Paulo
temoteojanice@gmail.com, an.cantarellim@gmail.com, walkiria.duarte@gmail.com

This work describes the construction of the project »My first Brazilian Sign Language Dictionary« that was originated with the need presented by deaf children and their teachers to have didactic material in Libras that could be used with their students at school and with family. As a theoretical framework, this work is based on issues related to Lexicography and Pedagogical Lexicography with the aim of developing studies to enhance the use of lexicographic works as pedagogical material to be used in the classroom. Therefore, the dictionary was constructed in a didactic proposal with the focus on the children's consultant, with vocabulary of basic signs of Libras that includes contents since the initial series to Elementary School I, as well as signs used in the daily life of a deaf student. The dictionary consists of 1,500 entries indexed in alphabetical order presented in Portuguese and in Libras, a bilingual dictionary and four children signing characters illustrate the signs. One of the objectives of this poster it is to show how was the dictionary-making process, since the composition of the entry that is composed of six main parts (1. Illustration of the meaning of the sign; 2. Digital spelling; 3. Glosa in Portuguese; 4. Definition and example of a sentence; 5. Sign illustration in Libras; 6. Sign language phonology) to the final structure of this sign language dictionary. Libras is a space-visual language and, for this reason, the register of a sign is presented in stages, with the help of arrows, in addition to the description in Portuguese of how the sign should be articulated. The dictionary has alphabetic index (A to Z) and semantic index that lists the signs grouped into semantics categories. According to the Census of the Brazilian Institute of Geography and Statistics (2010), there are more than 9,722,163 people with hearing impairment in the country. Therefore, it is expected that this dictionary could contribute to the education and citizenship of the Brazilian deaf, being an aid tool for Libras' teachers in the preparation of their classes, as well as contributing to lexicographic research's in Brazilian Sign Language.

Keywords: Brazilian sign language (Libras), sign language dictionary, dictionary, lexicography

References

Instituto Brasileiro de Geografia e Estatística (IBGE, 2010). In: https://ww2.ibge.gov.br/home/

Welker, A. H. Lexicografia Pedagógica: Definições, história, peculiaridades. In: ATARA, C., Bevilacqua, C. & Humblé, P. (org.). Lexicografia Pedagógica: pesquisas e perspectivas. Florianópolis: UFSC/NUT, 2008, p. 9 – 45.

Meu primeiro Dicionário de Libras

Maçã

Fruto da macieira de sabor doce com polpa firme e casca brilhante de cor vermelha, verde, ou amarelada, rica em vitamina B. Ex.: Comer uma maçã por dia ajuda na prevenção do câncer de pulmão.

Mão horizontal em C, diante da boca, movendo a mão para cima e para baixo.

Macaco

Animal primata, mamífero, salteador, de grande ou pequeno porte, de acordo com a espécie. Ex.: Os macacos são conhecidos por gostarem de comer banana.

Mãos abertas, dedos curvados. Mão direita tocando a cabeça, mão esquerda tocando a barriga, curvando a ponta dos dedos.

Macarrão

Massa feita geralmente de farinha, água e ovos, conhecida também como pasta, com vários formatos: espaguete, parafuso, penne, fettucine, entre outros. Ex.: O macarrão geralmente é servido com deliciosos molhos.

Mãos horizontais em 3, diante do corpo, mover as mãos em círculo para frente e depois para os lados.

Vertaalwoordenschat: an online platform for bilingual dictionaries of Dutch

Carole Tiberius, Koen Mertens & Bart Hoogeveen
Instituut voor de Nederlandse Taal
{carole.tiberius,koen.mertens}@ivdnt.org

This demo presents a free online bilingual dictionary platform for Dutch, called the Vertaalwoordenschat, which is available at http://www.vertaalwoordenschat.ivdnt.org. Around the turn of the century, a number of bilingual lexical resources has been produced with Dutch as the source or target language under the auspices of an intergovernmental committee of lexical experts (1993-2003). See Martin (2007: 228) for a full overview. It mainly concerned language pairs which would not normally be addressed on the commercial market, but which were socially speaking highly relevant (e.g. for cultural identification and social integration). The Committee set itself a goal to not only compile dictionaries, but to also develop multifunctional and reusable electronic lexical databases for these language pairs. As such the resulting lexical databases contain information useful for different types of lexicons. For instance, information such as free text definitions, lexicographic comments and descriptions are mainly useful for human use, whereas semantic type, example types and complementation patterns may be more useful for computational applications and information like lemma, word form, part of speech, pragmatic labels, collocations, idioms, translation equivalents can be used by both humans and computers. (Maks et al. 2008:1723; Tiberius et al. 2010)

The resources for the different language pairs vary as to size and contents, but as a rule they use the same Dutch source as a base, i.e. the Reference Database of Dutch (RBN; van der Vliet 2007). The RBN is a lexical database based upon modern Dutch written corpora; it contains about 45,000 entries and more than 90,000 example sentences; it has a rich and explicit microstructure describing orthographical, morphological, syntactic, collocational, semantic and pragmatic features of the Dutch entries (Martin 2007:230).

In addition, a subset of the bilingual resources was compiled using the dictionary tool OMBI (Omkeerbare Bilinguale Bestanden = Reversible Bilingual Lexical Databases) specifically designed for creating and editing rich multi-purpose bilingual resources (Maks 2007, Martin and Tamm 1996). One of the most distinctive features of this tool was the reversal of source and target language at sense level. As a result the resources created in OMBI have a very similar data format and could potentially be linked following Martin's hub-and-spoke model (Martin 2013).

The compilation of the bilingual lexical databases started in 1998 and the first was completed in 2002. Several bilingual printed paper dictionaries have been derived from these resources and have been published since. However, fifteen years on, many of the print dictionaries have gone out of print and will not be published as a print dictionary again. Consequently, these resources and the information contained in them are no longer available to users. To overcome this, the Dutch Language Union has handed the resources over to the Dutch Language Institute at the beginning of 2017, commissioning the institute to make the resources available to the public through an online platform within the limits of prevailing copyright and publication licences. In September 2017, the online platform, the Vertaalwoordenschat, was released with Dutch – Modern Greek. In spring 2018, the next language pair has been released, i.e. Dutch – Portuguese and the release of Dutch-Estonian is expected at the end of 2018. The platform offers a simple but efficient search option and is used on a daily basis. In the

future we aim to add more languages and to extend the functionalities of the platform, i.e. linking the different databases to each other, linking to external resources, etc.

Keywords: bilingual lexicography, online, Dutch

References

Maks, I., C. Tiberius & R. van Veenendaal. 2008. Standardising bilingual lexical resources according to the Lexicon Markup Framework. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Maks, I. 2007. OMBI: The practice of Reversing Dictionaries. *International Journal of Lexicography* 20.3.

Martin, W. 2013. Reversal of bilingual dictionaries. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography.*1445-1455. Berlin/Boston: De Gruyter Mouton.

Martin, W. 2007. Government Policy and the planning and production of bilingual dictionaries: the 'Dutch' approach as a case in point. *International Journal of Lexicography* 20.3.

Martin, W. and Tamm A. 1996. OMBI: an editor for Constructing Reversible Lexical Databases. *Euralex '96 Proceedings I-II*, Göteborg University.

Tiberius, C., A. Aalstein & J. Hoogland. 2010. OMBI bilingual lexical resources: Arabic-Dutch / Dutch-Arabic. *Proceedings of the XIV Euralex International Congress (EURALEX 2010)*, Leeuwarden.

Vliet, H. van der. 2007. The Referentiebestand Nederlands as a Multipurpose Lexical Database. *International Journal of Lexicography* 20.3.

Romanian-Slavonic Lexicons from the XVIIth Century. The Project of a Comparative Study

Mădalina Ungureanu, Mihai Alex Moruz
»Alexandru Ioan Cuza« University, Iași, Romania
madandronic@gmail.com, mmoruz@info.uaic.ro

Romanian lexicography of the XVIIth century is, for the most part, inspired by the most prestigious Slavonic lexicon of the era, the lexicon written by Pamvo Berynda (Lexikon slavenorosskij i imenь tlьkovanije, Kiev, 1627). This lexicon contains 6982 words, in two parts. The first part (approx. 4980 words) contains Russian and Slavonic words, but also words from Greek, Latin, Polish, German, Slovak and Czech. Definitions vary from simple words to encyclopedic definitions, sentences, source citations, stylistic notes. The second part is a list of 2002 proper names, toponyms, names from Greek mythology, object and scientific terms in Hebrew, Greek and Latin. At the time of its publishing, lexicography in the Romanian principalities contained a set of Romanian glosses for a Slavonic text (known as the Bogdan Glosses) and two fragments of a Romanian-Slavonic vocabulary from the XVIth century. The Berynda lexicon is (directly or indirectly) the starting point for six Slavonic-Romanian dictionaries (one dating from the first half of the XVIIth century, the others from the second half), which have not been studied thus far, except for the oldest of them. The links between these lexicons are interesting, and have been, until now, only assumptions; some of the authorship questions are also unanswered. Because of this, a comparative study of these lexicons is necessary; we intend to carry out this analysis within a research project to be undertaken at the "Alexandru Ioan Cuza" University of Iași. Also, such an undertaking would allow easy access to five new texts, together with the already published one (the Lexicon of Mardarie Cozianul, studied in 1900), with numerous benefits for diachronic study of literary Romanian language (under phonetic and lexical points of view) and lexicology (given the descriptive passages in the lexicons). This is also an opportunity to better understand the translation of Slavonic texts into Romanian and to extract the sense of a given word at that time, which can lead us to better understand the semantic intent of the translator. Lastly, a comparative edition of all these texts will allow determining the circulation of a text in the Romanian area, the role of the scribes, the manner of distribution of information and the repercussions of such distribution in translated texts.

The project, therefore, intends to create a comparative edition of the Slavonic-Romanian lexicons mentioned above. We will start with the prototype, Berynda's lexicon, to determine how a word is reflected in Romanian lexicons. For further disambiguation of the Slavonic word, it will be accompanied by the corresponding entry in *Lexicon Palaeoslovenico-graeco-latinum* by F. Miklosich, or another Slavonic dictionary. The edition will be provided with an index of Romanian words, with their Slavonic equivalent; such an instrument would be useful for Romanian lexicography, both as a source for first usage of a word and for sense checking. The comparative edition will allow for the study of the relations between the Romanian lexicons, and also the relation to the source. On a more general level, the publishing of the six lexicons will contribute to the better understanding of the age and the manner in which the orthodox counter reform of Petru Movilă, which implied creating tools for writing in Slavonic (grammars and dictionaries), influenced the Romanian principalities, especially since this does not hinder the orientation towards Greek culture just few decades later. The project also intends to offer an online version of the edition, a tool that will be very useful for other projects for digitizing old Romanian texts.

Keywords: Pamvo Berynda, Lexicon slaveno-rosskij, Romanian lexicography, XVIIth century

Exploring the treatment of informal usage in general bilingual dictionaries: the case of English and Estonian

Enn Veldi
University of Tartu
Enn.Veldi@ut.ee

The aim of the present paper is to explore the best lexicographic practice in English–Estonian and Estonian–English bilingual lexicography with a view to improving the coverage of informal usage and contributing to lexical enrichment of general bilingual dictionaries.The treatment of informal usage in general bilingual dictionaries is of interest for a number of reasons. First, both English and Estonian monolingual dictionaries include thousands of meanings with the usage label 'informal'. For example, the search 'informal usage' in CALD 4 provided 6102 matches; a similar search in the Estonian Explanatory Dictionary (EKSS) yielded 5147 matches. These lists were used as a starting point for selecting the meanings for the subsequent study of bilingual dictionaries. Second, informal usage reveals its characteristic patterns of word-formation, such as clipping (often in combination with affixation), initialisms, acronyms, compounding, and conversion. Informal vocabulary and slang are closely related in that their word-formation patterns are similar (for the discussion of word-formation in slang see Mattiello 2008 and Widawski 2015). Third, informal senses often form networks of expressive near-synonyms, where a good knowledge of the repertoire of near-synonyms in both languages is required for the selection of suitable equivalents. The analysis of the material for the present paper proceeded from two perspectives that complement each other.. One method is to explore informal meanings from the perspective of word-formation patterns; this method is also useful for the discovery of cross-linguistic regularities between the two languages. The second method proceeds from meaning, and its focus is on finding networks of near-synonyms in both languages and selection of suitable interlingual equivalents.

It is generally known that longer words tend to be shortened in spoken language, which results in monosyllabic and disyllabic informal equivalents. Some of these words are international, such as the English 'crocodile' and the Estonian *krokodill*), 'croc' and *kroku* being their informal equivalents. However, the practice of cross-referring clippings to their longer counterparts, but not the other way round, without any cross-linguistic equivalents, as in **croc** *fam* = **crocodile** (Silvet 2002), represents reduced treatment where the short Estonian equivalent is not provided. The Estonian–English dictionary no entry for *kroku*, and the English 'croc' does not appear under **krokodill** *zool* crocodile (Aule 2003). In the latter dictionary the treatment of the crocodile continues with a list of ten species of crocodiles together with their English equivalents, which is characteristic in that it shows that in recent decades Estonian lexicographers have focused on the treatment of scientific terminology. Another criticism concerns the case when only neutral interlingual equivalents are provided for an informal word, e.g. **bike 1** *s fam* jalgvõi mootorratas; 2 *v i fam* jalgvõi mootorrattaga sõitma (Silvet 2002), and the existing informal equivalents are neglected. In fact, Estonian has two established informal equivalents *tsikkel* and *motikas* denoting a motorcycle, and the shorter word *ratas* is used for a bicycle. Inclusion of informal equivalents can enrich bilingual dictionaries considerably. However, one has to bear in mind that synonymy is a language-specific phenomenon (Gouws 2013: 349), and instances of cross-linguistic asymmetry are not rare. For example, the English words 'decaf', 'OJ' meaning 'orange juice', and 'BLT' 'a bacon, lettuce, and tomato sandwich' have no informal counterparts in Estonian. Moreover, English is a pluricentred language with multiple standards, which needs

to be taken into account as well. The informal words ‘sarnie’ and ‘butty’ both denote a sandwich but should carry the regional label ‘British’.

An English suffix of recent origin is ‘-aholic’ (for its treatment as a suffix see OED). It appeared that while ‘workaholic’ and its Estonian equivalent *töönarkomaan* were listed in dictionaries, only one dictionary had another entry with this suffix, **shopaholic** *s fam* paadunud poeskäija (Silvet 2002). However, better Estonian equivalents are available for this word, e.g. *ostleja, poodleja, ostusõltlane.*

Among compounds it is worth studying exocentric compounds from the cross-linguistic perspective because their meaning is usually figurative. For example, Estonian has *hädavares* ‘lit misery crow’*, külmavares* ‘ lit frost crow’, *sitavares* ‘lit shit crow’, and *valge vares* ‘white crow’. While *hädavares* can be rendered as ‘lame duck’, *sitavares* ‘shithead’, and *valge vares* ‘the odd man out’, there is no English equivalent for *külmavares* ‘ a person who does not feel comfortably in cold weather’.

Noun–verb conversion in English deserves attention because sometimes the verb is used figuratively as ‘to rabbit on about sth’ in British English.This meaning is missing in English–Estonian dictionaries although suitable equivalents can be found by means of corpus research.

The coverage of informal usage in bilingual dictionaries can be improved considerably.

Keywords: informal usage, bilingual dictionaries, English, Estonian

References

Aule, A. (2001). The Contemporary Estonian-English Dictionary. Vol. 1. A–J. Tallinn: Estonian Language Foundation.

Aule A. (2003). The Contemporary Estonian-English Dictionary. Vol. 2. K. Tallinn: Estonian Language Foundation.

CALD 4 = Cambridge Advanced Learner’s Dictionary on CD-ROM. (2013). 4th edition. CUP.

EED = Estonian–English Dictionary. (2005). First edition. Tallinn: TEA Publishers.

EKSS = Eesti keele seletav sõnaraamat (Explanatory Dictionary of Estonian). Available at http://portaal.eki.ee/dict/ Accessed on 7 December 2017.

Gouws, R.H. (2013). Contextual and Co-textual Guidance Regarding Synonyms in General Bilingual Dictionaries. International Journal of Lexicography. 26.3; 346–361.

Mattiello, E. (2008). An introduction to English slang. A description of its morphology, semantics and sociology. Monza: Polimetrica.

OED = Oxford English Dictionary. Available at www.oed.com Accessed on 7 December 2017.

Saagpakk, P. F. (2000). Estonian–English Dictionary.Tallinn: Koolibri.

Silvet, J. (2002). English–Estonian Dictionary. 4th enlarged and revised edition. Tallinn: TEA Publishers.

Widawski, M. (2015). African American slang: a linguistic description. CUP.

Abstracts of papers

On the Detection of Neologism Candidates as a Basis for Language Observation and Lexicographic Endeavors: the STyrLogism Project

Andrea Abel, Egon W. Stemle
Institute for Applied Linguistics, Eurac Research, Italy
{andrea.abel, egon.stemle}@eurac.edu

The goal of the project STyrLogisms is to semi-automatically extract candidate neologisms (new lexemes) for the German standard variety used in South Tyrol. We use a list of manually vetted URLs from news, magazines and blog websites of South Tyrol, and regularly crawl their data, clean and process it. We compare this new data to reference corpora, additional regional word lists and all the formerly crawled data sets. Our reference corpora are DECOW14, with around 60 million word forms, and the South Tyrolean Web Corpus, with around 2.4 million word forms; the additional word lists consist of named entities, terminological terms from the region and specific terms of the German standard variety used in South Tyrol (altogether around 53,000 word forms). Here, we will report on the method employed, the first round of candidate extraction with an approach for a classification schema for the selected candidates, and some remarks on the second extraction round.

Keywords: neologism, web corpus, dictionary of variants

Lexicographie et terminologie au XIX^e siècle : *Vocabularu romano-francesu* [Vocabulaire roumain-français], de Ion Costinescu (1870)

Maria Aldea
Université Babeș-Bolyai de Cluj-Napoca (Roumanie)
aldea_maria@yahoo.com

Dans cette étude, nous nous proposons d'analyser la manière dont un lexicographe roumain entend définir certains termes au sein d'un corpus choisi : il s'agit du *Vocabularu romano-francesu* [Vocabulaire roumain-français] élaboré par Ion Costinescu et paru en 1870 à Bucarest. Publié suite à une initiative privée quelques années après la fondation de l'Académie roumaine, ce dictionnaire trouve ses modèles déclarés dans la lexicographie française, surtout dans le Dictionnaire de l'Académie française et le Dictionnaire de Napoléon Landais. L'angle d'approche que nous privilégierons nous permettra, d'une part, de saisir la manière dont s'est déroulée l'une des étapes « pré-terminologiques » de la nouvelle discipline qui fait son apparition vers le milieu du XX^e siècle, à savoir la science des termes et, d'autre part, de mesurer les termes y enregistrés à l'aune des principales tendances du développement culturel et scientifique de la société roumaine pendant la deuxième moitié du XIX^e siècle.

Keywords: dictionnaire, terminologie, langue roumaine, langue française, modernisation, emprunt linguistique, Ion Costinescu

Nathanaël Duez lexicographe : l'art de (re)travailler les sources

Antonella Amatuzzi
Università degli Studi di Torino
antonella.amatuzzi@unito.it

La production lexicographique de Nathanaël Duez (1609-1660), maître de langues actif à Leyde, aux Pays Bas, comprend une édition de la *Janua linguarum reserata* de Comenius, comportant les versions française, italienne, allemande et latine, la *Nova nomenclatura quatuor linguarum, gallico, germanico, italico et latino idiomate conscripta*, le *Dictionnaire françois-alleman-latin et aleman-françois-latin* et le *Dittionario italiano e francese Dictionaire italien et François.*

L'objectif du présent travail est de l'analyser dans son évolution (elle commence par un répertoire plurilingue pour terminer avec un véritable dictionnaire, riche d'informations et soigneusement construit) pour mettre en évidence la démarche que Duez suit pour la réalisation de ses ouvrages (notamment la sélection et le remaniement de sources préexistantes). L'étude de l'intertextualité montre qu'il intervient de plus en plus sensiblement pour créer des outils pédagogiques qui répondent aux besoins de ses élèves, clairs et facilement consultables. L'apport de Duez à l'histoire de la lexicographie devrait être réévalué.

Keywords: lexicographie historique, français, italien, allemande, latin, Duez, Pays Bas

The Dictionary of the Learned Level of Modern Greek

Anna Anastassiadis-Symeonidis, Asimakis Fliatouras, Georgia Nikolaou
Aristotle University of Thessaloniki, Democritus University of Thrace, Aristotle University of Thessaloniki
ansym@lit.auth.gr, afliatou@helit.duth.gr, ngeorgia@smg.auth.gr

The aim of this paper is to discuss the theoretical background and methodological tools for the elaboration of a specialized dictionary, the Dictionary of the Learned Elements of Modern Greek (DILLEMOG). The learned level of Modern Greek (MG), which originates from the natural diachronic inheritance and from the prototyping of Ancient Greek, includes segments, structures and processes which pertain to all levels of linguistic analysis. DILLEMOG will constitute an innovative lexicographical database which will provide the user with all the necessary information on the [+ learned] linguistic items of MG, such as definitions, collocations, degree of learnedness, lexical and morphological classification, functionality and usage.

Keywords: learned level, DILLEMOG, lexicographical project

Thesaurus of Modern Slovene: By the Community for the Community

Špela Arhar Holdt[12], Jaka Čibej[123], Kaja Dobrovoljc[13], Polona Gantar[2], Vojko Gorjanc[2], Bojan Klemenc[1], Iztok Kosem[23], Simon Krek[23], Cyprian Laskowski[2], Marko Robnik-Šikonja[1]

Affiliations: [1]Faculty of Computer and Information Science, University of Ljubljana,
[2]Faculty of Arts, University of Ljubljana,
[3]"Jožef Stefan" Institute
spela.arharholdt@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, kaja.dobrovoljc@ijs.si, apolonija.gantar@guest.arnes.si, vojko.gorjanc@ff.uni-lj.si, bojan.klemenc@ fri.uni-lj.si, iztok.kosem@ff.uni-lj.si, simon.krek@ijs.si, cyprianadam.laskowski@ff.uni-lj.si, marko.robnik@fri.uni-lj.si

By presenting the Thesaurus of Modern Slovene, the largest open-access collection of Slovene synonyms, this paper describes the concept of a responsive dictionary, a dictionary that allows its data to continuously respond to the changes in language and the feedback from the language community. We begin by briefly summarizing the method of its construction and its technical aspects. A great deal of deliberation and work has been put into interface design, with the aim to make the Thesaurus as user-friendly as possible for all digital media. This is followed by a more detailed description of the types of user input (e.g. synonym suggestions, synonym votes) and feedback (interface improvement suggestions) collected as part of development, as well as the methodology for their implementation. We also touch upon a series of dissemination activities aimed specifically at community building and user involvement. In conclusion, we describe our plans for the future, such as updates to be implemented in version 1.1 of the Thesaurus.

Keywords: responsive dictionary, digital lexicography, community, crowdsourcing, thesaurus, Slovene

An Overview of FieldWorks and Related Programs for Collaborative Lexicography and Publishing Online or as a Mobile App

David Baines
SIL International
david_baines@sil.org

The FieldWorks ecosystem provides open-source tools for linguists whether working alone or in distributed teams.

FieldWorks is a comprehensive tool for managing linguistic data. It has an extensive selection of fields for each lexical entry and areas for storing grammatical data and interlinear texts. The bulk editing tools can save hours of work by operating on many entries at once. FieldWorks can be used to create mono- or multi-lingual dictionaries and has excellent support for complex scripts. Comprehensive help and resources are available within the tool, which is designed for trained linguists.

Language Forge is an online dictionary creation tool that allows collaborators to browse, comment or contribute to a lexicon. The project manager can control the roles for each team member. Language Forge shares the FieldWorks data allowing users of either tool to modify a shared lexicon. Language Forge can be used with minimal training as it exposes only a small subset of the FieldWorks data.

Webonary is an online platform for publishing dictionaries and their reversal index. The linguist can update the data on Webonary from within FieldWorks as often as desired. Dictionary App Builder facilitates the creation of Android and iOS apps from the FieldWorks data.

Keywords: FieldWorks, Language Forge, collaboration, multilingual, complex scripts, Online publishing, Webonary, mobile publishing, Dictionary App Builder

Dictionary of Verbal Contexts for the Romanian Language

Ana-Maria Barbu
Institute of Linguistics, Romanian Academy, Bucharest
anamaria.barbu@g.unibuc.ro

This paper presents a dictionary of verbal contexts for Romanian, which comprises 600 verbs and over 2,000 meanings with one or more valency patterns. It is manually built but based on corpus information, and is developed both for teaching Romanian to foreigners, by its printed version, and for computational linguistics, by its XML format and consistent principles and conventions of the design. The dictionary is rich in information, including lexical, grammatical and semantic features of the complements, morphosyntactic variants occupying an argument position, dependencies between complements induced by control, raising and predication phenomena and verbal alternations, as variants of valency patterns with the same meaning. The paper offers details about all this information, the building procedure and some problems that needed to be solved during our work. This enterprise is far from being finished, because further work has to be done to improve the actual encoding and add new types of information, such as semantic roles or diathesis uses, for growing the number of entries and for getting different kinds of generalizations.

Keywords: verb pattern, verbal context, valency, argument, complement, Romanian verbs

In Praise of Simplicity: *Lexicographic Lightweight Markup Language*

Vladimír Benko
Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics
vladimir.benko@juls.savba.sk

Our paper presents a simple markup language – *Lexicographic Lightweight Markup Language* (*LLML*) that has been used for almost the last three decades in the framework of two dozen lexicographic projects carried out by our Institute, as well as in several projects carried out in co-operation with commercial dictionary publishers. While initially trying to solve the problem of insufficient computing power of early *MS-DOS*-based personal computers in early 1990's only, *LLML* is even today the central component of lexicographic workstations our lexicographers work with. Central components of the *LLML* syntax are introduced and exemplified by a sample entry from the *Dictionary of the Contemporary Slovak Language* (*SSSJ*). The final part of the paper describes in short some components of the *LLML*-aware toolbox, i.e., programs that are used in our Institute during compilation, validation, proofreading and typesetting of the respective entries. Some of these tools, however, are just a "bonus", and "low-cost" projects could do even without them.

Keywords: lexicographic data representation, lightweight markup language, XML

Interactive Visualization of Dialectal Lexis Perspective of Research Using the Example of Georgian Electronic Dialect Atlas

Marine Beridze, Zakharia Pourtskhvanidze, Lia Bakuradze, David Nadaraia
Javakhishvili State University, Goethe-University Frankfurt/M, Javakhishvili State University, avakhishvili State University
marineberidze@yahoo.com, pourtskhvanidze@em.uni-frankfurt.de, lia.bakuradze@tsu.ge, david.nadaraia@gmail.com

This article presents a report of the results on the current situation in the development of two projects. These are (1) "The Large Georgian Dialect Lexicographic Database and the Georgian Electronic Dialect Atlas" and (2) "A Georgian Language Island in a Trans-Ethnic Area (GLITEA)"[4]. In the first project, the lexicographical component of the Georgian dialect corpus will be expanded and visualized cartographically. The second project examines the dialect of the Georgian – Fereydanian – spoken in Iran by the descendants of about 100 thousand Georgians, who were forcibly evacuated from east Georgia to Iran by Shah Abbas I in the period 1614 to 1616. The dialect is a typical case of a language island and offers the possibility for diverse linguistic research into language history, language contacts and language migration.

Keywords: dialectology, Linguistic Geography, dialectological lexicography, canonical visualization of linguistic data

4 Supported by Shota Rustaveli National Science Foundation (SRNSF) [grant numbers 217008/217438].

Semantic-based Retrieval of Complex Nominals in Terminographic Resources

Melania Cabezas-García, Juan Carlos Gil-Berrozpe
University of Granada
melaniacabezas@ugr.es, jcgilberrozpe@ugr.es

In English, specialized concepts frequently take the form of complex nominals (CNs), e.g. *greenhouse gas emissions*. The syntactic-semantic complexity of these multi-word terms (MWTs) highlights the need for a systematic treatment in specialized resources. This paper explores how semantic patterns in CNs can be applied to retrieve information in terminological knowledge bases, specifically in EcoLexicon (http://ecolexicon.ugr.es), the practical application of Frame-based Terminology (Faber 2012). For that purpose, we extracted the 250 most frequent CNs in an English wind power corpus. Structural disambiguation was performed to identify the internal groups linked by semantic relations. *Ad-hoc* semantic categories were then assigned to the elements of CNs with a view to studying the formation of CNs and allowing semantic-based queries in EcoLexicon. Then, the semantic relations between the CN constituents were analyzed by means of knowledge patterns and paraphrases. Our preliminary results showed recurrent semantic patterns in CN formation. This facilitates the inference of semantic relations, which is one of the main difficulties of MWTs. Furthermore, a semantic-based view of the CN module of EcoLexicon is presented, which allows different types of semantic query.

Keywords: complex nominals, semantic patterns, semantic categories, terminological knowledge bases

Investigating the Dictionary Use Strategies of Greek-speaking Pupils

Elina Chadjipapa
Democritus University of Thrace
elinaxp@hotmail.com

The purpose of this large-scale study was to determine the profile of Greek pupils as dictionary users. In particular, the study investigates the dictionary use strategies that Greek pupils adopt, and records those that they prefer in total and by category while using a dictionary. A total of 745 pupils attending the last three years of primary school and the first three years of junior high school participated in a survey that was carried out in 2014. The data was collected by using the S.I.D.U., a self-report questionnaire. The results revealed that Greek pupils cannot be characterized as strategic dictionary users, as the mean scores of all categories of the dictionary use strategies were below 3.4, which is considered to reflect medium usage. Furthermore, the participants stated that they prefer to employ the look-up and selection strategies more than the lemmatization and the awareness strategies. The medium scores of strategic dictionary use indicate that Greek pupils need training in order to become strategic users.

Keywords: dictionary use strategies, selection strategies, awareness strategies, lemmatization strategies, look-up strategies, strategic dictionary users

Lexicography in the Eighteenth-century Gran Chaco: the Old Zamuco Dictionary by Ignace Chomé

Luca Ciucci
Language and Culture Research Centre, James Cook University
luca.ciucci@jcu.edu.au

The *Vocabulario de la lengua zamuca* is the only extant dictionary of Old Zamuco, an extinct Zamucoan language spoken in the 18th century in the abandoned mission of *San Ignacio de Samucos*, located in the northern Chaco lowlands of South America. This document was until now inaccessible to scholars, but has now been thoroughly studied by the present author, and found to contain very rich data, which establish its fundamental importance for linguistic studies on Zamucoan and Chaco languages. The critical edition of the dictionary is currently under publication. The original author of the dictionary, the Jesuit Father Ignace Chomé (1696-1768), reveals his brilliant linguistic intuition and remarkable skills in the collection and representation of linguistic data. One of the theoretical challenges he had to face was the threefold system of nominal suffixation, which is an absolute linguistic rarity. The present paper will show how the author dealt with the main word classes of Old Zamuco, that is verbs, nouns and adjectives. Chomé structured the entries for these lexemes in such a way as to provide plenty of information on inflectional and derivational morphology. The data from the dictionary permit new and interesting insights on the grammar of Old Zamuco.

Keywords: historical lexicography, lexicography of extinct languages, morphology, South American languages, Zamucoan

A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined

Lut Colman, Carole Tiberius
Dutch Language Institute (INT), Leiden
lut.colman@ivdnt.org, carole.tiberius@ivdnt.org

Woordcombinaties (*Word Combinations*) is to be a new online lexicographic resource in which a Dutch collocation and idiom dictionary will be combined with a pattern dictionary. We believe that the combination of these dictionary types will be of great value to language learners and teachers. In this paper we present the three-year pilot in which we design the project and start with the description of the combinatorics of a selection of verbs for advanced learners of Dutch as a second language. We will merge a pattern dictionary of Dutch verbs, following the example of the *Pattern Dictionary of English Verbs* (*PDEV*),[1] with a collocation application, following the example of *Sketch Engine for Language Learning* (*SkeLL*).[2] In a follow-up to this pilot, more verbs and the combinatorics of nouns and adjectives will be dealt with. The long-term purpose of *Woordcombinaties* is a fully-fledged phraseological resource for Dutch.

Keywords: word combinations, collocations, idioms, proverbs, conversational routines, patterns, e-dictionary for learners of Dutch as a second language, Corpus Pattern Analysis (CPA)

1 http://pdev.org.uk

2 https://skell.sketchengine.co.uk/

Neologisms in Online British-English versus American-English Dictionaries

Sharon Creese
Coventry University, The Alan Turing Institute
creeses@uni.coventry.ac.uk

A common source of publicity for modern-day dictionary publishers is the regular (usually quarterly) release of lists of neologisms that have recently been added to their online dictionaries.[1] The publishing of updated versions of these sites every few months means it may no longer take years for new words to be included in a dictionary. However while different dictionaries may utilize neologisms in similar ways in order to improve brand awareness, the way in which these new words are presented and used in the dictionaries themselves can vary widely, including amongst those of differing varieties of English. This paper will describe differences in the approach and treatment of British-English neologisms in online editions of British-English dictionary *OED* (the *Oxford English Dictionary*) and American-English dictionary *Merriam-Webster*. In particular, the way in which each dictionary responds to potential new words will be discussed, as will the comprehensiveness of the resulting new entries and the differences found in the types of information each contains.

Keywords: neologism, lexicography, dictionaries, dictionary components, British-English, American-English, OED, Merriam-Webster

1 See for example the September 2017 updates for the *OED* (https://public.oed.com/the-oed-today/recent-updates-to-the-oed/september-2017-update/new-words-list-september-2017/) and *Merriam Webster* (https://www.merriam-webster.com/words-at-play/new-words-in-the-dictionary-sep-2017).

Everything You Always Wanted to Know about Dictionaries (But Were Afraid to Ask): A Massive Open Online Course

Sharon Creese, Barbara McGillivray, Hilary Nesi, Michael Rundell, Katalin Sule
Coventry University, The Alan Turing Institute/University of Cambridge, Lexical Computing, Macmillan Dictionaries
creeses@uni.coventry.ac.uk, bm517@cam.ac.uk, h.nesi@coventry.ac.uk, michael.rundell@lexmasterclass.com, k.sule@macmillaneducation.com

We have created a Massive Open Online Course (MOOC) about dictionaries and dictionary-making, to be hosted by FutureLearn. This paper discusses the design and development of this course, which is pitched at high school and undergraduate level participants as well as language enthusiasts around the world. The MOOC will answer questions such as: how dictionaries are made and how this process has changed over time; what goes into a dictionary and who decides; and what kinds of language evidence underpin the information which dictionaries provide. Participants will be encouraged to compare the quantity and quality of information in different types of dictionary, and will investigate corpus-based and computational lexicographic methods. It will also consider dictionary users' attitudes and common misconceptions, taking into account the requirements and habits of English language learners as well as fluent speakers. By the end of the course, participants will know about some of the latest trends in lexicographic research, the roles of language technology, corpora and crowdsourcing in the dictionary compilation process, the range of possible dictionary entry components, lexicographical choices and computational methods surrounding the selection and ordering of word meanings, and the content and wording of definitions.

Keywords: MOOC, massive online open course, dictionary skills, lexicography, history of dictionaries, dictionary-making, corpus linguistics, neologisms, dictionary inclusion criteria, dictionary typologies, lexicographic evidence, crowdsourcing, meaning and definition, corpus-based lexicography

Researching Dictionary Needs of Language Users Through Social Media: A Semi-Automatic Approach

Jaka Čibej, Špela Arhar Holdt
Jožef Stefan Institute, Faculty of Computer and Information Science
Centre for Language Resources and Technologies, University of Ljubljana
jaka.cibej@ijs.si, spela.arharholdt@fri.uni-lj.si

With the rise of digital media in the last decades, many language-related discussions have found home on various fora and social media such as Facebook, where users can participate in a shared-interest group to discuss language use, problems and resources. The posts in these groups are formulated by language users as a genuine response to a specific disruption in language use and offer an empirical starting point for studying language problems. We propose an automatic approach to extracting questions from language-related Facebook groups and describe the procedure in consecutive steps. We also address the issues of copyright, privacy and ethical constraints, and propose ways to overcome them. We present the extraction method on a case of two Slovene language-related Facebook groups: *Za vsaj približno pravilno rabo slovenščine* and *Društvo ljubiteljskih pravopisarjev in slovničarjev*. Both groups allow users to discuss language-related problems and find answers to their questions within the community. Our first extraction from these groups yielded approximately 1,900 posts (written by approximately 500 users) and 13,000 comments (posted by more than 900 users), providing ample material that can be analyzed to reveal the users' most frequent language problems.

Keywords: lexicographical user research, language problems, social media, automatic extraction, Facebook, Slovene

Corpus-based Cognitive Lexicography: Insights into the Meaning and Use of the Verb *Stagger*

Thomai Dalpanagioti
Centre for the Greek Language
dalpanagiotith@yahoo.gr

Situated within the framework of "cognitive lexicography", this paper aims to demonstrate how lexical meaning and usage patterns can be represented in a coherent and principled manner by applying cognitive semantic theories to corpus data. The focus of attention is on the lexicographic tasks of establishing lexical units, capturing usage patterns and providing definitions. The proposed corpus-based and cognitively-oriented approach is applied to a lexical item from the semantic field of motion, the verb *stagger*. Monolingual learner's dictionaries (MLDs) are examined as to their *stagger* entry in order to specify in what respects this approach can improve EFL lexicography. The paper is not restricted to a theoretical discussion of lexicographic issues or a critical review of existing entries; rather, a new version of the *stagger* entry is offered.

Keywords: frame semantics, conceptual metaphor and metonymy theory, word sense disambiguation, usage patterns

Polysemy and Sense Extension in Bilingual Lexicography

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

Polysemy often poses problems for the dictionary representation of word meaning, because the discrimination of senses is seldom clear-cut. In the past twenty years, corpus linguists, notably Kilgarriff (1997) and Hanks (2000), have argued that the concepts of "word sense" and "word meaning" are problematic to the extent that they invite a "checklist" view of meaning that is not borne out by corpus evidence, although precisely that "checklist" view is encouraged by dictionary representation in that dictionaries describe meaning as discrete items in lists (Fontanelle 2016). The challenges associated with representing polysemy are particularly acute for bilingual dictionaries, because patterns of polysemy associated with cross-linguistic equivalents display differing degrees of what has been called "overlapping polysemy" (Alsina & DeCesaris 2002; Boas 2009). This paper considers the treatment in bilingual dictionaries of two small sets of words in English and their equivalents in Spanish, French and Italian which display varying degrees of overlapping polysemy. We suggest ways of incorporating sense extension and partial parallelisms into bilingual dictionary entries, specifically by subdividing senses according to the semantic types as found in corpora.

Keywords: bilingual lexicography, overlapping polysemy, sense extension

A Workflow for Supplementing a Latvian-English Dictionary with Data from Parallel Corpora and a Reversed English-Latvian Dictionary

Daiga Deksne, Andrejs Veisbergs
Tilde, University of Latvia
daiga.deksne@tilde.lv, andrejs.veisbergs@lu.lv

The lexicon of contemporary languages is changing rapidly, mostly by acquiring new loans and derivations. The change in lexicon is best reflected in the corpora of contemporary languages. Nowadays many collections of parallel-aligned texts are available electronically. To satisfy user needs for a modern, complete, up-to-date dictionary, we created a workflow for enriching the existing Latvian-English dictionary with data from parallel corpora containing lexis commonly used in contemporary language, as well as data from the reversed English-Latvian dictionary. While revising the existing Latvian-English dictionary, we identified some issues, for example, missing feminine forms of the nouns naming nationalities and occupations, representation of the words with optional parts or spelling variations. The task of dictionary improvement was done semi-automatically by the joint work of a lexicographer, computational linguists and programmers. Such natural language processing tools as a tokenizer, part-of-speech tagger, lemmatizer and spell-checker were used to reduce the manual work. As a result, the number of entries has increased by 32%, and the number of translations by 28%.

Keywords: electronic dictionaries, parallel corpora, NLP tools, XML format

Multilingual Generation of Noun Valency Patterns for Extracting Syntactic-Semantical Knowledge from Corpora (MultiGenera)

María José Domínguez Vázquez, Carlos Valcárcel Riveiro, David Lindemann
Universidade de Santiago de Compostela, Universidade de Vigo, Universität Hildesheim
majo.dominguez@usc.es, carlos.valcarcel@uvigo.es, david.lindemann@uni-hildesheim.de

The aim of the project is to develop a prototype for a generator of argument structure or valency realizations in terms of syntagmatic and paradigmatic combinations of Spanish, German and French nouns. The two main applications of the tool prototype we are aiming to develop, are (1) the generation of noun phrases as argument structure realizations that follow patterns related to semantic features, for the creation of corpus and web query strings; and (2) the knowledge-based generation of simple and complex noun phrases that are acceptable in a coherent sentence context. An essential step in developing these applications is the systematic description of the valency-related syntagmatic and paradigmatic properties of argument combinations. To this end, we have devised a methodology based on bidirectional mutual enrichment (bottom-down and bottom-up). With the aim of generating argument surface realizations, we will mainly use lexical knowledge represented in wordnets for Spanish, German and French for the semantic annotation of lexical prototypes and their subsequent paradigmatic expansion.

Keywords: noun valency, argument structure, combinatory patterns, corpus lexicography, ontologies, wordnet, natural language generation

Developing a Russian Database of Regular Semantic Relations Based on Word Embeddings

Ekaterina Enikeeva, Andrey Popov
Saint Petersburg State University
protoev@yandex.ru, hedgeonline@gmail.com

Recent computational semantic models yield high-quality results with regard to semantic relations extraction tasks, and thus may be applied as a baseline for semantic lexicon construction. Moreover, the stochastic information about lexical compatibility is useful for reducing ambiguity and detecting anomalies during syntactic parsing. We prove that this approach is reasonable and describe a Russian semantic lexical database, acquired in an unsupervised manner and employed as a semantic component of a syntactic parser and a fact extraction system.

Keywords: distributional semantics, word vector representations, semantic lexicon, Meaning ↔ Text model

Towards a Representation of Citations in Linked Data Lexical Resources

Anas Fahad Khan, Federico Boschetti
Istituto di Linguistica Computazionale „A. Zampolli“, CNR, Pisa
{fahad.khan, federico.boschetti}@ilc.cnr.it

In this article we look at the modelling of citations in lexical resources in linked data. We start by discussing the treatment of citations in linked data and in TEI; we also look at the idea of different conceptual levels as posited by models such as TEI and FRBR. We argue that in representing citations in lexical resources it is important not to confuse different levels of information, and that at least in the case of attestations it is important to model the purpose of a citation, or the claim that is being made by that citation, separately. We develop this point with two separate examples before presenting *lemonBib*, our extension of the lemon model based around the idea of a lexical attestation. We also give a treatment of part of one of the examples described previously in the article.

Keywords: linked data, citations, bibliographic data, lexical resources

Wortschatz und Kollokationen in „Allgemeine Reisebedingungen“. Eine intralinguale und interlinguale Studie zum fachsprachlich-lexikographischen Projekt „Tourlex“.

Carolina Flinz, Rainer Perkuhn
Universität Pisa, Institut für Deutsche Sprache, Mannheim
c.flinz@ec.unipi.it, perkuhn@ids-mannheim.de

Zur Vorbereitung eines zweisprachigen Fachwörterbuchs zur Tourismusfachsprache werden korpuslinguistische Verfahren eingesetzt, um Auffälligkeiten in der jeweiligen Fachsprache im Vergleich zum allgemeinsprachlichen Gebrauch aufzuspüren. Neben den hervorstechenden Elementen des Vokabulars, den Schlüsselwörtern als potentiellen Stichwörtern, geht es vor allem um sprach- und fachsprachspezifische typische Formulierungen und deren Übersetzungsäquivalente. Für die gemeinsame, interlinguale Betrachtung des Sprachenpaars Deutsch-Italienisch wurde ein kleines Fachsprachenkorpus aufgebaut und innerhalb der Sketch Engine-Umgebung unter Zuhilfenahme der darin integrierten Referenzkorpora ausgewertet. Für eine weitere intralinguale Untersuchung der deutschsprachigen Komponente wurde auf das Deutsche Referenzkorpus DeReKo und weitere, intern zu Verfügung stehende Instrumente des Instituts für Deutsche Sprache zurückgegriffen. Neben üblichen Verfahren der quantitativen Ein- oder Mehrwortbewertung wird ein Ansatz ergänzend getestet, der der dünnen Datengrundlage im fachsprachlichen Bereich Rechnung trägt: Diese ergibt sich nicht nur aus der Korpusgröße, sondern auch daraus, dass bestimmte feste Floskeln (wie ‚eine Reiserücktrittsversicherung abschließen‘) selten rekurrent, vielmehr eher nur einmal pro Text verwendet werden. Auch wenn dieser Ansatz aufgrund infrastruktureller Artefakte in Einzelfällen an seine Grenzen stößt, die hier selbstkritisch nicht verschwiegen werden sollen, so zeigt sich doch an vielen Stellen auch das große Potential.

Abschließend wird beispielhaft illustriert, wie Evidenzen dieser und der anderen korpuslinguistischen Auswertungen lexikographisch umgesetzt wurden.

Keywords: Korpuslinguistik, Fachlexikographie, intralingual, interlingual

Towards a Glossary of Rum Making and Rum Tasting

Cristiano Furiassi
University of Turin (Italy)
cristiano.furiassi@unito.it

A lexicographic work exclusively dedicated to the making and tasting of rum has not been published to date. With the ambitious aim of filling this editorial gap in mind, this article focuses on the implementation stage of a specialized glossary of rum-related terms in the English language. Preceded by an overview of the historical, geographical and linguistic factors that made rum a renowned global product, the computer-assisted terminology acquisition procedures applied in order to extract rum-related terms from an *ad hoc* corpus are described. By merging computer-assisted term extraction with data collected from experts' knowledge, fieldwork and the existing specialized literature on rum, a list of candidate headwords was drafted. The replicability of the methodology applied makes this pilot study generalizable, thus fostering the compilation of specialized glossaries connected to other fields or disciplines.

Keywords: computer-assisted term extraction, glossary, rum, specialized lexicography

Synonymy in Modern Tatar reflected by the Tatar-Russian Socio-Political Thesaurus

Alfiia Galieva
Tatarstan Academy of Sciences
amgalieva@gmail.com

This paper discusses some aspects of lexical synonymy in the modern Tatar language that have been revealed in the course of compiling the bilingual *Russian-Tatar Socio-Political Thesaurus.* Building the thesaurus is aimed at fixing all Tatar single words and multiword items related to the socio-political sphere with their Russian equivalents. A distinguishing feature of the contemporary Tatar lexicon is a great deal of absolute synonyms which emerged due to a combination of intralinguistic and extralinguistic factors. We disclose social and linguistic causes of the emergence of synonyms, describe the main structural types of synonymous items, and present corpus data on their frequency. Corpus data prove that synonymy in socio-political terminology is rather an artificial and superficial phenomenon. Currently most Tatar socio-political terms are coined by calquing the corresponding Russian terms, and lexical preferences of translators and terminology developers may differ, which leads to a large number of competing items of different origin and structure. On the level of multiword items, lexical variation is complicated by the factor of syntactic variation, which in its turn multiplies the number of synonymous compounds. Parallel denominations are used for a wide range of phenomena, including official names of state structures and social institutions.

Keywords: lexical synonymy, absolute synonyms, socio-political vocabulary, bilingual thesaurus, the Tatar language

Semantic Classification of Tatar Verbs: Selecting Relevant Parameters

Alfiia Galieva, Ayrat Gatiatullin, Zhanna Vavilova
Tatarstan Academy of Sciences, Tatarstan Academy of Sciences, Kazan State Power Engineering University
amgalieva@gmail.com, agat1972@mail.ru, zhannavavilova@mail.ru

This paper describes the methodology and current results of the ongoing classification of the Tatar lexicon in the process of developing databases of semantic classes of verbs. Our previous work included a semantic classification of Tatar verbs according to their basic meaning and thematic class. As a result, there have distinguished 59 basic semantic classes, with a semantic tag, or a set of tags, attributed to each of 3,200 verbs.

If the thematic classification is universal and may be applied to any language, the currently developed classification described in this paper is based on the parametric principle and includes a set of morphological, syntactic, semantic, and derivational characteristics that are relevant for Tatar grammatical and semantic systems. In a sense, the work is aimed at creating a Tatar analogue of B. Levin's verb classes, taking into account language-specific features.

In the database, each semantic class, or subclass, is supposed to be provided with a set of admissible diathesis alternations and syntactic descriptions, depicting the verb valency, thematic roles of the arguments and semantic restrictions on them. By now we have created a detailed classification of speech, behavior, sound emissions, weather, emotions, mental states and actions verbs; when selecting the pertinent parameters and verifying their relevance, verbs of other classes were also considered.

Keywords: Tatar verb, semantics, corpus, semantic classes

Revision and Extension of the OIM Database – The Italianisms in German

Anne-Kathrin Gärtig
Paris Lodron Universität Salzburg
anne-kathrin.gaertig@sbg.ac.at

The paper presents the *Osservatorio degli italianismi nel Mondo* (OIM), an online database and a homonymous research project on Italianisms in various European target languages, and the revision of the existing data and their structure which has been underway since 2017.

The OIM is the digitized version of the *Dizionario di italianismi in francese, inglese, tedesco* (Stammerjohann et al. 2008). The paper focusses on the Italian loanwords in German registered in two opera, and outlines how further loans in this target language are being systematically integrated during the revision process, and how gaps and weaknesses in their lexicographical description are being filled.

Keywords: multilingual lexicography, specialized dictionaries, online lexicography, language contact, history of the Italian/German lexicon

Russian Borrowings in Greek and Their Presence in Two Greek Dictionaries

Zoe Gavriilidou
Democritus University of Thrace, Greece
zoegab@otenet.gr

This paper focuses on Russian loanwords, loanblends and loanshifts that entered the Greek lexicon during various historical periods and how they have been recorded into two major dictionaries of Modern Greek (MG), the *Dictionary of Standard Modern Greek* (DSMG) (1998) and the *User's Dictionary* of the Academy of Athens (UD) (2014). It also aims at the analysis of the semantic fields of all documented Russian borrowings in the history of Greek, following a classification scheme which was originally used for typological comparison (Haspelmath & Tadmor 2009c). In the first part of the paper, we consider the contact situations that led to borrowing from Russian into Greek and the reasons for borrowing, which include, among others (a) response to major political events, such as the October 2017 Revolution, the Soviet era, the 1987 Perestroika; (b) literary translations of Russian masterpieces in Greek and Greek literature, such as the work of Nikos Kazantzakis or Miltiadis Karagatsis; and (c) religious affinity. Then we compare how these borrowings are recorded in DSMG and UD. In the next section, we offer a morphophonological analysis of borrowings. The semantic fields in which the borrowings belong to are also studied. Finally, the paper provides experimental data for supporting Anastassiadis's (1994) claim that lexical fields, in which loanwords abound, reflect a stereotypic image of the country where the donor language is spoken.

Keywords: loanword, borrowing, loanshift, loanblend, internationalism, loan translation, calque, structural borrowing

Frame-based Lexicography: Presenting Multiword Terms in a Technical E-dictionary

Laura Giacomini
Heidelberg University
laura.giacomini@iued.uni-heidelberg.de

In this paper, a frame-based approach to terminological variation is presented along a model for presentation of multiword terms and their variants in a technical e-dictionary. A case study concerning terminology related to semiconductor diodes is the background against which methods and goals of a larger study on the technical language (Habilitation thesis at Hildesheim University) are illustrated and compared with those of existing resources. At the core of the proposed model are three interrelated description layers (ontology – frame – terminology), with the frame layer serving as the semantic interface between ontological classes and terms, as well as a variation typology accounting for orthographical, morphological and syntactic term variants. The microstructural properties of the envisaged e-dictionary, which aims at supporting text production in the native language, are illustrated by means of the example entry *diode in forward bias*. The addressed users, technical writers and professional translators, are able to access all types of data separately from each other, in a modular way. The paper closes with an outlook on how future developments could include application of the model to further technical domains.

Keywords: frame-based terminology, frame-based lexicography, term variation, technical language, LSP dictionary

Bilingual Corpus Lexicography: New English-Russian Dictionary of Idioms

Guzel Gizatova
Department of European Languages and Cultures, Kazan Federal University, Russia
guzelgizatova@hotmail.com

The paper deals with the principles of constructing the first printed and on-line English-Russian dictionary of idioms based on corpus data. The need for a new dictionary of idioms is motivated by the fact that there is presently no corpus-based dictionary of English-Russian idioms built on authentic examples. Existing traditional bilingual dictionaries do not meet modern requirements of the present-day lexicography with respect to vocabulary and illustrative examples, which are often out of date. This is definitely connected with the fact that traditional English-Russian idiomatic dictionaries were constructed in the 'pre-corpus era'.

The purpose of the present research is thus to introduce a methodology for generating a comprehensive idiom list of the dictionary, to consider linguistic issues presenting difficulties in bilingual lexicography related to the concept of equivalence in idioms, and analyze the semantic asymmetry between English and Russian idioms.

Keywords: idioms, corpora, bilingual lexicography

On the Interpretation of Etymologies in Dictionaries

Pius ten Hacken
Leopold-Franzens-Universität Innsbruck
pius.ten-hacken@uibk.ac.at

Etymological information is an expected type of information in historical dictionaries, but it also appears in many general dictionaries, while it is the key information in etymological dictionaries. Etymologies are generally considered to trace the history of words. However, the notion of a *word* in this statement is an abstraction in more than one way. First, the questions of which forms and which meanings should be placed together as a word does not have an obvious answer. Moreover, the question of which words there are in a language at a particular time cannot be answered on a purely empirical basis. In the light of such observations, I show that what is recorded in an etymology can best be interpreted as the history of the motivation speakers had for the combination of a particular form with a particular meaning. This does not subtract from the value of etymological information, but gives a linguistically sound interpretation of what etymologists have tried to achieve.

Keywords: etymology, historical dictionaries, dictionary interpretation

Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages

Mika Hämäläinen, Jack Rueter
Department of Digital Humanities, University of Helsinki
mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi

We present our ongoing development of a synchronized XML-MediaWiki dictionary to solve the problem of XML dictionaries in the context of small Uralic languages. XML is good at representing structured data, but it does not fare well in a situation where multiple users are editing the dictionary simultaneously. Furthermore, XML is overly complicated for non-technical users due to its strict syntax that has to be maintained valid at all times. Our system solves these problems by making a synchronized editing of the same dictionary data possible both in a MediaWiki environment and XML files in an easy fashion. In addition, we describe how the dictionary knowledge in the MediaWiki-based dictionary can be enhanced by an additional Semantic MediaWiki layer for more effective searches in the data. In addition, an API access to the lexical information in the dictionary and morphological tools in the form of an open source Python library is presented.

Keywords: online dictionary, collaborative editing of XML, Semantic MediaWiki dictionary

Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings

Nicolai Hartvig Sørensen, Sanni Nimb
Society for Danish Language and Literature
nhs@dsl.dk, sn@dsl.dk

We describe the use of a tool that assists lexicographers with extending the lexical coverage of an online Danish dictionary. The tool is based on a word embedding model (word2vec) trained on a large Danish corpus, and it presents semantically related lemmas already included in the dictionary and, importantly, their definitions. Furthermore, lemma candidates, i.e. words from the corpus which are *not* included in the dictionary, are presented in the tool, supplemented by information on corpus frequency. The tool thereby facilitates the lemma selection as well as the process of writing consistent definitions across synonyms and near synonyms. We discuss the shortcomings of the tool and the semantic model when it comes to identifying words similar in meaning from different genres and registers. We also look closer into whether it does in fact benefit the dictionary-making process or not by studying a number of previously edited words, including their synonyms, and comparing them with the output data from the tool.

Keywords: lemma selection, word embedding, word2vec, semantic similarity, dictionary-making process

Wordnet Consistency Checking via Crowdsourcing

Aleš Horák, Adam Rambousek
Natural Language Processing Centre, Faculty of Informatics, Masaryk University, Botanicka 68a, Brno, Czech Republic
hales@fi.muni.cz, rambousek@fi.muni.cz

Large ontologies and semantic networks represent complex multilevel structures, which are incredibly resistant to standard proof checking procedures. Automatic consistency checks can discover system errors such as missing intralingual links, but to find a missing word sense is a difficult task. Standard solutions rely on successive consultations of multiple information sources in a multi-level review process. In this paper, we present a new approach of supplementing such multi-level reviews with engaging the dictionary users in WordNet error corrections and enhancement proposals via systematic crowdsourcing. This approach defines an early release phase with the full dataset published to the target audience followed by a continuous workflow consisting of structured adjustment suggestions obtained from the public users and of the complete editing process by expert reviewers. The review team members are handling prestructured review tasks organized in aggregated forms with correction proposals, the revision management and the appropriate editing of proposed changes. Both the users and reviewers have access to the complete revision history, which allows them to handle repeated proposals responsibly.

Keywords: WordNet, semantic network, ontology, consistency checking

Building a Lexico-Semantic Resource Collaboratively

Mercedes Huertas-Migueláñez, Natascia Leonardi, Fausto Giunchiglia[2]
University of Trento, University of Macerata
mdlm.huertas@unitn.it, natascia.leonardi@unimc.it, fausto@disi.unitn.it

Multilingual lexico-semantic resources are used in different semantic services, such as meaning extraction or data integration and linking, which are essential for the development of real-world applications. However, their use is hampered by the lack of maintenance and quality control mechanisms over their content. The Universal Knowledge Core (UKC) is a multilingual lexico-semantic resource designed as a multi-layered ontology that has a language-independent semantic layer, the concept core, and a language-specific lexico-semantic layer, the natural language core. In this paper, we focus on expert-based, collaborative workflow for building and maintaining our resource through lexicalization and evaluation of language elements via a dedicated User Interface (UI). We have run a three-month study to analyze the feasibility of the proposed solution. We interviewed participants to obtain a comprehensive vision with respect to different aspects related to the way they interacted with the UI and how the content presented through it was perceived. We concluded that this collaborative experience fostered not only the implementation of a resource, but also an improvement of its functionalities, and, above all, it represented an example of effective knowledge sharing which opened up the way to a network of collaborative intelligence.

Keywords: multilingual resource, collaboration, knowledge sharing, user study

2 The present study is the result of a close collaboration among the three authors. However, Mercedes Huertas-Migueláñez wrote Sections 1, 4 and 5. Natascia Leonardi wrote Section 3. Fausto Giunchiglia wrote Section 2. The Abstract and conclusions were a collaborative effort of the three authors.

The CPLP Corpus: A Pluricentric Corpus for the *Common Portuguese Spelling Dictionary* (*VOC*)

***Maarten Janssen*[1], *Tanara Zingano Kuhn*[1], *José Pedro Ferreira*[1], *Margarita Correia*[1,2]**

[1]CELGA-ILTEC, Universidade de Coimbra, [2]FLUL, Universidade de Lisboa

{maartenjanssen,tanarazingano,jpf,margaritacorreia}@uc.pt

The Pluricentric Corpus of the Portuguese Language (CPLP Corpus) aims to provide comparable corpora for the national varieties of the countries where Portuguese is an official language, making it possible to undertake corpus-based comparisons among the varieties of these countries. It is intended as a publicly available corpus for comparative linguistics and language resource development, but furthermore constitutes one of the pillars of the *Vocabulário Ortográfico Comum da Língua Portuguesa* (*VOC*), the official spelling dictionary for Portuguese. The headword list in *VOC* is partly derived from lexicographic tradition, which is to date based almost exclusively on the European and Brazilian varieties, and partly made up of words retrieved from the CPLP corpus, many of them included for the first time in official language resources for Portuguese. This double inclusion route aims at presenting an integral (i.e., non-contrastive) and increasingly balanced perspective on all the varieties. This paper describes the general design of the corpus, the challenges faced in its development, as well as the way it was used in the compilation of *VOC*.

Keywords: corpus, pluricentric languages, Portuguese, spelling dictionaries

Málið.is: A Web Portal for Information on the Icelandic Language

Halldóra Jónsdóttir, Ari Páll Kristinsson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
halldo@hi.is, aripk@hi.is, steinst@hi.is

Málið.is is a web portal on the Icelandic language and language use. It currently includes seven different resources, providing reliable information on the language. Six of the seven resources made available on the portal are living projects, constantly being updated. *Málið.is* provides easy and fast access to these, accessible on and designed for desktop and mobile devices. One of these resources contains a number of specialized terminologies. Thus it is a special characteristic of *málið.is* among language portals for other languages that it provides access to abundant domain specific vocabulary. Work is under way to add more resources to the portal, providing more diverse information. The aim of the project is to strengthen the Icelandic language in the digital era by making it easy to access information on the language, helping people becoming more proficient and confident language users.

Keywords: e-Lexicography, Icelandic, dictionary portal, orthography, phraseology, etymology, terminology

The Treatment of Politeness Elements in French-Korean Bilingual Dictionaries

Hae-Yun Jung, Jun Choi
Kyungpook National University
haeyun.jung.22@gmail.com, c-juni@hanmail.net

Expressions of politeness in French and Korean are lexically as well as conceptually different. For example, there is no equivalent word to the French *s'il te/vous plaît* ('please') in Korean. This particular expression of politeness can be translated in many ways in Korean, which are not necessarily lexical and can also correspond to syntactic elements. At the same time, polite terms of address in French such as *Monsieur/Madame* ('Sir/Madam') have a variety of equivalents which depend on the relationship between the interlocutors and denote the various levels of politeness in Korean. Because of these discrepancies, the lexicographical description of politeness related lexical items presents many issues and shortcomings with regard to their practical use from the perspective of French learners of Korean. The objective of this paper is thus to analyze how lexical items of politeness are treated in French-Korean lexicography and what issues the relevant entries present. The equivalents and examples analyzed are extracted from the Naver dictionary portal, which gathers information from existing published bilingual dictionaries. As a result of the analysis, this study proposes a solution to the issues which the entries of such expressions present by suggesting a model for the lexicographical treatment of politeness elements.

Keywords: bilingual dictionary, French-Korean lexicography, politeness, pragmatic information, microstructural model

A Lexicon of Albanian for Natural Language Processing

Besim Kabashi
Friedrich-Alexander-Universität Erlangen-Nürnberg
Ludwig-Maximilians-Universität München
besim.kabashi@{fau,lmu}.de

For a lot of applications in the field of natural language processing a lexicon is needed. For the Albanian language a lexicon that can be used for these purposes is presented below. The lexicon contains around 75,000 entries, including proper names such as the names of inhabitants, geographical names, etc. Each entry includes grammatical information such as part of speech and other specific information, e. g. inflection classes for nouns, adjectives and verbs. The lexicon is a part of a morphological tool and generator, but can also be used as an independent resource for other tasks and applications or can be adapted for them. Both information from some traditional dictionaries, e. g. spelling dictionaries, and a balanced linguistic corpus using corpus-driven methods and tools are used as sources for the creation and extension of the presented lexicon. The lexicon is still a work in progress, but aims to cover basic information for the most frequent tasks of natural language processing.

Keywords: Albanian, NLP lexicography, lexicon updating, corpus linguistics

Associative Experiments as a Tool to Construct Dictionary Entries

Ksenia S. Kardanova-Biryukova
Moscow City University, Moscow, Russia
kardanova81@yandex.ru

Associative experiments have been used in psychology and psycholinguistics for over a hundred years and have proven to be an efficient tool in identifying those components of a cognitive structure which are relevant for a contemporary language user and arranging them into a hierarchy. We argue that the findings obtained through such an experimental approach can serve as the basis for compiling a dictionary entry that reflects the language in use now and does not lag behind by failing to depict changes in the semantic structure of a word or modifications in its use. To demonstrate how it works we have considered a rather complex notion of *empire* and its representation in the linguistic consciousness of Russian and American English native speakers. Relying on the findings of the associative experiment held with 148 Americans and 434 Russians we suggest ways of drafting dictionary entries that would reflect those semantic components which are relevant for language users given the current political and economic environment.

Keywords: associative experiment, modelling cognitive structures, stimulus and reaction

The Case of Reciprocity in Czech

Václava Kettnerová, Markéta Lopatková
Charles University
kettnerova@ufal.mff.cuni.cz, lopatkova@ufal.mff.cuni.cz

Reciprocity has been the focus in much theoretical research in recent years. It has been primarily studied as a grammatical property, which is not of high relevance for the description of the lexical stock of a language. At the same time, however, it has been widely accepted that languages substantially differ with respect to the inventory of words allowing for reciprocity, and that the applicability of reciprocity is rarely derivable from the semantic and/or syntactic properties of these words. The integration of the information on reciprocity into lexicons would thus be highly beneficial for both human users (esp. for foreign speakers) and for natural language processing tasks. In this paper, we demonstrate how the reciprocity of Czech verbs can be represented in a lexicon in a comprehensive and systematic way. Czech represents a language where reciprocity is a highly productive phenomenon. We show which semantic and syntactic properties are relevant for the description of reciprocal verbs, and based on this a user (be it human or computer) can acquire their reciprocal constructions.

Keywords: reciprocity, reciprocal verbs, valency structure of verbs, lexicon, syntax, Czech

Building a Gold Standard for a Russian Collocations Database

Maria Khokhlova
St. Petersburg State University
m.khokhlova@spbu.ru

In the last decade, linguists have become increasingly interested in corpus material, which allows for a fresh approach to the phenomena that have already been extensively described in academic works. The dual nature of the co-occurrence phenomenon itself lies, on one hand, in its linguistic component and, on the other, in the probabilistic (combinatorial) characteristics. The former has been described in numerous papers and explicitly defined in dictionaries, while the latter can be identified by a statistical approach. The present paper focuses on the process of building a gold standard that will include data from Russian dictionaries and corpora. The standard is being prepared for a Russian Collocations Database that already includes information on words' collocability and was extracted from text corpora by statistical measures and linguistic filters. The gold standard will be also used for the evaluation of the extracted collocations and for marking them as "true" collocations with references to the dictionaries.

Keywords: database, collocations, corpora, dictionaries, Russian language

Rethinking the Role of Digital Author's Dictionaries in Humanities Research

Margit Kiss, Tamás Mészáros
Institute for Literary Studies, Hungarian Academy of Sciences, Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics
kiss.margit@btk.mta.hu, meszaros@mit.bme.hu

Although it is true that computers make one's work significantly easier during some of the work phases of creating an author's dictionary, our goal is to completely revise conventional computer-based dictionary work processes as well as to extend the possible range of applications for author's dictionaries. In our study we examine the role of the digital author's dictionaries in humanities research. We developed the extended digital author's dictionary based on the oeuvre of Kelemen Mikes (1690-1761). We present the advantages of the extended digital author's dictionary and demonstrate its benefits in literary and linguistic research. The main advantages of our system include easily accessible up-to-date encyclopedic information and the improved efficiency of historical text analysis methods. The benefits of the extended functions of the digital author's dictionary help scholars answer more specialized and complex research questions, and reconsider expectations towards the new generation of author's dictionaries.

Keywords: author's digital dictionary, historical text analysis, Kelemen Mikes

New German Words: Detection and Description

Annette Klosa, Harald Lüngen
Institut für Deutsche Sprache, Mannheim, Germany
klosa@ids-mannheim.de, luengen@ids-mannheim.de

In this paper, we discuss an efficient method of (semi-automatic) neologism detection for German and its application for the production of a dictionary of neologisms, focusing on the lexicographic process. By monitoring the language via editorial (print and online) media evaluation and interpreting the findings on the basis of lexicographic competence, many, but not all neologisms can be identified which qualify for inclusion in the *Neologismenwörterbuch* (2006-today) at the Institute for the German Language in Mannheim (IDS). In addition, an automated corpus linguistic method offers neologism candidates based on a systematic analysis of large amounts of text to lexicographers. We explain the principles of the corpus linguistic compilation of a list of candidates and show how lexicographers work with the results, combining them with their own findings in order to continuously enlarge this specialized online dictionary of new words in German.

Keywords: detection of neologisms, description of neologisms, corpus linguistics, lexicography

Collocations Dictionary of Modern Slovene

Iztok Kosem[12], Simon Krek[2], Polona Gantar[1], Špela Arhar Holdt[1], Jaka Čibej[123], Cyprian Laskowski[1]

[1]Faculty of Arts, University of Ljubljana

[2]Jožef Stefan Institute

[3]Faculty of Computer and Information Science, University of Ljubljana

iztok.kosem@ff.uni-lj.si, simon.krek@guest.arnes.si, apolonija.gantar@guest.arnes.si, Spela.ArharHoldt@ff.uni-lj.si, jaka.cibej@ff.uni-lj.si, CyprianAdam.Laskowski@ff.uni-lj.si

The paper presents the compilation of the Collocations Dictionary of Modern Slovene, a new resource targeting the language production needs of Slovene speakers. An important aspect of the compilation of the dictionary is the immediate publication of all the entries, from automatic, post-processed, finalized by lexicographers and so on, and indicating to the users their status, i.e. the stage in the compilation process. Furthermore, we discuss the introduction of crowdsourcing into the lexicographic workflow. The paper also focuses the development and presentation of the interface, which introduces new approaches to collocation presentation. The aim was to develop a collocation-driven interface that would allow different types of users a great deal of flexibility and customizability in exploring collocational information about words. In this way, the interface represents a hybrid between a more corpus-based presentation of collocations (e.g. in tools such as Word Sketch) and a traditional sense-driven presentation of collocations as found in existing collocations dictionaries.

Keywords: collocations, dictionary, database, interface

European Lexicographic Infrastructure (ELEXIS)

Simon Krek[1], Iztok Kosem[1], John P. McCrae[2], Roberto Navigli[5], Bolette S. Pedersen[6], Carole Tiberius[4], Tanja Wissik[3]

[1]Jožef Stefan Institute, Slovenia

[2]Insight Centre for Data Analytics, National University of Ireland Galway,

[3]Austrian Academy of Sciences, Austria,

[4]Dutch Language Institute, Netherlands

[5]Sapienza University of Rome, Italy

[6]University of Copenhagen, Denmark

simon.krek@ijs.si, john@mccr.ae, iztok.kosem@ijs.si, tanja.wissik@oeaw.ac.at, carole.tiberius@ivdnt.org, navigli@di.uniroma1.it, bspedersen@hum.ku.dk

In the paper we describe a new EU infrastructure project dedicated to lexicography. The project is part of the Horizon 2020 program, with a duration of four years (2018-2022). The result of the project will be an infrastructure which will (1) enable efficient access to high quality lexicographic data, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. One of the main issues addressed by the project is the fact that current lexicographic resources have different levels of (incompatible) structuring, and are not equally suitable for application in in Natural Language Processing and other fields. The project will therefore develop strategies, tools and standards for extracting, structuring and linking lexicographic resources to enable their inclusion in Linked Open Data and the Semantic Web, as well as their use in the context of digital humanities.

Keywords: lexicography, research infrastructure, natural language processing, computational linguistics, semantic web, artificial intelligence, linked open data, digital humanities

Computer-aided Analysis of Idiom Modifications in German

Elena Krotova
Institute of Linguistics, Russian Academy of Sciences
elena_krotova@inbox.ru

This paper deals with corpus approaches to the study of modifications of idiomatic expressions in German. It concentrates on one group of phraseological units, idioms. In spite of a high degree of stability, idioms still undergo different modifications. To get reliable results about idiom modifications, a large number of modified target structures is crucial. Therefore, a Python-program has been created that obtains information about the usage of idioms and about their possible modifications from a corpus. It also summarizes the data in the form of graphs. The report will look further into the program's opportunities to acquire information about idiom usage, how idiom modifications correspond to the syntactic behavior of their paraphrases or free phrases containing the same verb as the idiom under discussion and in what ways such data can facilitate the work of a phraseologist.

Keywords: corpus linguistics, phraseography, modifications of idiomatic expressions

The Sounds of a Dictionary: Description of Onomatopoeic Words in the Academic Dictionary of Contemporary Czech

Magdalena Kroupová, Barbora Štěpánková, Veronika Vodrážková
Czech Language Institute of the Czech Academy of Sciences
kroupova@ujc.cas.cz, stepankova@ujc.cas.cz, vodrazkova@ujc.cas.cz

This paper is focused on the description of onomatopoeic interjections and onomatopoeic verbs in a monolingual dictionary. Compared to its predecessors, the emerging monolingual dictionary of Czech (ASSČ) provides more space for the description of all dictionary entries, including onomatopoeic interjections, and their treatment is more detailed. The study concentrates on primary meanings of words related to natural sounds which are imitated, e.g., *bú* (*moo*), *bučet* (*to moo*), as well as on secondary meanings vaguely related to the sound. The paper compares possible ways of treatment of definitions, especially determination of the defining vocabulary for the specific part of the lexicon, the reflection of imitating, and the structure of individual components of the explanation of the meaning (e.g., typical producers, supporting adjectives and adverbs etc.), in addition, proposals of the treatment in the ASSČ are presented. To provide a complex description of onomatopoeic interjections, their syntactic functions and their specific semantic features are discussed.

Keywords: Czech, definition, interjection, (lexical) meaning, monolingual dictionary, onomatopoeia, verb

Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How

Gabriele Langer, Anke Müller, Sabrina Wähl, Julian Bleicken
University of Hamburg
{gabriele.langer, anke.mueller, sabrina.waehl, julian.bleicken}@uni-hamburg.de

Within the DGS-Korpus Project, a corpus-based dictionary of German Sign Language (DGS) is compiled. Dictionary entries describe signs, their meanings and uses as they are reflected in the corpus. The dictionary entries include authentic examples taken directly from the original corpus recordings. Without a functional writing system for sign languages (SL), corpus building as well as SL usage examples in dictionaries have to resort to videos as representations of SL use. Examples can either be taken from the original corpus material (authentic examples) without the option of even mild editorial changes, or they can be re-recorded with a different signing model. While the latter allows for editing as well as constructed examples, it also entails very drastic changes to the appearance of the original authentic examples they are based on.

In the article our reasons for inclusion of authentic examples are discussed and criteria for example selection listed. To compensate for the challenges that authentic examples removed from their original contexts entail, translations and context information are added to the entries. The practical steps for example preparation, namely selecting the segment, providing context information and adjusting translations are described.

Keywords: authentic examples, sign language dictionary, spoken language (as opposed to written), corpus-based lexicography

The EcoLexicon English Corpus as an Open Corpus in Sketch Engine

***Pilar León-Araúz*[1], *Antonio San Martín*[2], *Arianne Reimerink*[1]**
[1]*Department of Translation and Interpreting, University of Granada*
[2]*Department of Modern Languages and Translation, University of Quebec in Trois-Rivières*
pleon@ugr.es, antonio.san.martin.pizarro@uqtr.ca, arianne@ugr.es

The EcoLexicon English Corpus (EEC) is a 23.1-million-word corpus of contemporary environmental texts. It was compiled by the LexiCon research group for the development of EcoLexicon (Faber, León-Araúz & Reimerink 2016; San Martín et al. 2017), a terminological knowledge base on the environment. It is available as an open corpus in the well-known corpus query system Sketch Engine (Kilgarriff et al. 2014), which means that any user, even without a subscription, can freely access and query the corpus. In this paper, the EEC is introduced by describing how it was built and compiled and how it can be queried and exploited, based both on the functionalities provided by Sketch Engine and on the parameters in which the texts in the EEC are classified.

Keywords: specialized open corpus, terminology, corpus exploitation

When Learners Produce Specialized L2 Texts: Specialized Lexicography between Communication and Knowledge

Patrick Leroyer & Henrik Køhler Simonsen
Aarhus University & Copenhagen Business School
pl@cc.au.dk & hks.msc@cbs.dk

This article discusses the theoretical distinction between communicative- and cognitive-oriented dictionary use situations and explores whether or not this sharp distinction is still valid at a time when users do not use dictionaries but instead online language resources, particularly in learning environments. The paper seeks to answer this research question based on empirical data from a user study conducted at Copenhagen Business School in 2017. We carried out a controlled experiment involving ten test persons and the user study produced ten screen recordings, ten specialized texts, ten self-assessments and ten teacher-assessed rubrics. On the basis of our empirical data we found that the sharp distinction between communicative and cognitive-oriented dictionary use situations does not seem to make much sense anymore when users, to an increasing extent, do not use dictionaries but instead online language resources. We found that specialized language and specialized knowledge are completely intertwined, mutually interdependent and form a dialectic relation, which in fact can be identified by analyzing the test person's information search and retrieval processes. We also found that new, modern language resources make it possible to make searches in text directly and to take full advantage of the dialectic relation between specialized language and specialized knowledge.

Keywords: specialized lexicography, cognitive functions, communicative functions, situation distinction, learning situations, functional interdependence

ColloCaid: A Real-time Tool to Help Academic Writers with English Collocations

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

[1]Faculty of English, Adam Mickiewicz University in Poznań

[2]School of Literature and Languages, University of Surrey

[3]School of Computer Science, Bangor University

rlew@amu.edu.pl, a.frankenberg-garcia@surrey.ac.uk, g.rees@surrey.ac.uk, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk

Writing is a cognitively challenging activity that can benefit from lexicographic support. Academic writing in English presents a particular challenge, given the extent of use of English for this purpose. The ColloCaid tool, currently under development, responds to this challenge. It is intended to assist academic English writers by providing collocation suggestions, as well as alerting writers to unconventional collocations choices as they write. The underlying collocational data are based on a carefully curated set of about 500 collocational bases (nouns, verbs, and adjectives) characteristic of academic English, and their collocates with illustrative examples. These data have been derived from state-of-the-art corpora of academic English and academic vocabulary lists. The manual curation by expert lexicographers and reliance on specifically Academic English textual resources are what distinguishes ColloCaid from existing collocational resources. A further characteristic of ColloCaid is its strong emphasis on usability. The tool draws on dictionary-user research, findings in information visualization, as well as usability testing specific to ColloCaid in order to find an optimal amount of collocation prompts, and the best way to present them to the user.

Keywords: writing assistant, collocation, academic writing, English for academic purposes

LexBib: A Corpus and Bibliography of Metalexicographical Publications

David Lindemann, Fritz Kliche, Ulrich Heid
Universität Hildesheim
{david.lindemann, fritz.kliche, heid}@uni-hildesheim.de

This paper presents preliminary considerations regarding objectives and workflow of LexBib, a project which is currently being developed at the University of Hildesheim. We briefly describe the state of the art in electronic bibliographies in general, and bibliographies of lexicography and dictionary research in particular. The LexBib project is intended to provide a collection of full texts and metadata of publications on metalexicography, as an online resource and research infrastructure; at the same time, LexBib has a strong experimental component: computational linguistic methods for automated keyword indexing, topic clustering and citation extraction will be tested and evaluated. The goal is to enrich the bibliography with the results of the text analytics in the form of additional metadata.

Keywords: bibliography, metalexicography, full text collection, e-science corpus, text analytics

"Brexit means Brexit": A Corpus Analysis of Irish-language BREXIT Neologisms in The Corpus of Contemporary Irish

Katie Ní Loingsigh
Fiontar & Scoil na Gaeilge, Dublin City University
katie.niloingsigh@dcu.ie

The primary objective of this paper is to map the introduction and adoption of various Irish-language BREXIT neologisms in The Corpus of Contemporary Irish (CCI)[3]. This study follows a corpus approach and aims to record the development of Irish-language BREXIT neologisms in a subcorpus of media texts compiled from CCI. Firstly, a general overview is given of the Irish language along with a background to the development of various Irish-language BREXIT neologisms. The corpus is utilised to examine these terms and to identify any lexicogrammatical patterns of use. The emerging linguistic patterns are surveyed and discussed in the results section. This is a narrow study and is limited to print media, as there is currently no comprehensive corpus of spoken data available for Irish. While BREXIT is still a relatively new term, it is highly topical and in widespread use at the time of writing. This analysis aims to provide an insight into the development of unique Irish-language neologisms along with providing a base for future comparisons of similar neologisms in other minority languages

Keywords: Irish language, corpus linguistics, lexicography, neologisms

3 https://www.gaois.ie/g3m/ga/

A Call for a Corpus-Based Sign Language Dictionary: An Overview of Croatian Sign Language Lexicography in the Early 21st Century

Klara Majetić, Petra Bago
Faculty of Humanities and Social Sciences in Zagreb
klaramajetic13@gmail.com, pbago@ffzg.hr

Many sign languages today are still not standardized nor accessible to a wider audience. Sign languages with high quality dictionaries are scarce. In this paper we give a brief description of sign languages and how they differ from spoken ones, in order to better understand the issues lexicographers might face when compiling dictionaries for these languages. We focus on Croatian Sign Language (HZJ), giving an overview of the current situation in HZJ lexicography following some criteria we find relevant for both online and printed sign language dictionaries. The criteria have been classified into twenty-five categories and applied to create a model for an online HZJ dictionary, briefly presented in this paper. By presenting an unsatisfactory status of HZJ lexicography, we are issuing an urgent call for the compilation of a HZJ corpus as a basis for a high quality dictionary that could benefit both the potential hearing and deaf users.

Keywords: Croatian Sign Language (HZJ), sign language lexicography, dictionary evaluation, e-dictionary model

New Platform for Georgian Online Terminological Dictionaries and Multilingual Dictionary Management System

Tinatin Margalitadze
Ivane Javakhishvili Tbilisi State University
tinatin.margalitadze@tsu.ge

The *English-Russian-Georgian Technical Online Dictionary* is the first 'digitally born' online dictionary of Georgian, created in a Multilingual Dictionary Management System (MDMS), specially developed for this project. The *Technical Dictionary* is the third specialized dictionary created by the same lexicographic team since 2009 (after the *English-Georgian Biology Online Dictionary* and *English-Georgian Military Online Dictionary*). Work on specialized vocabulary of different domains has revealed that terminology has evolved, particularly during the last 10 – 15 years. The traditional, standard requirements for monosemy and mononymy are not always observed in actual terminological work. There are numerous instances of terminological synonymy, many terms are polysemous, frequently developed as a result of metaphorical change of the primary meaning; there are many multiword terms consisting of two, three or even more words, giving rise to numerous terminological abbreviations; synonymous terms may belong to different stylistic registers, which requires the introduction of some stylistic labels in terminological entries. Rapid development of science and technology in the 21st century caused the appearance of an abundance of new concepts and consequently new terms. The resulting influx of new terminology into the Georgian language dictates the necessity to provide definitions of such terms alongside their Georgian equivalents. Introduction of collocations and examples of usage of terms is another issue that comes to the foreground of lexicographic description of terms.

These observations about modern terminology, which is discussed in the first part of the paper, became the basis for the development of a new platform for English-Georgian online bilingual terminological dictionaries and MDMS, as outline din this paper.

Keywords: structural and semantic characteristics of modern terminology, Multilingual Dictionary Management System, a platform for English-Georgian online bilingual terminological dictionaries

A Sample French-Serbian Dictionary Entry based on the *ParCoLab* Parallel Corpus

***Saša Marjanović*[1], *Dejan Stosic*[2], *Aleksandra Miletic*[2]**
[1] *Faculty of Philology, University of Belgrade, Serbia*
[2] *CLLE, Université de Toulouse, CNRS, Toulouse, France*
sasa.marjanovic@fil.bg.ac.rs, dejan.stosic@univ-tlse2.fr, aleksandra.miletic@univ-tlse2.fr

It has already been shown in the state- of the art in lexicography that the bilingual dictionary making process can be improved by relying on parallel corpora. The aim of this paper is to present such an application of the *ParCoLab* parallel corpus, a searchable trilingual 11 million token electronic database of aligned texts in French, Serbian and English, developed at the University of Toulouse (France) in cooperation with the University of Belgrade (Serbia). In this paper, we first point out the shortcomings of the leading general French-Serbian dictionaries, which were made using traditional lexicographic methods. We pay special attention to the treatment of the equivalents offered. Taking the case of the French adjective *sale* 'dirty' as an example, we show that the *ParCoLab* parallel corpus makes it possible to: 1) have quick and easy access to meanings missing from the existing dictionaries and to corresponding equivalents; 2) find new equivalents that are not included in any of the existing dictionaries, and which are in some cases the most common translation solutions; 3) order equivalents by their relative corpus frequency; and 4) disambiguate different usages through adequate contextual examples. The solutions we offer are shaped into a sample dictionary entry.

Keywords: parallel corpus, lexicography, dictionary, bilingual, French, Serbian

Building a Portuguese Oenological Dictionary: from Corpus to Terminology via Co-occurrence Networks

William Martinez, Sílvia Barbosa
SYLED Université Sorbonne Nouvelle Paris 3, NOVA FCSH - CLUNL
wmartinez68@gmail.com, silviabarbosa@fcsh.unl.pt

This paper focuses on the elaboration of a dictionary of terms in the Portuguese language which describe the wine-tasting experience. We present a corpus-based analysis aimed at designing an electronic dictionary: on the basis of a compilation of approximately 21,000 wine descriptions downloaded from a dozen Portuguese websites, we estimated both by frequency analysis and lexicographical study which terms were recurrent, relevant and representative of the "hard to put into words" occupation that is oenology. From the results thus obtained, a list was made of words that describe the sensory analysis in its three main aspects: visual, olfactive and gustatory. An exhaustive co-occurrence analysis then identified those terms which contribute most to structuring the text by way of their tendency to attract other words against statistical odds. When displayed in a co-occurrence network, these anchors emerge from the mesh as the foundational lexicon for wine tasting, and can be evaluated as prime candidates for a distributional thesaurus.

Keywords: collocations, co-occurrences, word network, corpus linguistics, oenology, terminology

Computerized Dynamic Assessment of Dictionary Use Ability

Osamu Matsumoto
Waseda University
matsumoto.kbx@gmail.com

This paper demonstrates a new web-based dictionary training resource, named the Computerized Dynamic Assessment of Dictionary use Ability (C-DADA) intended for Japanese learners of English. C-DADA is a computerized format of dynamic assessment (DA), which originated in the works of Vygotsky, and is developed for lexicographic purposes. The fundamental aspect of DA is the integration of assessment and instruction. In addition, DA distinguishes two levels of ability: actual and potential. The former is the level at which the individual can perform by himself, while the latter is the level at which he can solve a problem with others' assistance. Likewise, C-DADA is designed to assess the learners' actual and potential levels of receptive dictionary use ability through providing feedback according to individual learners' responsivity, and so further promote their dictionary use ability. The paper first introduces the theoretical and methodological framework of DA, and then discusses the design and mechanics of C-DADA. Lastly, it gives a general overview of an on-going lexicographic project to examine the effectiveness of C-DADA.

Keywords: dictionary use, dictionary training, dynamic assessment, computerized dynamic assessment

Exploring the Frequency and the Type of Users' Digital Skills Using S.I.E.D.U.

Stavroula Mavrommatidou
PhD Student, Democritus University of Thrace
stavrmav@hotmail.com

S.I.E.D.U. (Strategy Inventory for Electronic Dictionary Use) is a valid and reliable electronic instrument designed for assessing users' skills in electronic dictionary searches. It can be used for research purposes mainly for the detection of users' profiles in order to design appropriate intervention programs in classrooms. In the present paper, it has been used for collecting empirical data on users' dictionary skills, which is an important but poorly researched topic in language learning contexts. Seven hundred people (students from high schools and universities as well as teachers) participated in the investigation and completed the online questionnaire S.I.E.D.U., reflecting on their own digital dictionary use. It was found that not all users are familiar enough with the strategies required when using digital dictionaries and some of them lack the right skills to fully benefit from this useful source of information. In addition, there are differences in the skills applied by users depending on their level of education but not between university students in different study fields.

Keywords: digital lexicography, user skills, dictionary use

Shareable Subentries in Lexonomy as a Solution to the Problem of Multiword Item Placement

Michal Boleslav Měchura
Masaryk University, Brno, Czech Republic
michmech@mail.muni.cz

This paper introduces a new way of dealing with phraseology in dictionaries. A classical question in lexicography is whether multiword items such as *third time lucky* should be listed under *third*, *time* or *lucky*. The ideal answer is 'under all of them' but, until now, the only way to do that in conventional tree-structured dictionaries has been to keep multiple copies (of what conceptually is one and the same item) in several places throughout the dictionary. We present a way to achieve the same goal without copying. The multiword item becomes a semi-independent subentry which exists in only one copy but appears simultaneously in several places in the dictionary. The structure of the dictionary remains a tree but the lexicographer is empowered to occasionally 'break out' of the tree in order to avoid duplication. This paper explains the reasoning behind the concept of shareable subentries, and shows how this new functionality has been implemented in the dictionary writing system Lexonomy.

Keywords: subentries, phraseology, Lexonomy

Creating a List of Headwords for a Lexical Resource of Spoken German

Meike Meliss, Christine Möhrs, Dolores Batinić, Rainer Perkuhn
Institut für Deutsche Sprache, Mannheim
meliss@ids-mannheim.de, moehrs@ids-mannheim.de, batinic@ids-mannheim.de, perkuhn@ids-mannheim.de

Except for some recent advances in spoken language lexicography (cf. Verdonik & Sepesy Maučec 2017, Hansen & Hansen 2012, Siepmann 2015), traditional lexicographic work is mainly oriented towards the written language. In this paper, we describe a method we used to identify relevant headword candidates for a lexicographic resource for spoken language that is currently being developed at the Institute for the German Language (IDS, Mannheim). We describe the challenges of the headword selection for a dictionary of spoken language, and having made considerations regarding our headword concept, we present the corpus-based procedures that we used in order to facilitate the headword selection. After presenting the results regarding the selection of one-word lemmas, we discuss the opportunities and limitations of our approach.

Keywords: list of headwords, spoken German, corpus-based methods

Using Diachronic Corpora of Scientific Journal Articles for Complementing English Corpus-based Dictionaries and Lexicographical Resources for Specialized Languages

Katrin Menzel
Dept. of Language Science and Technology, Saarland University
k.menzel@mx.uni-saarland.de

As technology and science permeate nearly all areas of life in modern times, there is a certain trend for standard dictionaries to bolster their technical and scientific vocabulary and to identify more components, for instance more combining forms, in technical terms and terminological phrases. In this paper it is argued that recently built diachronic corpora of scientific journal articles with robust linguistic and metadata-based features are important resources for complementing English corpus-based dictionaries and lexicographical resources for specialized languages. The Royal Society Corpus (RSC, ca. 9,800 digitized texts, 32 million tokens) in combination with the Scientific Text Corpus (SciTex, ca. 5,000 documents, 39 million tokens), as two recently created corpus resources, offer the possibility to provide a fuller picture of the development of specialized vocabulary and of the number of meanings that general and technical terms have accumulated during their history. They facilitate the systematic identification of lexemes with specific linguistic characteristics or from selected disciplines and fields, and allow us to gain a better understanding of the development of academic writing in English scientific periodicals across several centuries, from their beginnings to the present day.

Keywords: English diachronic corpora, Graeco-Latin combining forms, scientific journal articles, scientific vocabulary, corpus-based dictionaries, lexicographical resources for specialized languages

The DHmine Dictionary Work-flow: Creating a Knowledge-based Author's Dictionary

Tamás Mészáros[1], Margit Kiss[2]
[1]Budapest University of Technology and Economics,
[2]Institute for Literary Studies, Hungarian Academy of Sciences
[1]meszaros@mit.bme.hu, [2]kiss.margit@btk.mta.hu

Digitalized author's dictionaries could play an important role in humanities research. Not only could they provide better ways to study an individual author's vocabulary, but they could also act as a knowledge source for other computer-based methods. We present the process of making an author's dictionary of headwords, writing variations, word forms and corpus citations extended with part-of-speech, linguistic, literary and semantic information. We also describe how this extended dictionary incorporates knowledge from linked open data sources and from critical annotations and builds an RDF knowledge base attached to the dictionary. The result is a vast knowledge source about an author's oeuvre that can be studied and used to enhance corpus analysis. We demonstrate our method on processing a large text corpora of 1.5 million words from the 18th century and on creating the digital author's dictionary of Kelemen Mikes.

Keywords: author's dictionaries, knowledge-based systems, corpus analysis, linked open data

fLexiCoGraph: Creating and Managing Curated Graph-Based Lexicographical Data

Peter Meyer, Mirjam Eppinger
Institute for the German Language, Mannheim, Germany
meyer@ids-mannheim.de, eppinger@ids-mannheim.de

We present the conceptual foundations and basic features of fLexiCoGraph, a generic software package for creating and presenting curated human-oriented lexicographical resources that are roughly modeled according to Měchura's (2016) idea of graph-augmented trees. The system is currently under development and will be made accessible as open source software. As a sample use case we discuss an existing online database of loanwords borrowed from German into other languages which is based on a growing number of language-specific loanword dictionaries (*Lehnwortportal Deutsch*). The paper outlines the conceptual foundations of fLexiCoGraph's hybrid graph/XML data model. To establish a database, XML-based resources may be imported or even input manually. An additional graph database layer is then constructed from these XML source documents in a freely configurable, but automated way; subsequently, the resulting graph can be manipulated and enlarged through a visual user interface in such a way that keeps the relationship to the source document information explicit at all times. We sketch the tooling support for different kinds of graph-level editing processes, including mechanisms for dealing with updated XML source documents and coping with duplicate or inconsistent information, and briefly discuss the browser interface for end users.

Keywords: graph-based dictionaries, editorial process, data modeling, linked data, historical lexicography

ELeFyS: A Greek Illustrated Science Dictionary for School

Maria Mitsiaki, Ioannis Lefkos
Assistant Professor, Democritus University of Thrace, Laboratory Teaching Staff, University of Macedonia
mmitsiaki@helit.duth.gr, lefkos@uom.edu.gr

This paper reports on the design and compilation of ELeFyS (Εικονογραφημένο Λεξικό Φυσικής για το Σχολείο, ΕΛεΦυΣ), a Greek specialized school dictionary of science. Since its conception ELeFyS has been intended as a reference tool for the parallel development of scientific and linguistic literacy in a school context. To fulfil such an objective, generic entries include scientific terms that fall within the school subject of physics and are likely to be encountered in the upper grades of primary and lower grades of secondary school; however, the dictionary coverage is not restricted to terminology, but is also expanded to the terms/headwords' respective general sense(s) and use(s). Moreover, encyclopedic and cultural material is given as further stimuli for critical thinking. Under this scope, ELeFyS works both as a lexicographic product and a multi-functional teaching resource. In sum, it constitutes a novel endeavor of combining pedagogy and specialization in order to meet the complex linguistic and cognitive/scientific needs of school children in the late primary and the early secondary school grades. Such a complex aim of determining both communication- and knowledge-oriented lexicographic functions is being realized thanks to the enduring collaboration of a linguist and a science expert, well-rooted in long teaching experience. In what follows, we focus on the policy decisions made at the outset of the lexicographic project and the entry-building process.

Keywords: Greek science dictionary, macro- and microstructure, content-based learning/instruction

The Virtual Research Environment of VerbaAlpina and its Lexicographic Function

Christina Mutter, Aleksander Wiatr
Ludwig-Maximilians-Universität München
christina.mutter@lmu.de, aleksander.wiatr@lmu.de

This paper describes the long-term research project VerbaAlpina of Munich University, which has been funded by the German Research Foundation (DFG) (http://gepris.dfg.de/gepris/projekt/253900505) since October 2014. The project investigates the Alpine lexis of three conceptual domains in the Alpine region where dialects and languages belonging to three large language families (Germanic, Romance and Slavonic) are spoken. This paper emphasizes one of the project's main functional areas, its lexicographic function, which serves to gather, process, access and visualize lexical data. To this end, data from traditional linguistic atlases and dictionaries as well as recent data gathered via the project's crowdsourcing tool first have to undergo a process of systematic data processing to fit the unified structure of the relational database (MySQL). This process can be subdivided into three major steps: transcription, tokenization and typification. Apart from the multi-directionality of the project which collects, documents and disseminates structured linguistic and ethnographic data, VerbaAlpina also provides an innovative online publishing platform that will prove sustainable and can be easily cited.

Keywords: digitalization, crowdsourcing, interlingual geolinguistics

From Standalone Thesaurus to Integrated Related Words in *The Danish Dictionary*

Sanni Nimb, Nicolai H. Sørensen, Thomas Troelsgård
Society for Danish Language and Literature
sn@dsl.dk, nhs@dsl.dk, tt@dsl.dk

This paper presents a method of integrating Danish standalone thesaurus data automatically into a monolingual dictionary of modern Danish (*Den Danske Ordbog*, 'The Danish Dictionary') and discusses the results, including some of the problematic cases. The method draws on the detailed semantic grouping with two types of keywords in a well-structured XML-manuscript of a recently published thesaurus of Danish (*Den Danske Begrebsordbog*, 'The Danish Concept Dictionary') and on the fact that the two resources are linked on sense level, allowing for the automatic identification of semantically related thesaurus extracts for any given sense in the dictionary. The paper also presents a study of similar integrations of thesaurus data in four online English dictionaries, namely the *Oxford English Dictionary*, the *MacMillan Dictionary*, the *Merriam-Webster English Dictionary* and the *Oxford Dictionaries:English*, which we carried out in order to compare the structure of the underlying English thesaurus data as well as the resulting dictionary presentations with the Danish case.

Keywords: thesaurus, dictionary, linked data, synonyms

Terms Embraced by the General Public: How to Cope with Determinologization in the Dictionary?

Jana Nová
Czech Language Institute of the Czech Academy of Sciences
novaj@volny.cz

The determinologization can be described as a process when specialized words (scientific terms) move into general vocabulary, followed by changes in their meaning. Czech and Slovak linguists have described two types or subsequent stages of determinologization. Firstly, the term is widely used outside scientific communication and it remains connected to the background scientific concept, but its meaning becomes less accurate in the use of laypeople. Secondly, the connection to the original concept is lost and a new figurative meaning of the word develops. The lexicographical approach to the language material reveals there is another, transitional type of determinologization where the word is used by laypeople in a blurred meaning, e.g. wider or narrower than if the term was used by domain specialists. The treatment of selected determinologized words in four dictionaries (Czech, Slovak and two English ones) is compared in this paper and various ways to mark the determinologized usage are presented, including a separate paragraph in the dictionary entry, an example sentence with an additional explanation or an additional note to the entry.

Keywords: Czech, determinologization, dictionary definition, general dictionary, lexicography, Slovak, terminology

Process Nouns in Dictionaries: A Comparison of Slovak and Dutch

***Renáta Panocová*[*], *Pius ten Hacken*[¶]**
Pavol Jozef Šafárik University in Košice[*], *Leopold-Franzens-Universität Innsbruck*[¶]
renata.panocova@upjs.sk, pius.ten-hacken@uibk.ac.at

Process nouns are deverbal nouns that designate the process indicated by the corresponding verb. Often, they have additional readings, such as a result reading. We present the productive mechanisms for the formation of process nouns in Slovak and Dutch. The two rules in Slovak and three rules in Dutch differ in the degree of regularity and the tendency to have additional senses.

In their process readings, process nouns are prototypical examples of what can be covered in a run-on entry, i.e. a sub-entry under the headword it is related to without a separate definition. The felicity of run-on entries depends on the regularity and predictability of the word. Some of the rules for process nouns are so regular that there is no reason to specify their output. Other rules are better suited to a representation of their output as a run-on entry, but only if the meaning is constrained to the process reading.

Keywords: nominalization, process nouns, process-result alternation, run-on entries, Slovak, Dutch

Definitions of Words in Everyday Communication: Associative Meaning from the Pragmatic Point of View

Svitlana Pereplotchykova
Taras Shevchenko National University of Kyiv
s.pereplotchykova@knu.ua

In their interactions with others, people try to be cooperative if they want to be understood by their interlocutor(s). Spontaneous speech presupposes a good command of the language in use, but the overall meaning of certain words, their recognizability (visibility), is closely connected with the associations that the real-world objects they denote or the words themselves have acquired and are true for particular social groups. In the word-based games of the original TV show *Hollywood Game Night* and its Greek version *Celebrity Game Night*, the players provide their associations with the objects of reality in question or the words denoting them in order to convey the information effectively. The results of this research comprise an overview of how speakers of different languages process and understand words, and how different associations and verbalizations can be not only among languages but also among those who speak the same language but belong to different social groups. The analysis of these associations allows us to explain which elements of the information they provide can help us to understand each other correctly, which are the factors that make interlocutors choose this or that association, and why sometimes they fail to elicit the information they ask for, to explain what they mean.

Key words: salience, association, associative meaning, definition of words, communication, English, Modern Greek, semantic

Exploratory and Text Searching Support in the *Dictionary of the Spanish Language*

Jordi Porta-Zamorano
Centro de Estudios de la Real Academia Española
porta@rae.es

Online dictionaries try to include search capabilities to meet most users' needs. Although users are not always aware of how to effectively use dictionaries, sometimes it is the interface that does not facilitate a friendly access to the dictionary information. This work aims at lowering the barrier in supporting onomasiological and semasiological advanced searches to the *Diccionario de la lengua española* (DLE) by combining text searches and faceted navigation into a user-friendly dictionary interface, allowing even non-experts to move through the dictionary in a natural and flexible manner. However, since the DLE is an electronic version of a printed dictionary, it contains related and unrelated abbreviations condensing different information that have to be properly converted into the set the facets and values provided by the search system.

Keywords: Dictionary interfaces, dictionary searching, online dictionaries

Analyzing User Behavior with Matomo[4] in the Online Information System *Grammis*

Saskia Ripp, Stefan Falke
Institut für Deutsche Sprache (IDS), Mannheim
ripp@ids-mannheim.de, falke@ids-mannheim.de

The grammatical information system *grammis* combines descriptive texts on German grammar with dictionaries of specific word classes and grammatical terminology. In this paper, we describe the first attempts at analyzing user behavior for an online grammar of the German language and the implementation of an analysis and data extraction tool based on Matomo, a web analytics tool. We focus on the analysis of the keywords the users search for, either within *grammis* or via an external search platform like Google, and the analysis of the interaction between the text components within *grammis* and the integrated dictionaries. The overall results show that about 50% of the searches are for grammatical terms, and that the users shift from texts to dictionaries, mainly by using the integrated links to the dictionary of terminology within the texts. Based on these findings, we aim to improve *grammis* by extending its integrated dictionaries.

Keywords: user behavior, online information systems, automated tracking, Matomo, online grammars, online dictionaries, keyword analysis

4 The web tracking system Piwik was renamed Matomo on January 9th, 2018.

A Universal Classification of Lexical Categories and Grammatical Distinctions for Lexicographic and Processing Purposes

Roser Saurí, Ashleigh Alderslade, Richard Shapiro
Oxford University Press
roser.sauri@oup.com, ashleigh.alderslade@oup.com, richard.shapiro@oup.com

We introduce COMO (Compositional Morphosyntactic Ontology), a classification of part-of-speech categories and their associated grammatical features, which aims to be valid across languages of very different typology. The work has been carried out within the context of the Oxford Global Languages programme, which has the goal of developing language knowledge for 100 languages, particularly those under-represented in the digital space. The requirements around this project are: to be able to describe languages of different typeS while respecting their grammatical tradition, and to be able to serve two main use cases that define our typical work, namely, the labelling of linguistic information in lexicographic products, and the provision of support for language processing tools and corpus annotation processes. These requirements determined the conception and design of COMO, created as a reference model within a broader data architecture in order to address issues of syntactic and semantic interoperability. Our proposal builds on top of previous initiatives in the field aiming at the same goals, but incorporates different features in order to accommodate for the requirements in the project.

Keywords: part-of-speech tagging, morphosyntactic information, language modelling, interoperability, multilinguality

Dictionaries of Linguistics and Communication Science / *Wörterbücher zur Sprach- und Kommunikationswissenschaft* (WSK)

Stefan J. Schierholz
Friedrich-Alexander-Universität Erlangen-Nürnberg
Stefan.Schierholz@fau.de

The "WSK" is a German online dictionary series which will be published in print in 2019/2020. Each of the dictionaries is a terminological special field dictionary on the subject of "Linguistics and Communication Science". The dictionaries will be partially translated to English, are intended for experts and semi-experts, and they will serve for comprehension of technical terms, information, and translation. Currently, 25 dictionaries are envisioned for the series, which will contain more than 50,000 lemmas. Eleven thousand of these dictionary articles have been published online since 2013. About 800 people are working on the project worldwide.

In this paper, the structure of the whole project, the organization and management, and the work flow of article writing will be presented. By taking volume 1, "Grammar", as a basis, the text compound structure, the function of the systematic introduction, and the article structure will be introduced.

Keywords: special field lexicography, terminology, WSK, grammar, dictionary project

Verifying the General Academic Status of Academic Verbs: An Analysis of co-occurrence and Recurrence in Business, Linguistics and Medical Research Articles

Natassia Schutz
Centre for English Corpus Linguistics
Université catholique de Louvain, Belgium
natassia.schutz@uclouvain.be

General academic vocabulary lists have been the subject of much debate. Because they focus on single words, they have been criticized for not considering "the importance of contextual environments which reflect different disciplinary practices" (Hyland & Tse 2007: 251). This study aims to provide insight into the reliability of such vocabulary lists by analyzing cross-disciplinary phraseological variation. To do so, I analyze the collocations and lexical bundles used with c. 30 academic verbs found in a 3-million-word corpus containing research articles in business, linguistics and medicine. The results seem to suggest that there are sufficient commonalities, both in terms of use and meaning, to justify the creation and use of general academic vocabulary lists. In addition to their discipline-specific uses, many of the verbs under focus also have general academic uses that relate to the core business of research, irrespective of the academic discipline (e.g. *provide* + *information*/*insight* and *as can be seen in*). The results of this study also demonstrate the benefit derived from adopting a bottom-up approach to phraseology, as it identified a considerable number of verb-based patterns that are not found in existing corpus-driven academic phraseology lists.

Keywords: academic vocabulary, verbs, collocations, lexical bundles

Lexicography in the French Caribbean: An Assessment of Future Opportunities

Jason F. Siegel
The University of the West Indies, Cave Hill Campus
jason.siegel@cavehill.uwi.edu

While lexicography in the Hispanophone Caribbean has flourished, and to a lesser extent in the territories of the Caribbean whose official language is English, dictionaries of the French-official Caribbean (except Haiti) have been quite limited. But for the rest of the French-official Caribbean, there remains much work to do. In this paper, I assess the state of lexicography in the French-official Caribbean, as well as the possibilities for future work. There are six principal areas of lexicographic documentation to be developed. The first, most urgent task is the documentation of the endangered St Barth French. The next priority is multilingual lexicography for the Caribbean region. The third priority is multilingual lexicography of French Guiana, home to endangered Amerindian, Creole and immigrant languages. Fourth, there is a largely pristine area of lexicographic work for the English varieties of the French Caribbean. The fifth area of work to be developed is monolingual lexicography of French-based Creoles. Lastly, there is exploratory work to be done on the signed language varieties of the French-official Caribbean. The paper concludes with a discussion of the role that the Richard and Jeannette Allsopp Centre for Caribbean Lexicography can play in the development of these areas.

Keywords: minority languages, bilingual lexicography, French Caribbean

Looking for a Needle in a Haystack: Semi-automatic Creation of a Latvian Multi-word Dictionary from Small Monolingual Corpora

Inguna Skadiņa
University of Latvia, Institute of Mathematics and Computer Science
inguna.skadina@lumii.lv

Multiword expressions (MWEs) are an indispensable part of almost any dictionary. However, the identification of missing MWEs that have recently appeared in a language is not a simple task. In this paper we describe automated methods for MWE identification in a rather small Latvian text corpora. We propose starting with the application of statistical measures to identify a wide range of MWEs and then applying linguistically motivated filters to clean the list of initially extracted MWE candidates. We show that for morphologically rich languages, such as Latvian, in cases with a small amount of language data better results can be achieved with lemmatized data. We also demonstrate that in the case of a small general domain (balanced) corpus, automatic methods can be used to find good MWE candidates – terminological units, named entities and some lexicalized phrases. However, finding idiomatic expressions in small, general domain corpora is looking for a needle in a haystack: only a larger or more expressive corpus can help in the identification process.

Keywords: multi-word expressions, low resourced languages, collocations, named entities, terminology

Comparing Orthographies in Space and Time through Lexicographic Resources

Christian-Emil Smith Ore, Oddrun Grønvik
University of Oslo and University of Bergen
c.e.s.ore@iln.uio.no, Oddrun.Gronvik@uib.no

Many languages require an improved factual basis to facilitate computer-supported analysis of language variation and diachronic change. The material collections for the scholarly dictionaries of Norway serve as a platform for exploring the development and variation of Bokmål, the Norwegian written standard derived from Danish and modified towards the Norwegian vernacular through orthographic reforms that took place from 1901 to 2005. The development of modern Bokmål through usage should be analyzed by comparing corpora from different periods, lemmatized according to the then current orthography. This means building full form registers from time-bound orthographies. This plan is in process through digitizing orthographic dictionaries for Bokmål. The dictionaries are coordinated through the Dictionary Hotel, the electronic repository for retro digitized dictionaries and dialect collections at the Norwegian Language Collections, Bergen. At the lexical item level Bokmål and Nynorsk resources are coordinated through the Meta Dictionary, an electronic registry for the Norwegian lexicon. A common entry requires full identity in one headword form plus part-of-speech (POS). Preliminary results identify a core vocabulary for Bokmål of 6,900 lexical items, unchanged since 1938. More than 75,000 Meta Dictionary entries have a common identical form plus POS for Bokmål and Nynorsk. These numbers will increase when the Bokmål additions to the Meta Dictionary are quality controlled.

Keywords: dictionary, lexical item, full form register, computer assisted language analysis, corpus, lemmatizer, synchronic variation, diachronic change

The *Dictionary of the Serbian Academy*: from the Text to the Lexical Database

Ranka Stanković[1], Rada Stijović[2], Duško Vitas[1], Cvetana Krstev[1], Olga Sabo[2]
[1]University of Belgrade, [2]Institute for Serbian Language, Serbian Academy of Sciences and Arts
ranka.stankovic@rgf.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs, vitas@matf.bg.ac.rs, cvetana@matf.bg.ac.rs, olga011@yahoo.com

In this paper we discuss the project of digitization of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language*. Scanning and character recognition were a particular challenge, since various non-standard character set encoding was used in the course of the almost 60-year long production of the dictionary. The first aim of the project was to formalize the micro-structure of the dictionary articles in order to parse the digitized text of and transform it into structured data stored in relational lexical database. This approach is compatible with several standard structured forms and ontologies (TEI, LMF, Ontolex, LexInfo). A lexical database model was designed in compliance with these structured forms, following mostly the *lemon* model. Mapping of the lexical entry markers to LexInfo and TEI enabled export of the lexical data to the mentioned formats. A software solution for the dictionary text analysis, parsing and lexical database population was developed and tested on the first and the last published volumes of the dictionary (which contain 27,141 articles in total). An evaluation of the results shows that the developed model and software solution can be successfully used for the other volumes as well.

Keywords: computer lexicography, lexical database, language resources, dictionary, Serbian language

Commonly Confused Words in Contrastive and Dynamic Dictionary Entries

Petra Storjohann
Institut für Deutsche Sprache Mannheim
storjohann@ids-mannheim.de

This paper discusses changes in lexicographic traditions with respect to contrastive dictionary entries and dynamic, on-demand e-lexicographic descriptions. The new German online dictionary *Paronyme - Dynamisch im Kontrast* is concerned with easily confused words (paronyms), such as *effektiv*/*effizient* and *sensibel*/*sensitiv*. New approaches to the empirical analysis and lexicographic presentation of words such as these are required, and this dictionary is committed to overcoming the discrepancy between traditional practice and insights from language use. As a corpus-guided reference work, it strives to adequately reflect not only authentic use in situations of actual communication, but also cognitive ideas such as conceptual structure, categorization and knowledge. Looking up easily confused lexical items requires contrastive entries where users can instantly compare meaning, contexts and reference. Adaptable access to lexicographic details and variable search options offer different foci and perspectives on linguistic information, and authentic examples reflect prototypical structures. These are essential in order to meet all the different interests of users. This paper will illustrate the contrastive structure of the new e-dictionary and demonstrate which information can be compared. It also focusses on various dynamic modes of dictionary consultation, which enable users to shift perspectives on paronyms accordingly.

Keywords: paronyms, dynamic lexicography, contrastive entries, generating information on demand

Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX

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

The Institute of the Estonian Language is developing EKILEX, a new dictionary writing system for both semasiological dictionaries and onomasiological termbases. While the long-term vision is to have a single data source that provides consistent information about Estonian, the system also needs to cope with the multitude of existing datasets. In this paper, we present work in progress on modelling the data and importing an initial sample of legacy dictionaries. The data model is based on an m:n relation between words and meanings, which are both unified across dictionaries, even while there still are separate dictionaries in the system. What is dictionary-specific is only the mapping between word and meaning. The importing of dictionaries has revealed various issues with data quality: ambiguities, underspecification, inconsistencies and conflicts. These need to be dealt with, if the long-term vision is to be achieved. We also outline the next steps of human- and machine-readable publishing, corpus connection and quantification (frequency, salience measures, etc.).

Keywords: data modelling, dictionary portal, interoperability, linked data, Estonian

Historical Corpus and Historical Dictionary: Merging Two Ongoing Projects of Old French by Integrating their Editing Systems

Sabine Tittel
Heidelberg Academy of Sciences and Humanities
sabine.tittel@urz.uni-heidelberg.de

To combine corpus data with dictionary data has two advantages: (i) It embeds the vocabulary of the corpus texts within the overall system of the language, and it semantically disambiguates the texts. (ii) The corpus data enrich the dictionary and shed new light on the comprehension of the vocabulary. The retrospective integration of corpus data into a dictionary is a task that has to focus on two aspects, (i) on the integration of the word forms, and (ii) on the semantic integration of the words. This second aspect continues to be an important issue, particularly for historical languages. Automated solutions do not exist. In this paper, we present the retrospective integration – both with a graphical and a semantic focus – of the corpus of Old French legal texts, *Documents linguistiques galloromans* (with approx. 800,000 attestations of Old French lexemes), into the *Dictionnaire étymologique de l'ancien français* (with 83,000 dictionary entries). We have implemented a semi-automated process resulting in a time-saving editorial workflow to accomplish the data integration. Further, we have created a twofold publication concept for the dictionary entries that makes for a straightforward way of enriching the dictionary with the valuable material of the domain of Old French law.

Keywords: historical lexicography, corpus linguistics, Old French, dictionary writing system, scholarly digital text edition, history of law

Multimodal Corpus Lexicography: Compiling a Corpus-based Bilingual Modern Greek – Greek Sign Language Dictionary

Anna Vacalopoulou, Eleni Efthimiou, Kiki Vasilaki
ILSP-Institute for Language and Speech Processing / Athena RC
avacalop@ilsp.gr, eleni_e@ilsp.gr, kvasilaki@ilsp.gr

This paper describes the process of compiling NOEMA+, a bilingual dictionary of approximately 12,000 entries for the pair Greek Sign Language (GSL) - Modern Greek (MG) and of making it available openly online (http://sign.ilsp.gr/signilsp-site/index.php/el/noima/). The dictionary was based on several corpora that have been collected over the years, including information on compounding, GSL synonyms, classifiers, various lemma-related senses, semantic relationships, etc. These different corpora have been joined, normalized and translated into MG to form a parallel corpus of the language pair in question. In turn, this parallel corpus acted as the basis for the compilation of the bilingual dictionary described in this paper.

More specifically, among the issues to be discussed here are lemma identification, which proved far from intuitive for this particular language pair, lemma categorization, dictionary contents and structure, relations between entries as well as the corpus which was used for dictionary compilation. Finally, there will be a description of the different search choices offered, which cater for different user profiles and needs.

Keywords: bilingual lexicography, corpus-based lexicography, multimodal parallel corpus, sign language lexicography

Heritage Dictionaries, Historical Corpora and other Sources: Essential And Negligible Information

Alina Villalva
Faculdade de Letras & Centro de Linguística da Universidade de Lisboa
alinavillalva@campus.ul.pt

Contemporary dictionaries are often the result of accumulating the contents of previous dictionaries, namely heritage or patrimonial dictionaries, which are those that set a landmark in the historical lexicography of a given language. This is certainly the case for European Portuguese dictionaries, which offer word lists containing words in usage as well as rarely used words, words that belong to other language variants and phonetic/orthographic and morphological alternates, being apparently unable to distinguish among relevant, puzzling, or simply useless information.

If contemporary dictionaries were to be redesigned, should lexicographic ancestry be ignored, or should it be dealt with otherwise? The discussion on new lexicographic models is beyond the range of this paper, but any lexicographic innovation must be grounded in solid lexicological analyses that cannot ignore the consideration of heritage dictionaries and historical *corpora*. These sources provide a large quantity of information requiring specialized interpretation and critical trimming, which may nevertheless be insufficient to fully grasp a thorough knowledge of the related words. This paradox, which is the focus of the present work, will be addressed by advocating the need to adopt a lexicographic protocol that may help to select the right amount of information and combine it with identical information from other languages.

Keywords: heritage dictionaries, historical corpora, European Roots

Slovenian Lexicographers at Work[5]

Alenka Vrbinc, Donna M. T. Cr. Farina, Marjeta Vrbinc
University of Ljubljana, New Jersey City University, University of Ljubljana
alenka.vrbinc@ef.uni-lj.si, dfarina@njcu.edu, marjeta.vrbinc@ff.uni-lj.si

This paper reports part 1 of findings from a grant project between Slovenia and the U.S. that set out to understand the context and content of modern Slovenian lexicography. Interviews were conducted with six Slovenian lexicographers and one terminographer working on different projects within several institutions (the Slovenian Academy of Sciences and Arts; the University of Ljubljana; and Trojina, Institute for Applied Slovenian Studies). The grant's aim was to discern the philosophical underpinnings, most noteworthy accomplishments, and main projects of Slovenian dictionary work. The focus was on those aspects of lexicographic work that have the greatest significance for the general educated public rather than areas (such as dialectology, etymology, etc.) that might attract primarily language specialists.

The interview script consisted of thirteen narrative questions, designed to allow interviewees to reflect on their daily practice and their underlying vision of what lexicography or terminography is. This paper focuses on a single interview question that captured the interviewees' views on drudgery in lexicography, and on the social/ethical role of the lexicographer.

Keywords: harmless drudge, drudgery, interview, lexicographer, objectivity

5 An expanded version of this article is forthcoming in *Lexikos* 28 (2018) under the title: "Objectivity, Prescription, Harmlessness, and Drudgery: Reflections of Lexicographers in Slovenia".

Linking Corpus Data to an Excerpt-based Historical Dictionary

Tarrin Wills, Ellert Þór Jóhannsson, Simonetta Battista
University of Copenhagen
tarrin@hum.ku.dk, nkv950@hum.ku.dk, sb@hum.ku

A Dictionary of Old Norse Prose (*ONP*) is a digital dictionary that derives originally from an excerpt-based index of around 750,000 citations. This paper describes recent attempts to create two-way links between the growing body of digital texts encoded using TEI XML and the dictionary's word list, which forms the basis of the published dictionary. The process involves design challenges in bringing together very different digital structures, namely the text in an XML tree structure, and the dictionary in a relational database structure. Because of the very high levels of accuracy demanded by the end-users of the dictionary (particularly researchers in Old Norse studies), the linking process can only be automated for unambiguous cases, with remaining links entered manually. The application and interface that assists this process attempts to minimize the trade-off between automation and accuracy, and adds a range of tools to assist with the human lemmatizing process. We were able to achieve linking of lemmas in 90.4% of instances where the lemma was recorded in the TEI text, with very high levels of accuracy. Where no lemma was recorded, the application allowed an Old Norse scholar to link lemmas to previously unlemmatized words at an average rate of 4-7 seconds per word.

Keywords: online dictionary, corpus material, historical lexicography, Old Norse

Combining Quantitative and Qualitative Methods in a Study on Dictionary Use

Sascha Wolfer[1], Martina Nied Curcio[2], Idalete Maria Silva Dias[3], Carolin Müller-Spitzer[1], María José Domínguez Vázquez[4]

[1]Institut für Deutsche Sprache, [2]Universita degli Studie Roma Tre, [3]Universidade do Minho Braga, [4]Universidade de Santiago de Compostela

wolfer@ids-mannheim.de, martina.nied@uniroma3.it, idalete@ilch.uminho.pt, mueller-spitzer@ids-mannheim.de, majo.dominguez@usc.es

Many studies on dictionary use presuppose that users do indeed consult lexicographic resources. However, little is known about what users actually do when they try to solve language problems on their own. We present an observation study where learners of German were allowed to browse the web freely while correcting erroneous German sentences. In this paper, we are focusing on the multi-methodological approach of the study, especially the interplay between quantitative and qualitative approaches. In one example study, we will show how the analysis of verbal protocols, the correction task and the screen recordings can reveal the effects of intuition, language (learning) awareness, and determination on the accuracy of the corrections. In another example study, we will show how preconceived hypotheses about the problem at hand might hinder participants from arriving at the correct solution.

Keywords: research into dictionary use, observation study, language learners, quantitative and qualitative methods, online lexicographic resources

Methodological issues of the compilation of the Polish Academy of Sciences Great Dictionary of Polish

Piotr Żmigrodzki
The Institute of the Polish Language at the Polish Academy of Sciences
piotr.zmigrodzki@ijp.pan.pl

Polish Academy of Sciences Great Dictionary of Polish (pol. *Wielki słownik języka polskiego PAN*) is being created by a team of linguists, lexicographers and other specialists in The Institute of the Polish Language PAN. This process started in Kraków in 2006. The dictionary is published exclusively in the online format, free of charge. Before August 2018, 70,000entries, which include the most frequent words, idiomatic expressions and proper names of the Polish language, will have been finished. The paper provides a general description of the dictionary and focuses especially on following issues of its compilation: a) workflow and methods of organizing work, b) lexicographical system of the dictionary and some of its innovations, which are useful during the process of entry creation and the process of overseeing the entries by lexicographers, c) a new, introduced in 2015, graphical user interface and its features. The final part is devoted to a short presentation of the plans the team has for the future.

Keywords: Polish language, electronic lexicography, general dictionary of Polish, online dictionary

Abstracts of plenary lectures

Legal Interpretation via Dictionaries and Corpora: Can Judges Pass Lexicography 101?

Edward Finegan
University of Southern California
finegan@usc.edu

Appellate court judges in the U.S., including those on the Supreme Court, increasingly cite definitions from general dictionaries in their written opinions addressing issues that litigants have appealed from lower courts. While judges' cherry picking of dictionaries or of senses that appear to justify a predetermined conclusion is perhaps not surprising, recent research has highlighted the fact that opinions sometimes reflect judicial ignorance about how dictionaries are compiled, how senses are ordered within entries, even the fact that dictionaries differ in their micro-structural organization. In other words, some judicial opinions reflect inadequate knowledge of basic dictionary use. With the availability of a variety of corpora and increasing acquaintance with their use in legal settings, judges have recently begun citing corpus evidence, some of it manifestly ad hoc, in their opinions, and briefs filed before the U.S. Supreme Court have recently cited corpus evidence. These early stages of a move, if not away from dictionaries, certainly toward corpora in addressing questions of ordinary textual meaning in legal contexts should not leave observers sanguine. Despite possible advantages in relying on corpora and even occasional necessity (for example, in historical documents), no lexicographer or corpus linguist should regard this turn to corpora as an unalloyed blessing for judges seeking ordinary meaning of contextualized terms. This presentation identifies ways in which judges typically invoke dictionary definitions and the missteps they sometimes take. It also describes recent court opinions that rely on corpora instead of dictionaries and certain missteps apparent there. I will argue that the turn to corpora in pursuit of ordinary meaning in legal settings, if left solely to the courts and legal experts, will prove detrimental. The turn to corpora in legal settings calls for active involvement by lexicographers and corpus linguists to help ensure that such a turn improves upon the use of dictionaries as aids in striving for valid textual interpretation.

Has Lexicography Reaped the Full Benefit of the (Learner) Corpus Revolution?

Sylviane Granger
Centre for English Corpus Linguistics, University of Louvain
sylviane.granger@uclouvain.be

In 1992 Rundell and Stock wrote an extended three-part article on the "corpus revolution", in which they describe the rise of corpora and their impact on lexicography. Rundell (2008) revisited the topic and focused on the arrival of the web, which triggered a second stage in the corpus revolution. The purpose of my presentation is to look at some aspects of the current lexicographic scene and assess whether the corpus revolution has really fulfilled its promise. I will show that, while this is largely true of monolingual learners' dictionaries (especially in the case of English), the situation is much less favourable when it comes to general bilingual dictionaries, a particularly worrying fact given that bilingual dictionaries have been proved to be learners' favourite reference tool. The lack of representative translation corpora is partly responsible for this failure to keep pace. However, I will show that the judicious use of currently available monolingual and bilingual corpus resources, limited though they may be, can already bring about substantial improvements to bilingual dictionaries, in particular to one of their weakest aspects: their phraseological coverage. I will also highlight the value of adding learner corpus data to the lexicographer's monolingual and bilingual corpus base and illustrate the benefit with the *Louvain English for Academic Purposes Dictionary*, a customisable web-based tool designed to help non-native speakers of English write academic texts.

Keywords: phraseology, learner corpora, bilingual corpora, bilingual dictionaries

One Model, Many Languages? An approach to multilingual content

Judy Pearsall
Oxford University Press
Judy.Pearsall@oup.com

In 2015, Oxford Dictionaries launched the Oxford Global Languages (OGL) programme. OGL is an ambitious programme aimed at bringing lexical content online for 100 of the world's languages and making it available to developers, consumers, licensees, and researchers for a wide variety of uses. From the outset, the vision was to pursue a single model for creating, editing, storing, and delivering content across different languages in order to create lexical content that was sustainable, flexible, and agnostic, as well as capable of being interlinked and interoperable. We will explain the approach we have taken to address this challenge, from Domain Modelling and the creation of a single data model, to the migration of content to a new platform and data format, to the lexical framework, lexicographical methodology, and training involved. We will present some of the challenges and issues we have faced and how successful we have been in tackling them so far. Finally we will look at what we have learned and the possible future of further extending a single model for both lexical and broader language content.

Lexicography between NLP and Linguistics: Aspects of Theory and Practice

Lars Trap-Jensen
Society for Danish Language and Literature
ltj@dsl.dk

Over the last hundred years, lexicography has witnessed three major revolutions: a descriptive revolution at the turn of the 20th century, a corpus revolution in the second half of the 20th century, and the digital revolution which is happening right now. Finding ourselves in the middle of a radical change, most of us have difficulties orienting ourselves and knowing where all this is leading. I don't pretend to know the answers but one thing is clear: we cannot ignore it and carry on as normal. In this article, I will discuss how lexicography and natural language processing can mutually benefit from each other and how lexicography could meet some of the needs that NLP has. I suggest that lexicographers shift their focus from a single dictionary towards the lexical database behind it.

Keywords: e-lexicography, NLP, interoperability, data structure