Kernerman
DICTIONARY News

Number 25 • July 2017
kdictionaries.com/kdn

Adam Kilgarriff Prize, 2017

The Adam Kilgarriff Prize was set up in 2016 as a memorial to our multi-talented friend, who contributed so much to the fields of lexicography, corpus linguistics, and NLP. Along with the other Trustees, I'm delighted that the Prize has got off to such a terrific start. Not only did we receive eight excellent applications, but the projects they were based on reflected interesting and original work across the whole spectrum of fields in which Adam himself was engaged. There were submissions in areas as diverse as corpus building, software tools for language research, named entity resources, translation systems, machine learning, and dictionary-creation.

With so many high-quality applications, selecting a winner was a challenging process for the six Trustees, involving several lengthy Skype meetings. But our eventual decision was unanimous, and we're all thrilled that the first Adam Kilgarriff Prize is going to Dr. Paweł Rutkowski. Paweł and his team at the University of Warsaw have been working for over five years on the development of resources for users of Polish Sign Language, including an innovative corpus and a corpus-based dictionary. Dr. Rutkowski explains his project's goals on the last page of this issue. He has accepted an invitation from the organizers of eLex 2017 in Leiden to give one of the keynote talks at the conference, where he will receive the Prize in person from Adam's spouse, Gill Lamden.

We congratulate all the applicants for their impressive submissions, and for making the first iteration of Adam's prize such a success. Finally, I take this opportunity to thank my fellow Trustees for the great care and thoroughness which they devoted to the task of evaluating all the submissions. We are all confident that this year's Adam Kilgarriff Prize has a worthy winner, and we look forward to inviting applications for the 2019 Prize in due course.

Michael Rundell

The Trustees of the Adam Kilgarriff Prize are Miloš Jakubíček, Ilan Kernerman, Iztok Kosem, Michael Rundell (Chair), Pavel Rychlý, and Carole Tiberius.

25th jubilee issue with a focus on lexicography and computational linguistics

K DICTIONARIES LTD
8 Nahum Hanavi Street
Tel Aviv 6350310 Israel
+972-3-5468102
kdl@kdictionaries.com
http://kdictionaries.com

Editor | **Ilan Kernerman**

ISSN 1565-4745

Introducing LDL4HELTA: Linked data lexicography for high-end language technology application

Martin Kaltenböck and Ilan Kernerman

Martin Kaltenböck is Co-Founder, CFO and Managing Partner of Semantic Web Company. He leads the LDL4HELTA project, and leads and works in several national and international research, industry and public administration projects – mainly as regards project management, requirements engineering, and community and communication activities. He studied communication, psychology and marketing at the University of Vienna. https://linkedin.com/in/martinkaltenboeck/

Background

Business is becoming increasingly globalized, and enterprises as well as public organisations are increasingly acting in multiple areas, facing challenges of cross-lingual and inter-cultural barriers. English has been crowned a global language, yet regional identities flourish as well, and this trend correlates with a rise in human and machine solutions to facilitate and enhance communication across languages and cultures. Looking at the European market, for example, we see 24 working languages in EU28, which make cross-border services considerably complicated. This calls for powerful language technology (LT) and intense efforts to enable and materialize the vision of a multilingual digital single market (defined as a priority area of the European Commission, see http://ec.europa.eu/priorities/digital-single-market/).

As a result, we see a continuously growing LT market, primarily in Europe but also worldwide. The growth leads LT entrepreneurs to suggest solutions for data- and information-driven organisations to work internationally, to efficiently store, access, integrate and disseminate their data, and to allow for both inter- and intra-organisation communication across borders and language barriers by utilizing software tools.

Although the emerging LT industry is fairly young, it has rapidly changed the rules of the game and excluded major old-time players. It ranges as widely as machine translation (e.g. Google Translate, Translation Memory tools), speech technologies (e.g. Apple's Siri digital assistant), education (e.g. e-learning), text processing (e.g. MS Word / Office 365), text mining, content analysis, localization, etc. New major players consist of a range of world leading software and hardware corporations (Google, Apple, Microsoft, Facebook, SAP, IBM, Intel, Amazon, etc.) as well as newcomers offering innovative solutions.

Semantics and lexicography

To develop and provide satisfactory LT mechanisms and tools we need suitable software as well as high quality data and information in the form of corpora, dictionaries, word form lists and other fine lexical resources. Lexicography is a vital component in this ecosystem. It follows mainly a qualitative approach that tends to be both time-consuming and cost-intensive. Interoperating lexicography with emerging Semantic Web methods and technologies is already underway, mainly as part of academic research, but mainstream lexicographic applications still make little use of state-of-the-art linked data (LD) resources. LT, on the other hand, follows mainly a quantitative approach along statistical and machine learning technologies. Bridging the gap between these qualitative and quantitative approaches is a huge challenge, and new solutions that combine these fields successfully stand good chances to be useful for the market.

In this context, the emerging Semantic Web, with new LD and semantic technologies, offers innovative means for information and data retrieval, knowledge management systems, and other applications for lexical data exchange and integration. However, while existing methodologies are becoming mature, they still lack sufficient refinements for data quality mechanisms, provenance methods and security issues.

There are new RandD initiatives that aim to bridge the gap between these disciplines, but only a few commercial applications. Groundbreaking projects include LIDER (http://lider-project.eu/) and LOD2 – Creating Knowledge out of Interlinked Data, which included the development of 45+ LOD software components (http://lod2.eu). Several freely available sources (and semi-open lexicographic resources) have also been developed, such as the Linguistic Linked Open Data Cloud (http://linguistic-lod.org/llod-cloud), WordNet

(https://wordnet.princeton.edu/), BabelNet (http://babelnet.org/), or DBpedia (http://dbpedia.org). Although these sources are comprehensive and useful, as well as available in machine readable formats (often providing an API) that allow relatively easy and efficient data integration, their main drawbacks still regard the content quality and (in)completeness.

The need remains to combine such openly available sources with quality lexicographic resources, including monolingual, bilingual or multilingual dictionaries that offer comprehensive data such as precise definitions, examples of usage, and other grammatical and semantic information, among others.

The LDL4HELTA project

Linked Data Lexicography for High-End Language Technology Application (LDL4HELTA, https://ldl4.com/) attempts to deal with the issues described above by combining lexicography with LD and integrating closed data sources with open ones to develop new LT methods and tools. This project is part of the EUREKA bilateral Austria-Israel RandD framework (http://eurekanetwork.org/project/id/9898), endorsed and supported by the Austrian Research Promotion Agency and the Israeli Chief Scientist Office (Israel Innovation Authority). It is led by Semantic Web Company (SWC) and K Dictionaries (KD), with scholarly cooperation of the Austrian Academy of Sciences and Polytechnic University of Madrid.

The project brings together lexicographic resources of KD with SWC's expertise in semantic technologies for the development of new products and services, to help the international LT market meet the fast-growing demands for dedicated language-independent, language-specific and cross-language solutions. These, in turn, will enhance cross-lingual search and usage for multilingual data management and integration. This entails:

- Enhancing knowledge and technology transfer between the partners in lexical methodologies and LD and semantic technologies;
- Combining state-of-the-art lexicographic and LT resources with Semantic Web and LD mechanisms to bridge the gap between them and generate new cross-language lexical tools and services;
- Integrating existing and new tools of the partners to give way to improved enterprise-ready software and data solutions for a wider market;
- Developing new software components for to upgrade data quality.

In order to provide the above-mentioned solutions, an integrated multilingual metadata and data management approach is needed, and this is where SWC's PoolParty Semantic Suite plays a crucial role. As PoolParty follows W3C Semantic Web standards (http://w3.org/standards/semanticweb/), such as SKOS (http://w3.org/2004/02/skos/), it already incorporates language-independent-based technologies. However, as regards text analysis and extraction, the ability to process multilingual information and data is key for success – which means that such systems need to speak as many languages as possible.

The cooperation with KD in the course of LDL4HELTA will enable PoolParty Semantic Suite to continuously "learn to speak" more and more languages, and do so more precisely, by making use of KD's rich multi-language lexical content and its know-how in lexicography as a base for improved text analysis and processing.

The first goal of the LDL4HELTA project is to model and convert KD data into RDF format, make it enrichable by third-party sources, by applying the Linked Data Design Principles proposed by Tim Berners Lee (https://w3.org/DesignIssues/LinkedData.html), and to make use of a SPARQL endpoint as an API to enable complex and flexile data querying.

The second goal is to improve word sense disambiguation as regards entity extraction and semantic annotation. Several methods are combined to attain this purpose, including (i) using dictionary data (ii) using thesauri and knowledge models, (iii) making use of corpora and freely available lexical resources, and (iv) integrating users' first-choice mechanisms.

The project started in July 2015 and ends in September 2017. It is supported by an advisory board including Prof. Christian Chiarcos (Goethe University, Frankfurt), Mr. Orri Erling (Google, San Francisco), Prof. Asunción Gómez Pérez (Universidad Politécnica de Madrid), Dr. Sebastian Hellmann (Leipzig University), Prof. Alon Itai (Technion, Haifa), and Ms. Eveline Wandl-Vogt (Austrian Academy of Sciences).

Ilan Kernerman is CEO of K Dictionaries, leading strategic development and cooperation. He edits and publishes *Kernerman Dictionary News,* has co-edited conference papers and other collections, and been involved in international lexicography associations and projects, most recently Asialex president (2015-2017) and the Globalex initiative. His interests include lexicography and the interoperability with NLP and knowledge systems. http://kdictionaries.com/ilank_2015.pdf

K Dictionaries creates cross-lexical resources for 50 languages and cooperates with industry and academia partners worldwide. Based in Tel Aviv and incorporating cutting edge pedagogical and multilingual lexicography methodologies, it develops manually crafted and automatically generated linguistic data serving natural language processing technologies for human and machine use.
http://kdictionaries.com

Triplifying a dictionary: Some learnings

Timea Turdean and Shrikant Joshi

Timea Turdean is Technical Consultant at Semantic Web Company, Vienna. In her current position she supports clients and partners integrating semantic technologies and is involved in different research projects dealing with linguistic data, earth observation data and publication data. She holds a MSc. from Vienna University of Technology, and her background is in text mining and sentiment analysis. timea.turdean@semantic-web.com

1. Introduction

The Linked Data Lexicography for High-End Language Technology (LDL4HELTA) project[1] was launched in cooperation between Semantic Web Company (SWC)[2] and K Dictionaries (KD)[3], combining lexicography and computational linguistics with semantic and linked (open) data mechanisms and technologies. One of the implementation steps of the project was to create a language graph from the dictionary data. The input data consists of the Spanish lexicographic resource of KD, which is translated into multiple languages and is available in XML format. The data needed to be triplified (that is, converted to RDF[4]) for several purposes, including enhancing its enrichment with external resources.

Section 2 of this article describes previous work carried out in this domain. Section 3 discusses in detail the actual process of triplification of the dictionary XML into RDF. An interesting experiment was carried out by using and applying the same principles for the translation of a dictionary, as described in Section 4. Although the initial success has ratified the process, some work is still required to explore and enhance it further, which is described as part of the conclusions in Section 5.

2. Previous work

There are different initiatives and efforts that investigate the process and usefulness of triplifying lexicographic data. Terminesp[5] is a well-known database that was transformed into RDF following linked data best practices (cf. Gracia 2015). Our work builds on the findings of Klimek and Brümmer (2015), who have investigated the usage of the Lemon model[6] on KD's German lexicographic XML data, and demonstrated how it can be represented in RDF and noted some missing elements that needed to be reconsidered. Bosque-Gil et al. (2016) also report about combining linked data in lexicography, particularly regarding usage of the Ontolex model[7] on the monolingual part of KD's Spanish dataset mentioned above.

3. Triplification process

The triplification of a dictionary is a process of mapping its data (which in KD's case is in propriatory XML format) to RDF triples. Following the triplication process, the resulting data was stored in a database to facilitate further processing.

In previous works, an RDF lexicographic model was proven to work for KD's lexicographic resources. The present article reports on how this model was applied on the Global Spanish dataset (i.e. the monolingual core and its translations in other languages) and triplified. In the process we ensured that the RDF complied with Semantic Web (SW) standards[8].

3.1. Nature of a dictionary entry

The XML format of KD's Global Spanish dataset consists of a complex structure containing nested components. Each word constitutes an entry, containing information such as: pronunciation; inflections; range of application; sense indicators; compositional phrases; translations (of different components); alternative scripts; register; geographical usage; sense qualifier; version; synonyms; lexical sense; examples of usage; homograph information; language information; specific display information; identifiers; and more…

Entries can have predefined values that can recur, but their fields can also have so-called free values, which can vary too, including: Aspect; Tense; Subcategorization; Subject Field; Mood; Grammatical Gender; Geographical Usage; Case; and more…

3.2. Constructing a lexical model

After studying the entry structure, it was necessary to construct a model representing the entries in the SW conceptual form to go from the dictionary's XML format to its triples. The model was designed by Bosque-Gil et al. (2016), and an example representing two Spanish words having senses that relate to each other is presented in Figure 1.

1 https://ldl4.com/
2 https://www.semantic-web.at/
3 http://kdictionaries.com/
4 https://www.w3.org/RDF/
5 http://linguistic.linkeddata.es/terminesp/
6 http://lemon-model.net/
7 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
8 http://semanticweb.org/wiki/Semantic_Web_standards.html

Usually, when modelling linked data or just RDF it is important to make use of existing models and schemas to enable easier and more efficient use and integration. A well-known lexicon model is Lemon[9], whose core path can cover some of this dictionary's needs (cf. Klimek and Brümmer, 2015), but not all of them. The Ontolex model[10], which is more complex and considered to be the evolution of Lemon, offers more capabilities in this regard. However, also after adapting the KD data to the OntoLex model, some pieces of information were still missing and an additional ontology was needed to be created to cover all such elements and catch the specific details that did not get sufficiently treated (such as the free values). We named this model extension OntolexKD.

The process used to do the mapping from KD's XMLs to RDF consists of several steps. This can be visualised as a processing pipeline which manipulates the XML data. The tool that we used for this mapping was UnifiedViews[11]. This is an ETL (Extract, Transform and Load) tool with which you can configure your own data processing pipeline to generate RDF data. One of its use cases is to triplify different data formats and store the resulting RDF data in a database. Our processing pipeline appears in UnifiedViews as displayed in Figure 2.

The pipeline is composed of data processing units (DPUs) which communicate with each other iteratively. In a left-to-right order, the process outlined in Figure 2 represents:

- A DPU used to upload the XML files into UnifiedViews for further processing;
- A DPU which transforms XML data to RDF using XSLT[12]. The style sheet is part of the configuration of the unit;
- The .rdf generated files are stored on the filesystem;
- Finally, the .rdf generated files are uploaded into a triple store, such as Virtuoso Universal Server[13].

3.3. URIs

Complexity increases also through the URIs (Uniform Resource Identifier) that are needed for mapping the information in the dictionary since linked data requires every resource to have a clearly identified and persistent identifier. The start was to represent a single word (headword) under a desired namespace and build on it to associate it with its part of speech, grammatical gender and number, definition and translation.

The base URIs follow the best practices recommended in the ISA study on persistent URIs[14] following the pattern: http://{domain}/{type}/{concept}/{reference}.

An example of such URIs for the forms of a headword is:

- http://kdictionaries.com/id/lexiconES/entendedor-n-m-sg-form
- http://kdictionaries.com/id/lexiconES/entendedor-n-f-sg-form

9 http://lemon-model.net/
10 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
11 https://unifiedviews.eu/
12 https://www.w3.org/Style/XSL/
13 https://virtuoso.openlinksw.com/
14 http://philarcher.org/diary/2013/uripersistence/

Shrikant Joshi holds a PhD in Linguistics from Université de Lausanne, with a focus on the semantics of affixation, its formalisation and subsequent computational processing, and a BE in Electronics Engineering and an MA in French from University of Pune. He has been teaching courses in NLP, French and German Linguistics at the University of Pune as a visiting lecturer. Currently he is working as Technical Consultant and Researcher at Semantic Web Company. shrikant.joshi@semantic-web.com

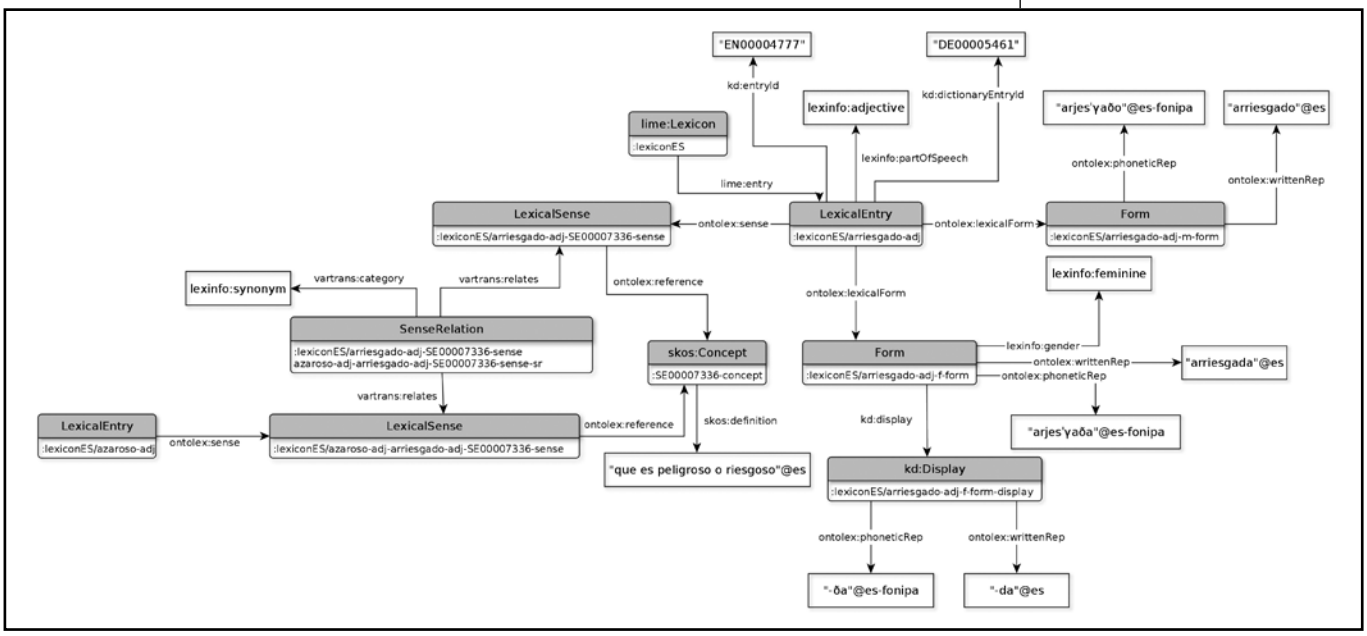


Figure 1: Language model example

TIAD shared task 2017 – Translation Inference Across Dictionaries

The first shared task on Translation Inference Across Dictionaries was aimed to explore best methods and techniques for automatic generation of new bilingual dictionaries based on existing resources. It relied on extracts from 15 bilingual dictionaries of K Dictionaries (KD) for developing three new language pairs that were validated against existing KD data and by human translators.

TIAD 2017 was organized by Noam Ordan, Morris Alper and Ilan Kererman (KD) and Jorge Gracia (OEG, Madrid Politechnic University). The results were presented in a workshop co-located with the Language, Data and Knowledge conference at NUI Galway on June 18, 2017 by four teams:

- Kathrin Donandt, Christian Chiarcos and Maxim Ionov; Goethe University, Frankfurt
- Tom Knorr; Neurocollective, San Francisco CA
- Thomas Proisl, Philipp Heinrich, Stefan Evert and Besim Kabashi; Erlangen University
- Uliana Sentsova; National Research University Higher School of Economics, Moscow

The papers are published as part of the LDK 2017 Workshop Proceedings http://ceur-ws.org.

Noam Ordan

https://tiad2017.wordpress.com/

These two URIs represent the singular masculine and singular feminine forms of the Spanish word *entendedor*.

- http://kdictionaries.com/id/lexiconES/entendedor-adj-form-1
- http://kdictionaries.com/id/lexiconES/entendedor-adj-form-2

If the dictionary contains two different adjectival endings, as with *entendedor* which has different endings for the feminine and masculine forms (*entendedora* and *entendedor*), and they are not explicitly mentioned, then we use numbers in the URI to describe them. If the gender is explicitly mentioned, then the URIs would be:

- http://kdictionaries.com/id/lexiconES/entendedor-adj-form
- http://kdictionaries.com/id/lexiconES/entendedora-adj-form

In addition, it should be considered that the aim of triplifying the XML was for all these headwords with senses, forms and translations, to connect and be identified and linked following the SW standards.

One of the last steps of complexity was to develop a generic XSLT which can triplify all the different languages of this dictionary series and store the complete data in a triple store. The question remains whether the design of such a universal XSLT is possible while taking into account the differences in languages or the differences in dictionaries.

4. Application and exploration

We tried to investigate also whether the automated resource linking could help with the translation of one dictionary into another the language. Two bilingual dictionaries were considered - English(en)-German(de) and German(de)-English(en).

For the word *bank* the following translations are found:

Bank (de) – *bank* (en) – German to English
bank (en) – *Bank* (de) – English to German

The URI of the translation from German to English was designed to look like:

- http://kdictionaries.com/id/tranSetDE-EN/Bank-n-SE00006116-sense-bank-n-Bank-n-SE00006116-sense-TC00014378-trans

And the one for the translation from English to German would be:

- http://kdictionaries.com/id/tranSetEN-DE/bank-n-SE00006110-sense-Bank-n-bank-n-SE00006110-sense-TC00014370-trans

In this case, both represent the same translation but have different URIs because they were generated from different dictionaries (in accordance with the translation order) that need to be mapped to each other so as to represent the same concept.

The word *Bank* in German can mean either a bench or a bank in English. When either of these English senses is translated back into German the result is the German word *Bank*. It is, however, not possible to determine which sense out of the two was translated unless the URI that contains the sense ID is included. It is also important to maintain the order of translation (source-target) but later map both translations to the same sense and same concept. This is difficult to establish automatically.

5. Future work

The actual overlap and automatic linking of the dictionary resources remains to be tested. There are also some lexicographic elements which were not covered by the new OntolexKD model and need to be added.

There is also the necessity to verify and check for differences between KD's XML dataset and the derived KD's triplified dataset. For this, SPARQL queries need to be created that validate and verify the resulting RDF.

References

Bosque-Gil, J., J. Gracia, and A. Gómez-Pérez. 2016. Linked data in lexicography. *Kernerman Dictionary News* 24, 19-24.

Bosque-Gil, J., J. Gracia, E. Montiel-Ponsoda, and G. Aguado-de Cea. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *Proceedings of GLOBALEX 2016 Workshop at the 10th Language Resources and Evaluation Conference (LREC 2016), Portorož, (Slovenia).*

Gracia, J. 2015. Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News* 23, 1-4.

Klimek, B., and M. Brümmer. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News* 23, 5-10.

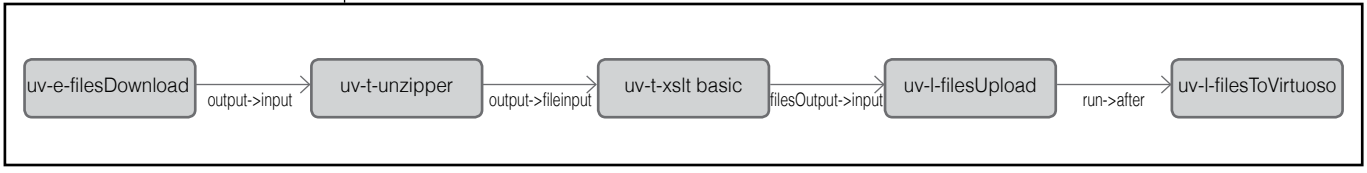


Figure 2: UnifiedViews pipeline used to triplify XML

Towards a module for lexicography in OntoLex

Julia Bosque-Gil, Jorge Gracia and Elena Montiel-Ponsoda

1. Introduction

Over the past few years, more and more efforts are being devoted towards the conversion of dictionaries into Linguistic Linked Data (LLD), based on Lemon (McCrae et al., 2012) and its more recent version OntoLex-Lemon[1], a de facto standard to represent ontology-lexica on the Web. These works aim both to enrich the so-called Linguistic Linked (Open) Data cloud[2] with lexical information to be consumed by natural language processing (NLP) tools, and to build bridges between the lexicography and semantic web communities. Recent projects such as LIDER[3], or on-going ones such as ENeL[4], LDH4HELTA[5] and LiODi[6], promote the adoption of linked data technologies in the work with lexicographic resources focusing on language technologies, e-lexicography and linguistic research, respectively.

Nonetheless, the conversion of a lexicographic resources to OntoLex is not always straightforward. Lemon was initially developed to enrich a given ontology with a lexical layer, and not with the idea of rendering any already existent dictionary to LLD. A majority of scholars working on this field, however, are turning to Lemon or OntoLex in pursuit of the latter objective. The more numerous and resource-specific the annotations in a dictionary are, the more complex the modeling solutions are, especially if until then the dictionary was targeted at human users. We are aware that some solutions exceed the needs of lexical information that some NLP tools require. However, if we are also aiming to bring linked data to lexicography, all dictionary content must be taken into account and must be retrievable once converted to LLD, i.e, migrating to LLD should imply no information loss. This means that structural aspects of the dictionary, as for instance senses and homographs order, along with the sub-sense hierarchy some dictionaries display, should be kept in mind when offering modeling solutions. There is a range of dictionary annotations (domain of usage, region, frequent use tags, restrictions on number and gender depending on a sense, etc.) that affect word meaning and language usage and are not structural in nature. Collocations, idioms, context indicators, semantic selection, etc. are presented differently in dictionaries and modeling them is not trivial.

The natural doubt that would be entertained by many experts is whether OntoLex is supposed to provide the means to model all aspects of a dictionary or whether this is outside its scope of ontology lexicalization, and therefore should be tackled by another initiative. In this paper we motivate our insights on OntoLex to enable dictionary representation as LLD in all its granularity, and advocate for the creation of a lexicography-specific module that would gather elements concerning dictionary structure and annotations. The module could also link to other modules that might be proposed, such as an etymology-oriented one to support etymological dictionaries.

The rest of the paper is organized as follows: Section 2 goes through state-of-the-art of LLD and lexicography and some of the problems encountered during the representation of dictionaries as LLD. Our motivation for OntoLex to be able to tackle those and the issues presented throughout the paper are stated in that section as well. Section 3 describes five of a series of issues we identified in our work modeling and analyzing dictionary entries, and which we argue serve as input for discussion on the need for a module for lexicography. Our initial approaches towards such a module and a description of how it would solve the described issues are outlined in Section 4, while Section 5 offers some concluding remarks.

2. Background and motivation

There have been several reports in the literature on the conversion of dictionaries to LLD, most of them relying on Lemon or OntoLex. However, proprietary formats, such as that of K Dictionaries (KD)[7], often have XML tags used in their annotation schemes that refer to linguistic categories or features which are not present in available repositories of linguistic categories or which lack a compatible definition that prevents us from reusing the ontology entity at hand. Ad

1 http://www.w3.org/2016/05/ontolex/
2 http://linguistic-lod.org/llod-cloud
3 http://lider-project.eu/
4 http://elexicography.eu/
5 http://ldl4.com/
6 http://acoli.informatik.uni-frankfurt.de/liodi/
7 http://kdictionaries.com/

Julia Bosque Gil is a PhD student at the Ontology Engineering Group (OEG) since October 2014. She holds a B.A. in German Linguistics and English from the Humboldt University of Berlin and an M.A. in Computational Linguistics from Brandeis University, Waltham. Her interests include the lexicon-ontology and the syntax-semantics interfaces, relations between the lexicon and syntax, semantic annotation and the representation of (multilingual) language resources as linked data. She has been working in the conversion of multilingual terminologies to RDF as part of the LIDER project and in the modeling of lexicographic data as linked data in collaboration with K Dictionaries and Semantic Web Company. For her PhD thesis she is investigating the use of linguistic linked data for research in linguistics and lexicography.
jbosque@fi.upm.es

Jorge Gracia is a postdoctoral researcher at Ontology Engineering Group, Universidad Politécnica de Madrid. He got his PhD in Computer Science at University of Zaragoza in 2009, with a thesis about heterogeneity issues on the Semantic Web. His current research interests include multilingualism and linked data, linguistic linked data, and cross-lingual matching and information access on the Semantic Web. His current research focus is on exploring how to move language resources (lexica, dictionaries, corpora, etc) from their data silos into the multilingual Web of Data and make them interoperable, in order to support a future generation of linked data-aware NLP tools.
http://jogracia.url.ph/web/

hoc vocabularies were defined to migrate content from the German monolingual dictionary of KD's Global Series (Klimek and Brümmer, 2015) and its Spanish multilingual set (Bosque-Gil et al., 2016). These works approached issues which affect, for example, the relation between a lexical sense and the lexicalized phrases and idioms in which it occurs, regional restrictions, lexical and semantic selection (in general) of lexical entries, groups of homographs, tone and register indications, inflection groups, context of use, frequency modifiers to register, etc. Multilingual dictionaries pose further problems due to the modeling of examples and translations of examples, as well as alternative forms of those translations (e.g. an example in English translated to Japanese in kanji and hiragana, and that translation in turn with a transliteration in rōmaji). The set of thirteen dictionaries (dialectal, bilingual, monolingual, historical, etc.) converted as part of the ENel Action (Declerck and Wandl-Vogt, 2015) required the definition of new properties to encode different types of temporal information and etymological aspects.

Structures typically found in dictionaries, such as the sense and sub-sense hierarchy in an entry, are not trivial to model either. polyLemon (Khan et al., 2016), developed as part of the conversion of the Liddell-Scott Greek-English Lexicon to Lemon, was suggested in order to capture the sense and sub-sense structure in dictionaries using properties such as `senseChild` and `senseSibling` to relate senses and their parent senses in the dictionary entry.

The accurate representation of etymological information as LLD is key in the conversion of historical and etymological dictionaries. An extension to Lemon, Lemonet, to represent etymological information of lexical entries was proposed (Chiarcos and Sukhareva, 2014) and, more recently, a revisited version builds upon the properties suggested for the modeling of the etymological WordNet[8] to undertake the conversion of the Tower of Babel (Starling) in the LiODi project (Abromeit and Fäth, 2016). Some recent work on the conversion of the classical Arabic dictionary *Al-Qamus* to Lemon and LMF has been undertaken (Khalfi et al., 2016), but no pointers or traceback to the original structure are given in the work.

Alternatives to the use of OntoLex are available as well. The Oxford Global Languages Ontology (OGL) (Parvizi et al. 2016) has been developed to model and integrate multilingual linguistic data from Oxford Dictionaries and emerges as an ontology exclusively created to meet dictionary representation requirements. It accounts for a range of information found in dictionaries, from inflected forms to semantic relations, pragmatic features and etymological data. The focus is laid on the representation of grammatical information with cross-linguistic validity and the respect towards grammar traditions. However, some modeling decisions and class definitions differ from those suggested in OntoLex (e.g. `Form` in OntoLex vs. a `Form` in OGL) and the emphasis is not set on the reuse of available ontology entities.

In this position paper we do not focus on a particular kind of lexical information present in dictionaries (e.g., etymology or morphology) but we aim to highlight some difficulties in the modeling of dictionary entries without information loss. Thus, we will not target the representation of resource-specific features of particular dictionaries. Taken into account the problems reported in the literature, and after analyzing dictionary entries in e-dictionaries of English (*Oxford Living Dictionaries Online*; *Merriam Webster Dictionary Online*; *American Heritage Dictionary Online*; *COBUILD Advanced English Dictionary* and *Collins English Dictionary*), German (*Duden Online Wörterbuch*; *PONS Deutsch als Fremdsprache Online Wörterbuch*), and Spanish (*Diccionario de la Lengua Española*; *Clave*: *diccionario de uso del español actual*), we report on some of the issues we gathered which may pose problems for the modeling with OntoLex and which we believe call for the definition of a new module to account for them. Future steps include the analysis of dictionaries in languages that are underrepresented in the LLOD cloud (e.g. Japanese) to identify further representation challenges.

We ground our proposal for a lexicography module on the following four points: (1) the use of OntoLex by the majority of the community to convert linguistic resources to LLD instead of to lexicalize ontologies, (2) the nature of Lemon being descriptive but not prescriptive and the respect towards different lexicographic views, (3) the coming together of the lexicography and the Semantic Web communities and potential benefits that LLD may bring about to lexicography, assuming it involves no information loss, and (4) the reuse of already available mechanisms in OntoLex.

3. Issues

In the following we report on some of the issues we have come across after our experiences in converting dictionaries to

8 http://www1.icsi.berkeley.edu/~demelo/etymwn/

LLD and our analysis of dictionary entries in English, German and Spanish. Here we restrain ourselves to issues that reveal current limitations of the OntoLex model, i.e, cases in which applying the Lemon core implies a different view on the data than the one provided in the original resource and, therefore, an information loss (type 1, hence T1), and missing entities, e.g. a property or a class, to account for information mostly found in dictionaries (type 2, hence T2). We have already raised some of these issues as input for discussion to the OntoLex community.[9]

Issue 1 (T1). Headwords that can take different parts-of-speech
Both Lemon and OntoLex specify a lexical entry as a word, a multiword expression or an affix with a single part-of-speech, morphological pattern, etymology and set of senses.[10] However, a headword in a dictionary may occur with different parts-of speech depending on context and its senses are nonetheless defined in the same dictionary entry, all of them derived from the same etymology (no homonymy involved). Applying the OntoLex model would imply the generation of several `ontolex:LexicalEntr`[ies], one per each part-of-speech the headword can take. Splitting the dictionary entry into several lexical entries would cause loss of information (shared etymology, pronunciation, senses implicitly related) and does not keep track of the dictionary representation. Examples: *poison*, *bread*, *water* (noun and verb), and Spanish *lento* 'slow, slowly' (adjective and adverb) and *alto* 'tall, loudly, height' (adjective, adverb and noun).

Issue 2 (T1). Lexical sense requiring a particular form
Some senses of a dictionary headword require a particular form, e.g. in English a plural form or in Spanish a masculine or feminine one. Since the meaning in these cases is associated with the form and it may differ significantly from other senses that do share gender or number features, splitting the dictionary entry into different lexical entries would be an option (see Issue 1). An alternative is the linking of that sense to elements in a catalog of grammatical categories which encode those grammatical restrictions, but we would need an exhaustive list of them for this option to be applicable. Examples: *refreshment(s)*, Spanish *cometa* (m.) 'comet', (f.) 'kite'. In these cases, the dictionary entry can be a single one (e.g. *refreshment* in English or *cometa* in Spanish) but one of its senses indicates a preferred form. In the case of *refreshment*, the plural form is used if the intended meaning is snacks and beverages; with *cometa*, the feminine form is applicable when referring to a kite, the masculine when denoting a comet. Further examples are *good(s)*, *manner(s)*; and Spanish *frente* (m.) 'front', (f.) 'forehead'.

Issue 3 (T2). Usage examples and their translations
Usage examples of a word or multiword expression are often provided in the definition of each of a dictionary entry's senses. Lexinfo[11] includes a property `lexinfo:senseExample` to describe an example of a sense (as a subproperty of `Lemon:definition`) and which is linked to the example data category in ISOCat.[12] Nonetheless, due to it being a datatype property, it does not enable including further information on the example or to establish translation relations among examples, which is common practice in bilingual and multilingual dictionaries. The Lemon model included a `Lemon:UsageExample` class and a property `Lemon:example` to link to it, but OntoLex does not cover this aspect yet. Examples: Spanish *preocuparse* 'worry'; *no hay por qué preocuparse* 'there is nothing to worry about' (*Collins English-Spanish Dictionary*).

Issue 4 (T2). Sense and homograph order
The order of senses may be based on frequency of use, date of origin, concreteness (from the most concrete to most abstract sense, etc.). Homographs are also given according to some ordering criteria that may vary from dictionary to dictionary. Their order should be searchable and retrievable as to recover the information provided in the original resource. Examples: *Boa*: noun. (1) any of a family (Boidae) of large snakes that kill by constriction and that includes the boa constrictor, anaconda, and python (2) a long fluffy scarf (*Merriam Webster Dictionary*)[13]; *bat1*: n. 1. A stout wooden stick; a cudgel [. . .]; *bat2*: n. Any of various nocturnal flying mammals of the order Chiroptera [. . .] (*American Heritage Dictionary*).

9 http://w3.org/community/ontolex/wiki/Lexicography
10 http://w3.org/2016/05/ontolex/#lexical-entries
11 http://lexinfo.net/ontology/2.0/lexinfo.owl
12 http://isocat.org/
13 Example of logical order of senses inspired by *Diccionario de la Lengua Española, Guía de Consulta*, http://dle.rae.es/

Elena Montiel-Ponsoda
is Associate Professor at the Applied Linguistics Department at Universidad Politécnica de Madrid (UPM) since 2012, and member of the Ontology Engineering Group since 2006. She got her PhD on Applied Linguistics from UPM in 2011. Her research interests are at the intersection between translation (and terminology) and knowledge representation, including among others: ontology localization and lexicalization, lexico-syntactic patterns for ontology development, functional models for deep semantics analysis, sentiment analysis, and linguistic linked data for content analytics. She is currently working on the representation of lexical resources according to the linked data paradigm, specifically, on how translation relations can help in the construction of the multilingual Web of Data.
emontiel@fi.upm.es

OntoLex 2017
1st workshop on the OntoLex model

The W3C OntoLex Community Group was launched in 2011, with Paul Buitelaar (INSIGHT. National University of Ireland, Galway) and Philipp Cimiano (CITEC, Bielefeld University, Germany) as chairs, with the goal to define a model for allowing to represent lexical knowledge in connection to ontologies [1]. The rationale behind the model is that semantics are captured by ontologies, and the role of the lexicon-interface is to link lexical entries to ontological entities expressing their denotational meaning, following a principle called *Semantics by Reference*. Based on five years of intensive discussions and work, the new OntoLex-Lemon model was launched in May 2016 [2] and is becoming the primary method for representing linked lexical resources on the Web of Data, not only for capturing the lexicon-ontology interface but for the representation of lexicographic resources as well. The model facilitates bridging the gap between the NLP and data science communities by making available and linking large amounts of quality lexical information to the knowledge represented on the semantic web, for example in graphs such as DBpedia, applications

Issue 5 (T2). Semantic selection
Some dictionaries indicate the semantic features of the lexical items that an entry (in one of its senses) selects or even the exact lexical items with which it collocates. This is usually indicated either with a specific tag (e.g. KD's Range Of Application), or in-between parentheses at the beginning of a definition. Examples are, for instance, the dictionary entry for the German verb *dämmen*, which in its sense 'to insulate, absorb, mute' selects arguments that denote warmth or sound (German *Wärme*, *Schall*, etc.) (KD), the adjective *cozy*, meaning beneficial to all those involved and possibly somewhat corrupt if predicated from a transaction or an arrangement (*Google Dictionary*); or the collocational measure words of *luck*: stroke, piece of (*Oxford Collocations Dictionary*). The OntoLex Syntax-Semantics Module (*synsem*) class `synsem:OntoMap` allows to map a syntactic frame to an ontology entity, so that the frame and its arguments are linked to the ontology elements that they lexicalize. Even though dictionaries commonly include information on subcategorization (transitive/intransitive/reflexive etc. annotations for verbs, for instance), details on the syntactic frame are not always provided beyond those annotations. Since in dictionary conversion we often lack a given ontology and detailed syntactic information is not provided, the mapping between syntactic arguments and ontology entities seems difficult to establish automatically via `synsem:OntoMap`: how do we automatically represent that the adjective *cozy* has a meaning only applied to transaction or agreement or that the measure words that collocate with *luck* are *stroke* or *piece* if the morphosyntactic information provided in the dictionary is just that *cozy* is an adjective and *luck* a noun? Furthermore, `synsem:condition` (in its turn subsuming `synsem:propertyRange` and `synsem:propertyDomain`) enables us to state constraints on the arguments of a predicate in a given ontology.14 The possibility of reusing it to state the constraints on syntactic arguments even in cases in which we lack a given ontology and therefore are not mapping to given ontology properties has to be further analyzed. In addition, the potential links between the modeled entries (e.g. *piece* and *luck*), i.e. the links at the lexical level, are also to be considered, for instance, by taking into account recent proposals on the representation of lexical functions as LLD (Fonseca et al. 2016).

14 http://w3.org/2016/05/ontolex/#conditions

4. A module for lexicography
The previous section dealt with some of the issues we encountered in our work with dictionaries and the potential ones that may rise with other lexicographic works that have not been migrated to LLD yet. In the following we draft a potential solution which can serve as a basis for a new module in OntoLex specifically developed for the representation of dictionaries after thorough revision and improvement according to the community's feedback.

In order to keep track of the dictionary representation and prevent any loss of information mentioned in Issue 1, related to the splitting of dictionary entries in several lexical entries, we propose a new class `DictionaryEntry`. This new class would both enable to group together lexical entries as well as to associate any information shared by all of them. In our view, we distinguish lexical entries and lexicons (as containers of lexical entries), from the original dictionary entry (the new class `DictionaryEntry`) and the original dictionary resource (`Dictionary`), which would serve in turn to record the provenance of each dictionary entry. Mirroring the `lime:Lexicon-ontolex:LexicalEntry` relation we suggest a `Dictionary-DictionaryEntry` one. Any lexical entry created during the conversion to LLD but not originally provided in the resource would then belong to a `lime:Lexicon`, but not to the instance of `Dictionary` representing that resource. A `lime:Lexicon` in English, for example, could aggregate lexical entries created on the fly by the LLD expert or original ones coming from as many English dictionaries as desired. These dictionaries can in turn differ in their modeling and their views on the data, their criteria of sense ordering or their structure.

Regarding Issue 2, the `DictionaryEntry` class would allow to divide a single lexical entry into several ones if desired, each with a different preferred form, while maintaining the original dictionary representation. If the dictionary entry is not split, the option of linking a sense to a grammatical restriction on gender or number from an external catalog would solve the issue, although the implications of this solution (its benefits and drawbacks) will need further analysis.

In order to represent usage examples and their translations (Issue 3) we propose to go back to `lemon:UsageExample` and link it to a `LexicalSense`. A new class `ExampleCluster` would link to `UsageExamples` that are translations from each other. The use of the vartrans

module to model translations among senses would imply the creation of lexical senses for each example, and therefore treating the example as a lexical entry, which we deem is beyond the definition of lexical entries.

Issue 4 was concerned with the order of senses in a dictionary entry and the order of homographs in the macrostructure of the dictionary. There are different possible approaches to resolve this: reusing already available RDF mechanisms, reifying the sense order in a new class `SenseOrder`, or defining a new property `senseOrder` attached to the lexical sense. The first option involves the reuse of `rdfs:Container`[s] to declare with e.g. `rdf:_1, rdf:_2` that a particular sense is the first or the second one. However, cases in which a set of senses allows for various orderings, depending on the ordering criterion, or in which some senses come from different dictionaries (each with its order), should also be accounted for. The second option suggests that the sense order is reified in a class `SenseOrder` linked to the lexical sense. This class would enable us to record the position of that sense, its provenance (presumably an instance of the class `Dictionary`), and, if desired, the ordering criterion. If repeated senses were identified (e.g. senses that share a definition in both dictionaries), `SenseOrder` would allow us to have one single lexical sense with two different positions according to the two different orderings and dictionaries, in a similar fashion as two containers with two different sequences of senses. Alternatively, if we assume that a lexical sense always comes from just one dictionary source, a property `senseOrder` would suffice.

Issue 5, dealing with semantic selection, has been brought up for further discussion in this paper to see whether it could be covered by synsem module mechanisms or whether it would require new entities in the context of the lexicography module. As part of the conversion of the KD's *Global Spanish Multilingual Dictionary* (Bosque-Gil et al. 2016), the semantic selection information provided by KD's tag RangeOfApplication was captured by the use of `synsem:condition`. In that approach, `synsem:condition` would link a lexical sense to a blank node[15] with an `rdf:value` recording the strings given as arguments in the original data. This modeling allowed us to deal with the lack of a given ontology and detailed information on the syntactic frames of lexical entries for each of their senses. Thus, the focus was set on representing the data just as it was in its original format while being compliant with the OntoLex formal specification and reusing its elements as much as possible. We argue that the lexicography module should aim to set the basis to exploit at the dictionary's macro-structure level the potential benefits of establishing semantic relations among lexical senses based on lexical selection or among syntactic frames and arguments and the ontology entities that they denote. To this aim, overcoming the lack of detailed syntactic information in the dictionary as well as the lack of a given ontology to lexicalize becomes essential.

15 `synsem:condition` has `rdfs:Resource` defined as its range.

5. Conclusion

OntoLex is increasingly being used to convert linguistic resources to LLD outside the scope of ontology lexicalization. In this position statement we have drawn attention to a series of issues raised in the literature on LLD related to the conversion of dictionaries to LD and to five of the ones we came across in the same line of work and after a later analysis of several additional dictionaries. We argue that the OntoLex model should enable the preservation of the content and the structure of the original resource, even if the LLD expert opts for a different representation that is better suited to the data exploitation by external applications or is more in line with his or her view on the lexicon-ontology interface. We have outlined some of our insights on how to address these issues in a new module for lexicography. It would be compatible with the mechanisms suggested in the state-of-the-art on dictionaries represented as LLD, as of the moment of writing, and also with other potential modules for the encoding of specific lexical aspects (e.g. etymology). The final module is intended to be dictionary-agnostic in the sense that it should be applicable (and combined with other modules if necessary) to different kinds of dictionaries (e.g., general, collocations, learner's, etymological, historical, etc.). This would bring linked data (LD) closer to lexicography not only with the aim of leveraging already available dictionaries in LD for NLP tasks, but also for introducing LD in the work carried out in that discipline.

Acknowledgments

This work is supported by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R), the Excellence Network ReTeLe (TIN2015-68955-REDT), and the Juan de la Cierva program and by the Spanish

as part of the Global WordNet Association Collaborative Interlingual Index, or existing resources such as BabelNet, DBnary and commercial dictionaries.

OntoLex 2017 was co-located with the Language, Data and Knowledge conference [3, and see p.13] and presented the first opportunity for practitioners to meet to discuss the model, its applications and future development [4]. The participants have shown interest in continuing the development of OntoLex-Lemon, particularly with regard to lexicographic resources. In consequence, the group is starting to work on a new best practice document that will provide modeling examples and guidelines for how to use OntoLex specifically to represent lexicographic resources such as dictionaries. Then it will be decided whether to stipulate the proposed vocabulary elements or modeling solutions into the status of a new module or to keep the best practices proposal as an informal document.

Philipp Cimiano
Universität Bielefeld

[1] https://w3.org/community/ontolex/
[2[https://w3.org/2016/05/ontolex/
[3] http://ldk2017.org/
[4] http://ontolex2017.linguistic-lod.org/

Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program.

This contribution was presented at the 1st Workshop on the OntoLex Model, co-located with the Language, Data and Knowledge Conference (LDK 2017) in Galway, Ireland, on June 18.

Dictionaries cited

American Heritage Dictionary Online. Retrieved 25/05/2017 from http://ahdictionary.com/

Clave: diccionario de uso del español actual (online). Retrieved 25/05/2017 from http://clave.smdiccionarios.com/app.php

COBUILD Advanced English Dictionary. Retrieved 25/05/2017 from http://collinsdictionary.com/

Collins English Dictionary. Retrieved 25/05/2017 from http://collinsdictionary.com/

Duden Online Wörterbuch. Retrieved 25/05/2017 from http://duden.de/woerterbuch

Merriam Webster Dictionary Online. Retrieved 25/05/2017 from http://merriam-webster.com/

Oxford Living Dictionaries Online. Retrieved 25/05/2017 from https://en.oxforddictionaries.com/

PONS Deutsch als Fremdsprache Online Wörterbuch. Retrieved 25/05/2017 from http://de.pons.com/

Real Academia Española, Diccionario de la Lengua Española (DLE). Retrieved 25/05/2017 from http://dle.rae.es

References

Abromeit, F., and Fäth, C. 2016. Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF. In *LDL 2016 5th Workshop on Linked Data in Linguistics*, 11.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de-Cea, G. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *GLOBALEX 2016 Workshop at LREC 2016*, 65.

Chiarcos, C., and Sukhareva, M. 2014. Linking Etymological Databases. A Case Study in Germanic. *In 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, 41.

Declerck, T., and Mörth, K. 2016. Towards a Sense-based Access to Related Online Lexical Resources. In *Proceedings of Euralex 2016*, 342-355.

Declerck, T., and Wandl-Vogt, E. 2014. Cross-linking Austrian dialectal Dictionaries through formalized Meanings. In *Proceedings of Euralex 2014*.

Declerck, T., and Wandl-Vogt, E. 2015. Towards a Pan European Lexicography by Means of Linked Open. Data. In *Proceedings of eLex 2015*, 342-355.

Declerck, T., and Wandl-Vogt, E. 2014. How to semantically relate dialectal Dictionaries in the Linked Data Framework. In *8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities LaTeCH 2014*, 9-12.

Del Gratta, R., Frontini, F., Khan, F., and Monachini, M. 2015. Converting the Parole Simple Clips Lexicon into RDF with lemon. *Semantic Web 64*, 387-392.

El Maarouf, I., Bradbury, J., and Hanks, P. 2014. PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model. In *Proceedings of Language Resources and Evaluation LREC*.

Fonseca, A., Sadat, F., and Lareau, F. 2016. Lexfom: a lexical functions ontology model. In *COLING 2016*, 145.

Gracia, J., Villegas, M., Gómez-Pérez, A., and Bel, N. 2016. The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal*.

Khalfi, M., Nahli, O., and Zarghili, A. 2016. Classical dictionary *Al-Qamus* in lemon. In *Information Science and Technology CiSt., 2016 4th IEEE International Colloquium*, 325–330.

Khan, F., Díaz-Vera, J. E., and 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 co-located with 13th Extended Semantic Web Conference ESWC 2016.*, Heraklion, Greece, 37-46.

Klimek, B., and Brümmer, M. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News* 23, 5-10.

McCrae, J., Fellbaum, C., and Cimiano, P. 2014. Publishing and Linking WordNet using lemon and RDF. In *3rd Workshop on Linked Data in Linguistics*.

Parvizi, A., Kohl, M., González, M., and Saurí, R. 2016. Towards a Linguistic Ontology with an Emphasis on Reasoning and Knowledge Reuse. In *Proceedings of Language Resources and Evaluation LREC*.

Villegas, M., and Bel, N. 2013. PAROLE/SIMPLE "Lemon" ontology and lexicons. *Semantic Web Journal* II.

KDICTIONARIES

KD API

K Dictionaries is completing the development of an online API which will provide programmatic access to its rich cross-lingual lexicographic resources for 50 languages.

In addition to the well-formatted XML data, the API outputs developer-friendly JSON-LD, which complies with the familiar RESTful API standard. The JSON-LD encodes RDF linked data, making it highly compatible with complementary open linked linguistic data sets. Users will be authenticated and given access according to their account type.

There is also an editing pipeline in development in which editors, translators and crowdsourced suggestions will use API calls to consolidate suggested changes to K Dictionaries' data and direct them through quality assurance.

The API will be launched in September 2017. Registration is open at api@kdictionaries.com.

Morris Alper

1st Conference on Language, Data and Knowledge – Galway, 2017

On June 19-20, the first conference on Language, Data and Knowledge (LDK 2017) took place in Galway, Ireland, which was attended by over 100 participants from 27 countries. This new biennial conference series brings together researchers from across disciplines concerned with the acquisition, curation and use of language data in the context of data science and knowledge-based applications. Language datasets, such as corpora, typological resources and, of course, dictionaries, are of increasing importance to machine learning-based approaches in Natural Language Processing (NLP), linked data and Semantic Web research and applications that depend on linguistic and semantic annotation with lexical, terminological and ontological resources, manual alignment across languages or other human-assigned labels. The acquisition, provenance, representation, maintenance, usability, quality as well as legal, organizational and infrastructure aspects of language data are therefore rapidly becoming major areas of research that were at the focus of the conference. A further focus was the combined use and exploitation of language data and knowledge graphs in data science-based approaches to use cases in industry, including biomedical applications, as well as use cases in humanities and social sciences.

The LDK conference has been initiated by a consortium of researchers from the Insight Centre for Data Analytics (National University of Ireland Galway), InfAI (Leipzig University, Germany) and Goethe University Frankfurt, Germany, and a Scientific Committee of leading researchers in NLP, linked data and the Semantic Web, language resources and digital humanities. LDK 2017 was endorsed by several international organisations, namely DBpedia, the ACL Special Interest Group for Annotation (SIGANN), Global WordNet Association, CLARIN, Big Data Value Association (BDVA) and DARIAH Ireland.

The conference stands at the intersection of data science and NLP and brought together researchers from across computer science as well as linguistics and the humanities. The conference series is concerned with the creation of data resources as well as the metadata, development, evaluation, quality and legal aspects of publishing such data as well as knowledge graphs as a key focus for providing knowledge about the world in terms of ontologies, terminologies and linked data. Direct applications in NLP are of particular importance in certain semantic technologies, question answering, multilinguality and big data. Finally, applications of these technologies to domains as varied as digital humanities, enterprise data analytics and text mining of biomedical literature are of importance.

The conference featured invited talks from both industry and academia. Zoltan Szlavik from IBM Benelux gave a talk about cognitive computing and the Watson systems. Kathleen McKeown from the Data Science Institute of the Columbia University talked about the intersection of data science and NLP. Antal van den Bosch of Radboud University in the Netherlands and director of the Meertens Institute related data science technologies with new breakthroughs in digital humanities. Finally, Isaac Graham from NUI Galway talked about relevant features of the Irish and Welsh languages.

In addition to the main conference there were a number of workshops and tutorials. The 1st OntoLex Workshop discussed the development of the OntoLex model, a recent vocabulary from the World Wide Web Consortium, for representing dictionaries relative to ontologies as linked data on the Web. The Translation Inference Across Dictionaries (TIAD) shared task explored the automatic generation of bilingual dictionaries from existing resources and was co-organized by K Dictionaries and the Polytechnic University of Madrid. A further workshop was organized on Challenges for Wordnets, concerning the construction of dictionaries in the wordnet format, organized by the Global WordNet Association. Importantly, also a tutorial on text analytics for Digital Humanities and Social Sciences with CLARIN was organized jointly by CLARIN and DARIAH Ireland. Finally, a meeting of the DBpedia Association took place after the main conference.

The next edition of the LDK conference is planned to take place in Leipzig, Germany in 2019. More details about the conference can be found at http://ldk2017.org.

John P. McCrae, Paul Buitelaar, Christian Chiarcos and Sebastian Hellmann

GWC9
Singapore, 2018

The Ninth Global WordNet Conference (GWC 2018), will be held from 8 to 12 January, 2018, at Nanyang Technological University, Singapore. The conference focuses on wordnets, but is broad in scope, welcoming more general work on lexical semantics and lexicography, as well as word sense disambiguation. We especially invite papers addressing the following topics:

- Linguistics and lexical semantics
- Architecture of lexical databases
- Tools and methods for wordnet development
- Wordnets and applications
- Standardization, distribution and availability of wordnets and wordnet tools

Submissions are anonymous, and can be for long papers, short papers, project reports or demonstrations. The deadline for submissions is September 6, 2017, and acceptance will be announced to the authors by end of September. Final papers are due on October 11.

In conjunction with the conference we intend to hold two workshops: one on distributional semantics and one on technology enhanced learning.

Francis Bond
Nanyang Technological University

http://compling.hss.ntu.edu.sg/events/2018-gwc/

The advent of post-editing lexicography

Miloš Jakubíček

Miloš Jakubíček is an NLP researcher, software engineer and CEO of Lexical Computing (LC), a research company developing the Sketch Engine (SkE) corpus platform. His research interests are devoted mainly to effective processing of very large text corpora for lexicographic and linguistic tasks and syntactic parsing of morphologically rich languages. Since 2008 he has been involved in the development of SkE, in 2011 he became director of the Czech branch of LC leading the local development team of SkE, eventually becoming CEO of LC in 2014 after its founder Adam Kilgarriff. He is a fellow of the NLP Centre at Masaryk University where his interests lie mainly in morphosyntactic analysis of Czech and its practical applications.
milos.jakubicek@sketchengine.co.uk

The lexicographic landscape has been subject to two major disruptions over the past twenty years.

The first is related to the uptake of information technology and availability of text corpora. Lexicographers were on the forefront of the shift to empiricism in linguistics[1] and it was for good: a field that never seriously acknowledged any theoretic framework was starting to benefit more than any other linguistic discipline – practical needs for describing language as used were very high.

The second change, related to the first one, was without doubt the breakdown of traditional publishing business, manifested in the end of paper dictionaries (as well as by the fall of many renowned dictionary publishers). From the perspective of users, dictionaries are tools to be *used while doing something else*, to paraphrase Hilary Nesi.[2] The environment has drastically changed and so do need to change the tools.

The impacts of both of these changes are yet to be discovered: for the latter one, the status quo can be well described by quoting another heavyweight in the field, the long-time editor-in-chief of Macmillan dictionaries, Michael Rundell, whom I often heard saying: "After working in this field for 30 years, I thought I had a pretty good idea about how to create and publish a dictionary. But things have changed so dramatically in the last five years, that I have only a limited idea of what the future of lexicography will be."

The impact is a bit easier to be foreseen as regards technological innovations. Contemporary lexicography makes heavy use of corpora and increasingly also of many natural language processing tools that automate the analysis of morphology as well as syntax and semantics. Many tools for (semi-)automation of specific lexicographic tasks have been developed as well. In a review carried out in 2011, Rundell and Kilgarriff argue[3] that while word sense identification and definition writing remain to be tackled, many other tasks alongside the lexicographic workflow have been already solved with an accuracy that delivers time- (and therefore money-) saving solutions. This is deemed to be the case for devising dictionary headword lists, finding collocations and other multiword units, or extracting dictionary examples from corpora.

What is the next step? At the moment lexicographers query corpora (by means of many tools) for finding linguistic evidence in order to draft a dictionary entry which they continue working on and which is subject to a number of reviews in the lexicographic workflow.

The next step is to spare the lexicographers from such initial corpus query and entry drafting. Instead of starting with an "empty" dictionary, they will be able to begin with a dictionary database pre-populated with entries according to a big underlying reference corpus. These entries will contain suggested word sense clustering, with definitions (or explanations in alternative forms such as image media), labels and examples extracted from corpora. These entries will then be edited in an environment that includes direct links to underlying corpus evidence so as to allow manual inspection of the source texts, as well as mechanisms for easy and simple corrections of the entry (e.g. lumping and splitting of word senses, replacing dictionary examples, amending definitions and labels). Having all the evidence at hand, the next step is to leave the "easy" bits to the computer and have human editors spend their time on the more intellectually demanding parts of the job. This opens the way to *Post-Editing Lexicography*, in an analogy to the translation process. Translators used to use many independent tools (dictionaries, in the first place!) up to the moment when machine translation

1 As can be seen from early corpus development projects like COBUILD or BNC, which were both driven and devised (also) for lexicographers, who were themselves employed in empirical linguistic research (cf. Church, Ward and Hanks, 1990. Word association norms, mutual information, and lexicography. *Computational Lnguistics* 16.1, 22-29).

2 See e.g. *The Oxford Handbook of Lexicography*, 2016. Durkin, P. (ed.). Oxford: OUP, 584.

3 Rundell and Kilgarriff, 2011. Automating the creation of dictionaries: Where will it all end? In Meunier et al. (eds.), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: Benjamins, 257-281.

and translation memories became mature enough to be exploited for professional translation, and henceforth translators became post-translation editors.

An important lesson from the translation business concerning potential danger for the future of lexicography is that the transition to post-editing translation was by far not easy, partially because it may have actually begun too soon (pushed by translation and localization agencies pressing to cut costs), having machine translation do yet another "over-promised and under-delivered" U-turn.

Eventually, this adoption has further progressed as the technology became more mature, and, mainly, as the translation environment for professionals has become more suited for the task of post-editing, which is very different from translating from scratch.

However, the episode had a very undesired consequence that we want to avoid in lexicography. Translators were abandoning machine translation, for both technological reasons as well as for fear of becoming jobless. These fears remain, even though the former is being improved and the latter just did not prove to be the case: the translation industry is growing and some reports describe it as one of the fastest growing businesses today.[4] Moreover, as the whole process gets streamlined, there are good chances that lower per-word income of translators will eventually turn into a higher per-hour rate for them.

The lessons for lexicography are straightforward: the transition to post-editing must be backed by solid technology (which we believe we have), revisited workflows (which we need to work on), and with advocacy explaining that it is not meant to steal lexicographic jobs. The shift to *post-editing lexicography* might well be a fertilizer for the falling industry, showing faster (and hence, more affordable) and more effective workflows.

A specific account comes with addressing less-resourced languages – or those that are basically not resourced at all at the moment. There are plenty of them, often geographically located in areas with growing numbers of speakers. Many of these speakers live in poverty, which nevertheless does include the possession of a smartphone. Language resources will be one of the first data needed when these societies will approach information levels of the developed world. There will be a strong business need for them but no time for twenty-years-running lexicographic projects, another reason why human efforts need to be supplemented by automation as far as possible.

At the eLex 2017 conference we are going to present a proof of concept in the form of interconnecting Sketch Engine, a leading corpus query system, with Lexonomy, a new lightweight dictionary writing system.[5] We will show how a dictionary draft can be directly obtained from a reference corpus (as a One-Click Dictionary) in Sketch Engine and how it can be efficiently post-edited in Lexonomy.

The future of lexicography presents big challenges. It would be naïve not to realize that many of them pose real issues, problems and obstacles for all players in the field. However, the more so we need to look for those of them that present real opportunities – and, I believe, *post-editing lexicography* is one of them.

4 See, e.g., https://gala-global.org/industry/industry-facts-and-data.

5 Jakubíček M., Kovář V., Měchura M. and Rychlý P. One-Click Dictionary. In *Electronic lexicography in the 21st century, Proceedings of eLex 2017* (forthcoming).

Sketch Engine (SkE) started in 2004 as an academic and lexicographic product for corpus query and management and has since attracted a wide audience including translators, writers, marketers, brand naming and SEO professionals. To meet the challenge of guiding this variety of users to the functionality they need in a streamlined way, an all-new user interface is under development to not only bring in the latest Web technology but also change the way users interact with SkE. With a soft launch due in autumn 2017, users will enjoy a new friendly design which adapts to small touch screens of tablets and mobile phones. Input forms and selection screens will offer basic and advanced layouts, the former targeted at casual users without a profound knowledge of corpora or NLP and the latter serving academic and professional users. In this process, various controls have shifted to more intuitive spots, enabling the user to, for example, adjust the view on the result screen rather than make this decision beforehand as it is now, while preserving all the options and features that are currently available.
Hand in hand with developing the new look and feel, SkE will become more useful to anyone in need of glossaries or dictionaries. In addition to serving as an indispensable tool for gathering data, the new link with **Lexonomy** system will enable data conversion into a lexicographic product as part of an online dictionary writing tool, which doubles as a hosting service that can produce a dictionary and have it published online instantly. Lexonomy also features an easy-to-use XML editor suitable for users with no prior knowledge to create lexicographic products complying with current standards. Embedding Lexonomy in SkE will become vital for starting brand new lexicographic projects. Users will access a corpus to identify the most frequent words and have the list pushed to Lexonomy along with part of speech tags, usage flags, example sentences, collocations, synonyms, definitions or translation, thus generating a dictionary draft for post-editing. Likewise, a subject-specific glossary can be developed analogically from a terminology list extracted from a domain corpus. This push & pull model will dramatically change the way dictionaries are built, besides its beneficial time and money saving implications.

Ondřej Matuška
Lexical Computing

Phonetic transcription of dotted Hebrew

Alon Itai

Alon Itai is Professor Emeritus of the Computer Science Department of the Technion, Haifa. His research interests include Machine Learning and Natural Language Processing. He is the director of the MILA center for processing Hebrew (http://mila.cs.technion.ac.il/eng/index.html), dedicated to developing tools for processing Hebrew.
itai@cs.technion.ac.il

Abstract
The pronunciation of Israeli Hebrew mostly follows the pronunciation rules of dotted Hebrew script, though there are several systematic deviations. As part of the development process of a new Hebrew lexicographic resource by K Dictionaries, we have constructed and implemented an algorithm to deduce the pronunciation from dotted texts and tested it on a large manually tagged database. The database contains 35,443 dotted Hebrew words and their IPA (International Phonetic Alphabet) transcription. The program succeeded in correctly predicting the pronunciation of over 89% of the words in the database. 78.5% of the errors occurred when predicting the stress of loan words. Most of the remaining errors occurred in predicting the pronunciation of *schwa*. We found out that the traditional phonological explanation that is based on sonority theory correctly predicts 88.3% of all pronunciations of *schwa*. We constructed an alternative algorithm that correctly predicts the pronunciation of *schwa* in 99% of the words of the database.

1. Historical background

Hebrew is among the first languages transcribed by a phonetic alphabet. The original script constitutes an abjad, i.e., a script where the vowels are not represented. However, some consonants served as matres lectionis, namely the consonants ה,ו,י in addition to their role as consonants are used to indicate the vowels, a,u/o and i . During the first centuries A.D. Aramaic replaced Hebrew as the main spoken language of Jewish people in Palestine. However, the Holy Scriptures and mainly the Bible were canonized using the original abjad. Because of the need to correctly read the Bible, a system of diacritics, called dots, was added during the 10th century C.E. These symbols were small, so as to not change the holy texts, and reflected the way the scriptures were pronounced at the time.

All through the Middle Ages, scholars continued to write in unvocalized Hebrew using the matres lectionis more extensively, and this is the standard script of Israeli Hebrew.

With the revival of Hebrew as a spoken language, at the end of the 19th century, the Sephardic pronunciation was adopted. This pronunciation fused several diacritics and Israeli Hebrew further modified the pronunciation. Thus dotted texts are only a rough guide to pronouncing Hebrew, and various systematic deviations exist.

2. Pronunciation Rules

2.1 Consonants. The consonants follow a regular pattern. Table 1 shows a many-to-one map to IPA.

2.2 Vowels. Most vowels follow a many-to-one transcription. The two problematic cases are the diacritic *qamatz*, which is most often pronounced /a/ but sometimes /o/, and the diacritic *schwa*, which is in most cases silent but sometimes pronounced /e/.

2.3 Stress. In most Hebrew words the stress is on the last syllable, though some are penultimate. Even though stress is phonemic, it is not transcribed explicitly in Hebrew dotted script. In nearly all cases one can deduce the stress from the vowel pattern of the word or from its morphological analysis.

Some noun patterns also have penultimate stress. For example, the segolite word pattern class can be easily recognized since the last vowel is the diacritic *segol* (e.g. כֶּלֶב /ˈkelev/ *dog*). There are several related patterns which are easy to identify. Verbs in the past tense end with an unstressed suffix (e.g. כָּתַבְתִּי /kaˈtavti/ *I wrote*). These inflections can be identified by a morphological analyzer.

Loan words pose a greater challenge as their stress does not follow the rules of native words. The stress is most often penultimate and, contrary to native Hebrew words, its position does not change even in the presence of a stressed suffix. Thus, a prerequisite

Heb	א,ע	ב	ב,ו	ג	ד	ה	ז	ח,כ	ט,ת	י	כ,ק	ל	מ	נ	ס,ש	פ	פ	צ	ר	ש
IPA	ʔ	b	v	g	d	h	z	x	t	j	k	l	m	n	s	p	f	t͡s	ʁ	ʃ

Table 1: Transcription of Hebrew consonants

for determining the stress position of a word is to determine whether it is a loan word, and in some cases a morphological or semantic analysis is necessary.

For example, the word *bira* with ultimate stress is a native word meaning *capital* (city) and with penultimate stress is a loan word meaning *beer*. The dotted script renders both words identically. Thus, to correctly determine the stress position one first needs to disambiguate the word, which requires examining the context and performing a sematic analysis.

2.4 Miscellaneous. Some combinations of letters/diacritics do not follow the above rules. For instance, חַ /χa/ at the end of a word is always pronounced /aχ/. כְּח /χχ/ is often, but not always, pronounced /kχ/.

3. The experiment

We constructed an algorithm to transform dotted Hebrew to IPA.

3.1 The database. We tested our algorithm on a database of 36,358 dotted words provided by K Dictionaries. The database was created from the Hebrew dictionary core edited by Orna Ben Natan. Then the project managers, Anat Merdler-Kravitz and Yifat Ben-Moshe reviewed the automatically-transcribed words and amended them as necessary. As a result, each database entry consists of a word in both its dotted transcription and IPA counterpart.

The database consisted of 21,126 lemmas, represented by their base form. In addition there were some plural forms of nouns, verb inflections, and 184 multiword expressions. Many Hebrew conjunctions and prepositions are represented as prefixes in the standard script. Except for the latter, the words did not contain such prefixes.

3.2 Evaluation. The program correctly transcribed 32,736 words which consist of 90% of the database.

The miscellaneous category consists of 9 occurrences of כְּח that were incorrectly transcribed as /kx/, and some occurrences of עָ and הָ that were transcribed as the null character and /h/ instead of /ʔa/ and /ha/.

The main source of errors is the misplacement of stress.

The program correctly identified the stress position of all but 3% of the original Hebrew words in the sample. Loan words are the main source of errors. The program identified some of these words using heuristics, such as the existence of non-native consonants (ʤ, ʒ, ʃ), suffixes (t͡sija, nik,…), words starting with /f/ and other patterns that defy Hebrew phonotactics (a cluster of four consecutive consonants or three consecutive consonants at the beginning of a word). There are some diacritics (*hataf-qamatz*, *hataf-patax* and *hataf-segol*) that appear only in native words.

Total number of errors	stress	*schwa*	*qamatz*	miscellaneous
3,622	3,295	172	122	33
	91.0%	5.2%	3.4%	0.9%

Table 2: Distribution of errors

Total sample	loan words	program error	slang	database error
500	476	15	8	1
100%	95.2%	3%	1.6%	0.2%

Table 3: Distribution of stress errors on a random sample of 500 words

Since in most loan words (though not in all) the stress is penultimate, the program placed the stress there, thus eliminating a potential error.

As described above, the diacritic *Schwa* in Hebrew is sometimes pronounced /e/ and sometimes omitted (in Hebrew the diacritic *schwa* is never pronounced as a phonologist *schwa*). Hebrew phonologists used sonority theory to predict this behavior. Phonologists define sonority as the audible energy omitted with each phoneme (Burquest and Payne 1998, O'Grady and Archibald 2013). In each syllable the sonority rises until reaching the syllable's nucleus (usually a vowel) and then it falls. In English, the sonority scale, from highest to lowest, is the following:

a > e o > i u > r > l > m n ŋ > z v ð > s f θ > b d g > p t k.

Rosen (1957, following Segal), and later Boletzky (2007), postulated that when the onset of the syllable defies this order, i.e., the first phoneme is more sonorous than the second, the syllable is split by inserting the phoneme /e/ between the first and second phonemes. The sonority of the phonemes of the onset of each of the two syllables increases. Thus, for example, since /l/ is more sonorous than /v/, /lvi.ˈva/ becomes /le.vi.ˈva/, i.e., the syllable /lvi/ becomes two syllables /le/ and /vi/, thus causing the sonority of each syllable to increase.

To accommodate for the observed behavior of Israeli Hebrew, Rosen (1957) postulated the following sonority hierarchy for Hebrew:

a > e o > i u > j > l > m n > z v > f x χ > b d g > p t k > ʔ h

Rosen did not place /ʁ/ (r) and the silibants (s and ʃ) in this hierarchy. To properly place phonemes one should check whether an /e/ is inserted before or after the occurrences of the phoneme. Thus, we found that it is best

LOTKS 2017

Workshop on Language, Ontology, Terminology and Knowledge Structures

On September 19th the second edition of the Language, Ontology, Terminology and Knowledge Structures (LOTKS) workshop will take place as a satellite workshop of the 12th International Conference on Computational Semantics (IWCS) in Montpellier, France. Following on from a successful first edition as a joint workshop at LREC 2016, the intention is once again to provide a forum for different research communities to interact and discuss issues within the intersection of computational linguistics, ontology engineering, knowledge modelling and terminologies.

LOTKS grew out of the need for a workshop that dealt, on the one hand, with enhancing knowledge bases or conceptual schemes with linguistic knowledge, as well as on the other, the growing use of ontologies and concept schemes to enrich linguistic or lexical datasets -- in particular computational lexicons.

The workshop also offers showcasing the use of conceptual/terminological/ ontological resources in NLP or computational linguistics in general. This year we have introduced new themes relating to the use of terminology schemes and ontologies in the digital humanities. The workshop welcomes contributions from both academics and industry professionals.

Fahad Khan
Istituto di Linguistica Computazionale (A. Zampolli) – CNR

https://langandonto.github.io/langonto-termiks-2017/

to place /r/ together with /l/. However, since such a split never occurs before or after s, ʃ it is not possible to place the silibants to conform to this rule.

We tested this sonority rule on our database (omitting syllables with silibants in the onset). The theory successfully predicted the omission/inclusion of /e/ in 88.3% of words.

We have developed an alternative algorithm with better performance: When *schwa* immediately follows the first letter it is pronounced /e/ if and only if at least one of the following occurs:

- The first phoneme is a word prefix, such as b (=in) v (=and).
- The first phoneme is a verb conjugation prefix, e.g., tsaˈper te.sa.ˈper, *you will tell* = t (future, 2nd person)+saˈper.
- The first phoneme is j,l,m,n,r.
- The second phoneme is ʔ,h,ʕ .
- If *schwa* occurs elsewhere it is pronounced /e/ if and only if it is:
  - The second *schwa* in the pattern C1 *schwa* C2 *schwa* C3 (Ci a consonant).
  - Between two identical or similar letters (e.g., between /d/ and /t/)

The first two rules require a morphological analyzer to identify the correct analysis of a word in context. Since we did not have at our disposal a morphological analyzer for dotted texts, we could not apply these rules, which could have prevented at least 49 errors. The verb conjugation prefixes with *schwa* are t,j,l,n,m. With the exception of /t/ the prefix has high sonority and should, in most cases, cause a syllable break. Thus the second rule is often subsumed by the third. (This explains the low number of errors when rules 1-2 are ignored.) Since the number of remaining errors was small, we were able to manually identify when rules 1-2 were applicable, thus obtaining an error rate of less than 1%.

Qamatz

The diacritic *qamatz* is most often pronounced /a/ (*big qamatz*). The database contained 199 occurrences where *qamatz* is pronounced /o/ (*small qamatz*). We used two heuristics to identify (some of) them: The *qamatz* was followed by a consonant with the diacritic *hataf-qamatz* (which is always pronounced /o/). Thus, the pattern was /oCo/.

The consonant after the *qamatz* had a *schwa* and the following consonant had a *dagesh* (that indicates germination or strong pronunciation). Thus, the pattern was *qamatz* C1 *schwa* C2 *dagesh*. Hebrew grammar dictates that the *dagesh* is a light *dagesh* and C2 is either ב,ג,ד,כ,פ,ת.

This allowed us to identify 74 cases of /o/ (37.2%). The *small qamatz* is relatively rare, appearing in only 0.6% of all words of the database and in only 3% of the errors.

Conclusions

We have constructed an algorithm to transcribe dotted Hebrew texts to IPA conforming to the observed Israeli Hebrew pronunciation. The algorithm was implemented as a Python 3 program and is available from the author. The program was tested on a large database and the error rate was 11.2%.

We used the database to test how well sonority theory explains the pronunciation of *schwa*, and have formulated a simple alternative algorithm that outperforms the sonority theory algorithm.

References

Bolotzky, S. 2007. The sonority in the phonology of Israeli Hebrew. In *Hebrew and her sisters*, Efrat, M. (ed.). Haifa University Press, 239-248.

Burquest, D. A., and Payne, D. L. 1998. *Phonological analysis: A functional approach*. Dallas, TX: Summer Institute of Linguistics.

O'Grady, W. D., and Archibald, J. 2013. *Contemporary linguistic analysis: An introduction*. (7th ed.). Toronto: Pearson Longman, 70.

Rosen, H. 1957. *Ha-Ivrit Shelanu*, (*Our Hebrew*). Tel Aviv: Am Oved.

	Sonority theory w/o silibants	The alternative algorithm		
		Rules 3-6 w/o silibants	Rules 3-6 with silibants	Rules 1-6 with silibants
Sample size	7449	7449	8612	8612
# errors	871	125	126	77
% error	11.69%	1.68%	1.46%	0.89%

Table 4: Sonority theory and the alternative algorithm for words with ***schwa***

The Saga of *Norsk Ordbok*: A scholarly dictionary for the Norwegian vernacular and the Nynorsk written language

Oddrun Grønvik

The 9th of March 2016 saw the launch of *Norsk Ordbok*, a twelve-volume scholarly dictionary of the Norwegian vernacular and the Nynorsk standard language. *Norsk Ordbok* fills twelve volumes of 9,600 pages, has about 11 million words of text, holds 330,000 entries and ca 15,000 fixed phrases. It took 86 years to complete since the material collecting started and until volume 12 was out. Two thirds of the editing happened after 2000. The dictionary as well as much of the evidence (contained in the Norwegian Language Collections, cf. Grønvik 2020) is freely available on the web (http://www.norskordbok.uio.no).

The full story of a twelve-volume scholarly dictionary could easily fill another volume, but in this article only a few points will be adressed, i.e. (1) the linguistic backdrop, (2) the dictionary project and its source material, (3) the digitisation project NO2014, and (4) the future.

1. The linguistic and historical backdrop to *Norsk Ordbok*

Norway has a broken history; independence until the end of the 14th century, subordination under Denmark until 1814 and under Sweden from 1814 to 1905, and independence again since 1905. These political changes have had a profound influence on the Norwegian language, which in turn has affected the formation of written standards and the scholarly lexicography for Norwegian. The chief result is that Norway today has two written standards, Bokmål and Nynorsk, which are close cognates, and which are each documented in a major dictionary. *Norsk Ordbok* documents the Nynorsk written standard and all Norwegian dialects.

The historical background can be summarised as follows (cf. Haugen 1976; Vikør 1995 p. 51 ff. and 92 ff.):

The spoken languages of medieval Norway, Iceland, Sweden and Denmark must have been mutually comprehensible, but resulted in different written practices. In this period, the written language of Norway was what is now called Old Norse.

Once the administration of Norway was transferred to Denmark, Old Norse was gradually replaced by Danish, until by the end of the sixteenth century Danish became the language for civic administration. Norway was not allowed a university until 1811, so tertiary education meant a long stay in Copenhagen. Danish was the dominant written language in Norway until after 1905.

In the same period, the Norwegian vernacular changed so much that a revival of Old Norse as a written standard after 1814 was unthinkable – the spoken dialects and the old written standard were too far apart.

The 1814 Constitution states that legislation should take place in the Norwegian language, but this was a shield against a Swedish takeover. The choice of Danish as an administrative language nevertheless left Norway with a national legitimacy problem – the idea of a separate Norwegian national identity, and the need for an independent state, was questioned. The language issue became the question of the day from the mid-19th century, and two solutions were presented, though not shaped into opposing camps until the end of the nineteenth century.

The response to the legitimacy issue was initated by members of the Royal Norwegian Society of Sciences and Letters (DKNVS). The society looked actively for someone who could document the Norwegian vernacular language, and prove (a) its connection to Old Norse, and (b) its separateness from Danish and Swedish. Because of the diglossic situation - Norwegian and Danish were close cognates, and Danish spoken in Norway was phonologically adapted to Norwegian - the difficulty was finding a trained linguist who was close enough to ordinary people to gather trustworthy linguistic information. The problem was solved when the the self-taught linguist and lexicographer Ivar Aasen (1813–1896) presented himself for the task. Aasen was funded from 1840 onwards and throughout his lifetime, first by DKNVS, then by Stortinget. Within his lifetime, Aasen documented the grammatical structure and the lexicon of Norwegian in a series of works culminating in *Norsk Grammatik* (1864) and the dictionary *Norsk Ordbog med Dansk Forklaring* (1873). The orthography expressed in the headwords of Aasen's 1873 dictionary was also his proposed standard for a common, wholly Norwegian written standard, the forerunner to today's Nynorsk (New Norwegian). Aasen's work put an end to the legitimacy doubts – Norwegian was a

Oddrun Grønvik obtained a B.A. with honors in English Litterature and Language from St. Hilda's College, Oxford, has the degree Candidate of Philology at University of Oslo, and was in 2006 granted an honorary PhD from the University of Zimbabwe for her work on dictionaries for African languages. She has worked in the publishing industry (1970-1979) and as consultant at the Norwegian Language Council (1979-1987), and in 1987 became Asssistant Professor at the University of Oslo and joined the editorial staff at *Norsk Ordbok*, serving as Chief Editor between 2006-2016 with special responsibility for lexicographical and digital development and training. In addition she has served as academic project manager for the NORAD-funded African Languages Lexical Project, and on the standardisation committee of the Norwegian Language Council. Dr Grønvik has written widely on the Norwegian language, language development and standardisation, lexicography and digitisation of language resources.
oddrungronvik@iln.uio.no

Ivar Aasen:
Norsk Grammatik
§ 384 (1864)

"What especially concerns the word forms, is that the particularities belonging to each of them are so many, that they cannot easily be dealt with otherwhere than in a dictionary."

separate West-Nordic language, descended from Old Norse, while modern Danish and Swedish stem from East-Nordic.

Aasen's work was made possible by the development of the comparative methodology of 18th and 19th century historical philology. He systematically documented the Norwegian dialects, employed the comparative method to establish a common pattern for phonology, morphology, lexicon and syntax, using Old Norse as a touchstone, but including nothing that was undocumented in his time. In shaping his proposed standard language he also took the standards of Swedish and Danish into account, to avoid unnecessary differenciation from what people were used to seeing in print.

To the ruling classes of Norway, however, the idea of even trying to establish a wholly Norwegian written standard, for everyday use in competition with Danish, seemed ridiculous and unthinkable. This would mean giving cultural hegemony to an uneducated, though literate, country population. At the same time, something had to be done to nationalise the Norwegian version of Danish, clearly diverging from the Danish of Denmark. The counter-solution to Nynorsk favored adapting standard Danish to Norwegian phonology and including typically Norwegian words in the lexicon. A gradual transition from a Danish to a Norwegian written standard, based on "educated everyday speech" was envisaged. The first orthographic reform of Danish in Norway came in 1907 and established the forerunner of today's Bokmål.

At the end of a long and fierce political struggle, the Norwegian parliament in 1885 voted to give both standard languages official standing as languages of instruction and leave it to each school board to choose which one to adopt. An earlier parliamentary decision, in 1874, had tasked teachers with adapting their oral instruction in class to the dialect of their pupils, instead of the other way round. Since then, Norwegian has been expressed in two written languages. Since 1929 these have been termed Nynorsk and Bokmål (Vikør 1995; Hovdhaugen 2000).

2 The *Norsk Ordbok* dictionary project and its source material

Plans exist back to 1911 for scholarly lexicography for Norwegian, in the form of committee reports and applications for funding. From 1920 onwards some funding was achieved for starting language collections -– slip archives with indexed excerpts, according to the best practice of the times. The resulting lexicographical work was envisaged both as a joint presentation of all spoken and written Norwegian, and as a separate work for each of the standards. Separate scholarly dictionaries became the solution and resulted in the parallell projects of *Norsk Riksmålsordbok* and *Norsk Ordbok - Ordbok for det norske folkemålet og det nynorske skriftmålet*. *Norsk Riksmålsordbok* was published in four volumes 1928-1958, with a supplement in two volumes published in 1995.

The split into two projects had an ideological basis with inevitable practical implications. The scholarly dictionary for the Danish-derived standard was to be based on printed literature of Norwegian authorship from 1814 onwards, supplied by speech materials representing "educated everyday speech". The scholarly dictionary for Nynorsk was to be based on the Norwegian vernacular in all its varieties, as documented back to about 1600, and on Nynorsk literature, which came into being from the late 19th century onwards. The first project regarded speech as a supplementary category and a dialect label as a warning; the other one saw the dialect materials as primary source material, expanded and developed through literary use. Sources as well as the lexicographical treatment of them were to be too different for the projects to be compatible within one framework.

Norsk Ordbok got off to a belated start in 1930. A grand plan for material collection, with a small editorial staff supported by volunteers, was drawn up and drew a gratifying response: 600-700 volunteers came forward within a year or two, and by 1940 the collections encompassed one million slips, 20 percent documenting speech, the rest drawn from written sources. At the same time, a rough first version was drafted on the basis of existing Nynorsk dictionaries and some large dialect

collections. This draft manuscript held 130,000 entries and covered 13,500 pages of typescript. The plan was to expand this manuscript with the materials from the language collections and end with a modern dictionary of 4-5 volumes.

Work on *Norsk Ordbok* was halted during the second World War, and started again in 1946. A review period led to the following three decisions: (a) continue collecting, especially oral materials; (b) draw up a detailed plan for the dictionary microstructure, so as to do justice especially to the richness of the sources; (c) start editing all over again, focussing on full use of the Language Collections with the draft manuscript as a guideline. On the basis of this very ambitious plan, the first fascicle was completed in 1950 and the first volume in 1966. At this time, the completed dictionary was thought to reach eight volumes at the most.

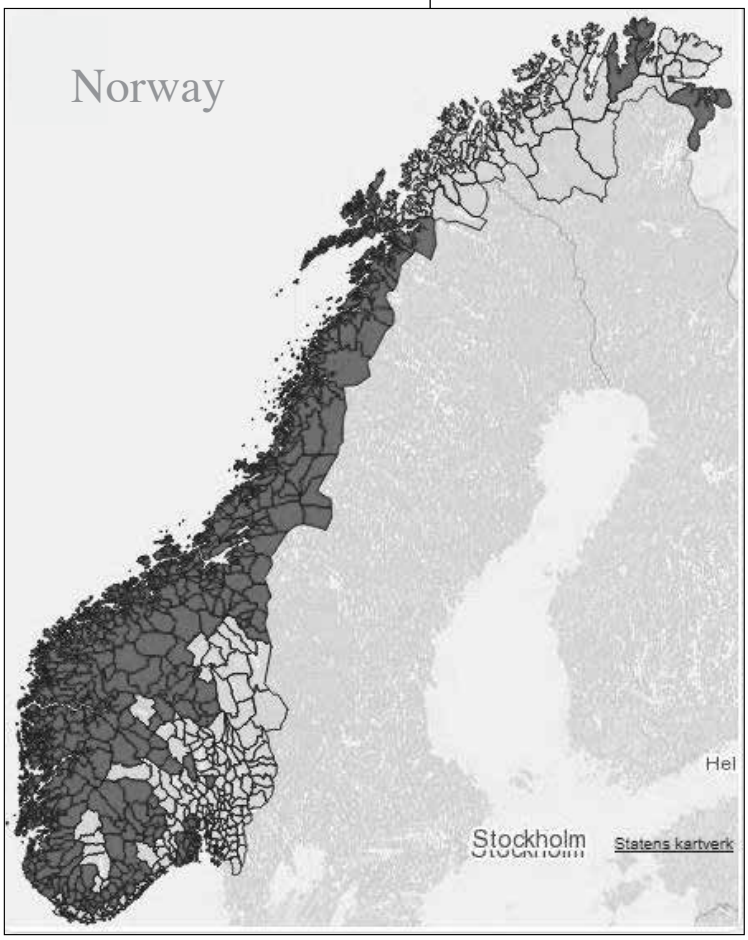
The group of editors increased slowly. When I was recruited in 1987, I became the eigth editor and the second woman editor. At that time two volumes were out and the third completed in manuscript. The Language Collections had quadrupled in size and the alphabet progress had slowed down. A little arithmetic showed that if work continued at the rate then current, *Norsk Ordbok* would be completed around 2060 and reach 16-20 volumes -– a plan which was unlikely to get funding. These facts were therefore kept quiet.

Norsk Ordbok needed a miracle, and the miracle turned up in the shape of a huge digitisation project for the university collections of the whole of Norway, designed to counteract nationwide unemployment when the Norwegian telephone and telegraph services went digital. Through the Documentation Project (1991-1997), run by Christian-Emil Smith Ore at the University of Oslo, key components of the national Language Collections were stored in databases and on the Web by 1997, i.e. the excerpt archives, the draft manuscript of 1940 and a number of other resources. All components were then coordinated in a digital index – the Meta Dictionary – with base forms and part of speech as in *Norsk Ordbok*. The Meta Dictionary at present holds about 550, 000 entries for Nynorsk.

3 The Digitisation Project NO2014 (2001–2016)

As the millenium approached, Norwegian authorities were planning another jubilee -– the bicentenary of the Norwegian constitution in 2014. *Norsk Ordbok* was chosen as one of the bicentenary projects, on the basis of a carefully worked out production plan. The basis for this plan was the conviction that trained linguists could become efficient scholarly lexicographers within one year, and after that meet production deadlines as planned. In order to succeed, *Norsk Ordbok* would need roughly 265 man years of efficient lexicography within 14 years -– 2001 to 2014. We thought we could do that, given (a) competent and tough management, (b) scholarly computer developers, (c) enough linguists, and (d) funding. All of these factors were equally important, and the planned project, named NO2014, would be doing a tightrope act from beginning to end. It was worth trying.

The conviction that scholarly lexicographers could be trained quickly and efficiently ran counter to traditional views of training needs for scholarly lexicographers. When I started in 1987, the general assumption was that training as a scholarly lexicographer would take at least five years, and the time would be spent in getting to know the collections, mastering a multitude of conventions, and accepting the need for extensive crosschecking and proofreading. Previous experience as a linguist would certainly be utilised, developed and challenged in handling very complex materials, but language analysis was only one of many tasks, and they all



The entry *støl* in the online version of *Norsk Ordbok*

seemed to hold equal significance.

Norsk Ordbok was one of many projects to embrace the computer at the end of the 20th century. The decisive experience for us on the issue of training enough lexicographers in a fairly short time, was participation in the ALLEX Project (1991-2006), a Norwegian and Swedish funded research project designed to provide monolingual dictionaries for the African languages of Zimbabwe. The ALLEX Project proved that mother-tongue linguists could become efficient scholarly lexicographers in a very short time, working through a lexicographical interface based on an analysis of the relevant language, storing results in databases, dealing with oral materials in corpora, etc.

Categorising and commenting on language through well worked-out software is not only a tool for efficiency, it is also an immensely effective tool for learning and mastery. This conviction combined with the assurance of computerisation support from EDD (the Unit for Digital Documentation at the University of Oslo), covered points (b) and (c) above. The Norwegian Parliament guaranteed point (d), with funding through the Ministry of Culture and the University of Oslo. NO2014 also had the immense good fortune to attract two directors[1], one after the other, who had all the qualifications one could wish for in professional and human terms. The positions as chief editors – a group of four -– were held by former staff members. Tasks were allocated according to project needs. I was made responsible for digitalisation and training, and this account is naturally coloured by my particular experience.

A premise for funding was moving the entire project to a digital platform. We took that to mean not only producing the dictionary itself, but also being able to access and sort digitalised materials, take care of the sorting, make sure that no entry lacked materials, and saving the finished product in a form that allowed different types of presentation of the finished product (Grønvik 2005).

An important decision concerning the software structure was to make a maximum format the standard, always allowing for the most extensive and complex entry, rather than having a more restricted basic format which might have to be extended. The standard sense unit therefore caters for definition, usage examples, sub-definitions with usage examples, multiword expressions with several senses, and finally compounds in which the entry headword appears as the initial or the final part, plus of course source tables for literature and geographical location.

The editorial interface was the first thing to get finished. By 2013, *Norsk Ordbok* in digital form was contained within one application which was able to (1) generate

1. The Project directors of *Norsk Ordbok* 2014 were Kristin Bakken 2002-2008 and Åse Wetås 2008-2015.

entries from indexed materials, with a link to the materials, (2) provide a tool for analysing linked materials and storing the analysis, (3) generate the entry head (the identifying information of the entry) from a separate full form register, (4) present a form where the edited text can be linked directly to the materials underlying each definition or description, (5) allow supervision of production flow at the micro and macro levels, (6) present the finished product in an optimally accessible fashion (paper: pdf with preset style sheet corresponding to the print typeface; web: settings for web presentation on different reading tools), and (7) provide a format for longtime storage of the product linked to its sources. The software package became a sort of lexicographical factory, designed to allow editors to concentrate on analysing and editing (Grønvik and Ore 2013).

The *Norsk Ordbok* software is now used in three other dictionaries, *Bokmålsordboka* and *Nynorskordboka* being the best known (see http://ordbok.uib.no/, cf. http://dictionaryportal.eu).

When the project NO2014 started in 2001, the alphabet stretch a-h was already edited, with great care and consistency, and deep respect for the materials contributed over the years, especially the oral materials. A primary task in the digitalisation process was to take care of what can be termed best practice in the pre-digital editorial work. Two tasks stood out: (a) the careful identification of formerly unstandardised dialect word forms, and finding for them a standard form consistent with modern Nynorsk orthography; (b) the treatment of multiword expressions (MWEs), which proved to have been a major difficulty in the pre-digital entry schema. The first task became a permanent concern for the project management, especially in offering all new editors training in handling Norwegian and Nordic dialectology, synchronically and diachronically, but also in giving particular attention to the standardising of new dialect materials which were added to the Language Collections during the project period. For the second task, training in identifying MWEs was offered on a permanent basis, but we also created a software template for the registration and editorial handling of MWEs, making them directly searchable.

Identifying both poorly documented word forms and MWEs was greatly helped by important additions to the digital Language Collections. The most important items were a corpus for Nynorsk literature covering the period 1866–2010, now at 105 million tokens (Nynorskkorpuset), and the digitalization of 65 dialect dictionaries (Norway has more than 400 dialects) in a common portal allowing cross-searches in headwords.

A Web edition of *Norsk Ordbok* was launched in 2012 (http://norskordbok.uio.no/), showing the section of the dictionary edited and completed in the relational database, today from "i" to "å" in the alphabet. From January 2014, the Web edition has been linked to a digital map of Norway, and thus been able to show geographical usage extent for word forms, senses and expressions. This edition is popular, and so is the map function.

When the NO2014 project was nearing completion, we also published our editorial handbook through the NO2014 website (*Redigeringshandboka* 2016). The project parole throughout was to encourage users to look behind the edited text into the materials of the Language Collections, raise questions and demand response. In public interaction, publishing the guidelines, which have the role of a method chapter in a dissertation, has turned out to be useful.

When *Norsk Ordbok* was completed, more than 200,000 headwords that had never been lexicographically treated, had an entry in a scholarly dictionary, while the central vocabulary of Norwegian had received in-depth treatment on the basis of written and oral materials covering the whole country and four centuries of documentation (Grønvik 2017).

A fully sourced account of the history of *Norsk Ordbok* (in Norwegian) will be found in the Festschrift published together with the final volume (Karlsen et al. 2016).

Launch of *Norsk Ordbok*, 9 March 2016

4 From the University of Oslo to a new life at the University of Bergen

In June 2014, the University of Oslo decided to end its commitment to Norwegian lexicography, and get rid of the Language Collections, which comprise archives going back to the 1880's and covering far more than the Nynorsk and dialect sections. NO2014 was half way through editing volume 12 when the project staff was sacked. Despite great difficulties, *Norsk Ordbok* did get finished in good order, but there was a delay of more than a year; volume 12 was sent to the publisher, Det Norske Samlaget, on November 24, 2015, and the launch came in March 2016.

By that time, the Norwegian government, through the Ministries of Education and Culture, had decided that the Language Collections, with the dictionaries *Norsk Ordbok*, *Bokmålsordboka* and *Nynorskordboka*, represent essential linguistic infrastructure, and therefore were too important to be left to the management of the University of Oslo alone.

Provided that funding was allocated, the University of Bergen had volunteered to house the Language Collections (comprising collections for Bokmål, Nynorsk, Old Norse and place names). After inventorying in the winter of 2015-2016, more than 70 tons of books and archives were moved in the summer of 2016. The transfer of the digital collections started about the same time, the first components being run from Bergen from September 2016. The total move is a very extensive operation still in process, involving recruiting and training of research, ICT and administrative staff. In February this year an application for revision and full digitisation of *Norsk Ordbok* a-h was submitted by the University of Bergen to the Ministry of Culture, and it was – so far and fingers crossed – well received (for revision plans, see Berg-Olsen et al. 2015).

This is how matters stand.

The cost of completing *Norsk Ordbok* through the Project NO2014 (2001-2016) stands at 260 million NOK, somewhere around 27.5 million Euro. This sounds like a lot of money, though it wouldn't buy many kilometers of road. However, it is enough not to be thrown away lightly, especially when there is visible and vocal public support for maintaining both the Language Collections and the dictionaries. In the future, Norwegian lexicographers will have to continue to serve the public, both in Norway and internationally, through developing Norwegian lexicography as best they can. At least we now know that we can do it!

References

Berg-Olsen, S. and Wetås, Å. 2015. Revision and digitisation of the early volumes of *Norsk Ordbok*: Lexicographical challenges. In Abel, A., Vettori, C., and Ralli, N. (eds.), *Proceedings of the 16th EURALEX International Congress*. Bolzano. EURAC research, 1075-1086.

Grønvik, O. 2005. *Norsk Ordbok* 2014 from manuscript to database – Standard Gains and Growing Pains. In Ferenc, K., Kiss, G., and Pajzs, J. (eds.), *Papers in Computational Lexicography Complex 2005*. Linguistics Institute, Hungarian Academy of Science. 60-70

Grønvik, O. 2020. The lexicography of Norwegian. In Hanks, P. and Schruyver, G.-M. (eds.), *International Handbook of Modern Lexis and Lexicography*. Berlin and Heidelberg: Springer-Verlag. Forthcoming.

Grønvik, O. and Ore, C.-E. S. 2013. What should the electronic dictionary do for you – and how? In Kosem, I. et al. (eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Eesti Keele Instituut, 243-260.

Haugen, E. I. 1976. *The Scandinavian languages: An introduction to their history*. Cambridge, MA: Harvard University Press.

Hovdhaugen, E. 2000. Normative studies in the Scandinavian countries. In Auroux, S. et al. (eds.), *History of the Language Sciences*. Berlin and New York: Walter de Gruyter I-II, 888-893

Karlsen, H. U., Vikør, L. S., and Wetås, Å. 2016. *Livet er æve, og evig er ordet*. Oslo: Det Norske Samlaget.

Vikør, L. S. 1995. *The Nordic languages. their status and interrelations, 2nd ed*. Oslo: Novus.

URLs

Bokmålsordboka and *Nynorskordboka*: http://ordbok.uib.no

The Documentation Project (1991-1997): http://dokpro.uio.no/engelsk/drawer_to_screen1.html

ENeL Dictionary Portal: http://dictionaryportal.eu/en/

Norsk Ordbok: http://no2014.uio.no/perl/ordbok/no2014.cgi

Nynorskkorpuset: http://no2014.uio.no/korpuset/conc_enkeltsok.htm

Redigeringshandboka 2016: http://no2014.uio.no/eNo/tekst/redigeringshandboka/redigeringshandboka.pdf

The First Century of English Monolingual Lexicography. Kusujiro Miyoshi

In a series of case studies ranging across English lexicography of the seventeenth century, Kusujiro Miyoshi calls the tenets of forensic dictionary analysis into action and proves its methodological productivity. Miyoshi's cast of characters is mostly familiar to historians of lexicography: Robert Cawdrey's *Table Alphabeticall* (1604), John Bullokar's *English Expositor* (1616), Henry Cockeram's *English Dictionarie* (1623), Thomas Blount's *Glossographia* (1656), Edward Phillips' *New World of English Words* (1658), Elisha Coles' *English Dictionary* (1676), and J. K.'s *New English Dictionary* (1702). The outlier is Richard Hogarth's *Gazophylacium Anglicanum* (1689), about which most of us knew next to nothing until we read Miyoshi's article about it in *Kernerman Dictionary News* (2008). De Witt Starnes and Gertrude Noyes addressed it in their classic study, *The English Dictionary from Cawdrey to Johnson 1604-1755* (1991; originally 1946), for just five pages, half of which discuss the *Gazophylacium*'s antecedents. If you really want to know about it, you'll necessarily turn to Miyoshi's treatment of it in his new book.

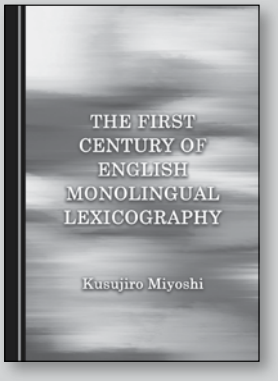
Miyoshi has long been a practitioner of forensic dictionary analysis. Julie Coleman and Sarah Ogilvie (2009) codified its "principles and practice" and included Miyoshi's *Johnson's and Webster's Verbal Examples* (2007) among their references. They proposed that one cannot trust lexicographers to tell the truth about those dictionaries in their prefaces. One can only ascertain the truth by digging elbow deep into dictionary data and analyzing them statistically. As Coleman and Ogilvie conclude (2009, 18), "Forensic dictionary analysis brings together statistical, textual and contextual approaches that allow dictionary researchers to examine, understand, and reconstruct lexicographic practices and policies." Miyoshi's investigations of seventeenth-century English dictionaries are statistical and textual by default — in most cases, we lack useful contextual evidence — and he is a master of the method.

As Miyoshi puts it, his "method, although simple, yields results: it is to dive directly into the contents of the dictionaries" in question, "relying little on descriptions in their title pages and introductory materials," which "tends to reveal the gulf between the front matter of dictionaries and their actual content" (xiv). Thus, his method is "greatly different from, or nearly diametrically opposed to, that of Starnes and Noyes," which "[leaves] a plenitude of historically significant facts undiscovered" (xii). Miyoshi sets out to discover those facts, some of them illuminating about relationships among dictionaries of the period, some about "the highly creative use of other dictionaries in one specific dictionary" (xiv), and altogether registering diverse approaches to representing lexical knowledge, through which later lexicographers would sort to identify and develop what we might call lexicographical best practices.

Chapter for chapter, Miyoshi tells us things of interest about these early English dictionaries, things we ought to know before proceeding to any advanced or speculative arguments about them. For instance, in his first chapter, Miyoshi shows that, although conventional wisdom says otherwise, Cawdrey's *A Table Alphabeticall* and Bullokar's *English Expositor* are structurally and textually closely related. Testing the overlap between the two across the alphabetical range I, L–P, and R–T, he discovers that the *Expositor* appropriates on average nearly 60% of the *Table*'s headwords and that nearly 37% of the *Expositor* derives from the *Table*. Beyond headwords, Miyoshi detects the *Table*'s influence on more than 25% of the *Expositor*'s definitions. At the very earliest stage, then, English lexicographers were aware of one another's work and built upon foundations laid by others.

Chapter 2 contrasts the *Expositor* and Cockeram's *English Dictionarie* on the treatment of derivatives. Miyoshi demonstrates clearly that while Cawdrey did not itemize derivatives, Bullokar — often further developing Cawdrey's entries — added some. Cockeram lists even more derivatives than Bullokar, sometimes where Bullokar lists none, and sometimes adding them to Bullokar's text. So, Cawdrey lists **liquid**; Bullokar lists **Liquid**, **Liquefaction**, and **Liquifie**; and Cockeram lists **Liquid**, **Liquable**, **Liquation**, **Liquator**, **Liquefaction**, and **Liquifie** (15).

Of course, we care whether the words in early English dictionaries reflected use. Even if some of these words were not in use — "Liquator. He which



The First Century of English Monolingual Lexicography

Kusujiro Miyoshi

Newcastle upon Tyne: Cambridge Scholars Publishing, 2017

Hardback, 180 pages, £58.99

ISBN 978-1-4438-5181-7

http://cambridgescholars.com/the-first-century-of-english-monolingual-lexicography

eLex 2017
Lexicography from Scratch

This year marks the fifth anniversary of the biennial Electronic Lexicography in the 21st Century conference series. The conference will take place in Leiden from 19 to 21 September and will be hosted by the Dutch Language Institute.

The theme of eLex 2017 is Lexicography from Scratch and the focus is on state-of-the-art technologies and methods for automating the creation of dictionaries. Over the past two decades, advances in NLP techniques have enabled the automatic extraction of different kinds of lexicographic information from corpora and other (digital) resources. As a result, key lexicographic tasks, such as finding collocations, definitions, example sentences or translations, are being increasingly transferred from humans to machines. Automating the dictionary creation process is highly relevant, especially for under-resourced languages, where dictionaries need to be compiled from scratch and where the users cannot wait for years, often decades, for the dictionary to be "completed".

This year we have received nearly 50% more submissions in comparison with the previous conferences. The Programme Committee has made a nice selection of papers for presentations, demos and posters. Each submission was reviewed by at least two members of the 69-member Scientific Committee.

Keynote lectures will be delivered by Frieda Steurs (Dutch Language Institute), Ivan Titov (University of Amsterdam/University of

melteth" — their inclusion can reveal something about language attitudes and lexicographical technique. Suppose that Cockeram did make some words up to see what an extended list of derivatives looked like — false evidence of English, but an important experiment in dictionary structure and, as Miyoshi points out, evidence that the notion of derivatives had taken hold of the linguistic imagination in early seventeenth-century England. Within Cockeram's entry, **Liquefaction** is defined as "That Liquation is," and the cross-reference, Miyoshi argues, may represent "Cockeram's attempt to present […] entries in a systematic way" (15), which I would emend to "increasingly complex entries." In the early English dictionaries, we see both macro- and microstructural features we now take for granted in the process of their invention.

The history of English dictionaries tends to mention Cockeram in passing. Miyoshi clearly sees him as perhaps the central figure in the development of English lexicography during the seventeenth century. Besides the treatment of derivatives in Chapter 2, Miyoshi considers his treatment of high-frequency verbs (Chapter 3), source material (Chapter 4), and entry structure (Chapter 5), and then contrasts his Anglicization of foreign words to that of Blount in *Glossographia* (Chapter 6). *The First Century of English Monolingual Lexicography* is a slim book: there are 38 pages of introductory material, including an elegant introduction by John Considine, and the chapters by Miyoshi comprise but 130 pages, not counting references and index. Half of the ten chapters and nearly half the pages — 62 of them — focus on Cockeram's work. It's the most detailed and concentrated analysis of Cockeram's *English Dictionarie* I've ever read — for that reason alone, the book is a valuable addition to the historical literature about early dictionaries.

In Chapter 3, Miyoshi investigates another aspect of Cockeram's "system," the treatment of phrasal verbs, which as an element of dictionary structure resembles and aligns well with treatment of noun derivatives — looking across lexical categories, one detects an inclination towards elaboration that would drive later lexicographical innovation until its practices were well established in dictionaries by John Kersey, Nathan Bailey, and of course Samuel Johnson. In Chapter 4, Miyoshi argues quite persuasively that the tradition of English–Latin dictionaries we have long assumed — under Starnes' and Noyes' influence — underlies the second part of Cockeram's dictionary, does not in fact. In the second part, entries open with colloquial terms then defined with Latinate or, as Miyoshi calls them, "refined" terms, the reverse of the "hard word" entry pattern. Because the defining terms are Latin or Latinate, it's easy to assume some connection to the vernacular–classical bilingual dictionaries that precede publication of *The English Dictionarie,* but Miyoshi's collations reveal that approximately 90% of the "refined" words come from the first part of Cockeram's *English Dictionarie*, Cawdrey's *Table*, or Bullokar's *Expositor*. With characteristic understatement, Miyoshi suggests, "We may now have to reconsider the influence of the English–Latin dictionary on the early English monolingual dictionary" (41).

If the second part of Cockeram's dictionary borrows so much material from Cawdrey and Bullokar, then what's new and interesting about it? Miyoshi explains in Chapter 5 that, as in the first part, Cockeram developed a more complex entry structure than those of his contemporaries — especially in presenting synonyms and information on word formation — such that the second part "is highly significant as a dictionary of its time" that contained "the precursors of techniques which are indispensable for the development of English monolingual dictionaries after it" (49). In Chapter 6 he wraps up his inquiry into Cockeram by comparing his approach to Anglicizing foreign words, especially Latin or Latinate ones — **litispendence**, for example — to Blount's in *Glossographia* and concluding that both lexicographers blotted English with inkhorn terms.

The book under review is a series of case studies operating a certain methodology; it concludes only that close attention to the data of seventeenth-century English dictionary texts leads us to re-evaluate relationships among them — both the data and the dictionaries, I suppose. Miyoshi's argument about Cockeram opens into a sort of teleological arc of lexicographical development in the period. Cockeram experiments with systematic approaches to the lexicon in dictionary form, we're told. Cockeram leads us to Chapter 7, titled "Edward Phillips's *New World of English Words* (1658): The First Systematic Treatment of English Vocabulary." Whereas, Blount, for instance, "still saw naturalized foreign words as the primary object of lexicography […] Phillips was coming to realize that what matters is the systematic treatment of the vocabulary of English, whatever its origins" (84), which is a necessary step towards the lexicographical professionalism of

Elisha Coles and John Kersey, who bring the seventeenth century to a close.

Elisha Coles's *English Dictionary*, the next chapter's focus, would summarize and harmonize all lexicographical developments of the century, looking backwards over his predecessors and extending lexicographical system to the internal linking of entries — Coles took the dictionary as a book of miscellaneous entries and made it whole. In John Considine's formulation, "In English and Latin, his work was a milestone in the establishment of the genre of the compactly printed, fully alphabetized classroom dictionary which draws on larger and more learned dictionaries. I think it would be possible to argue that he was one of the founders of that genre, although of course that is a simplification" (2012, 53). It is, rather, a simplification and a truth at the same time.

Chapters 9 and 10 rightly conclude the seventeenth century as it tips into the eighteenth, with John Kersey's *New English Dictionary* (1702), which pivots away from the hard words tradition towards the modern dictionary. Chapter 9 argues, on close comparison of Hogarth's *Gazophylacium* with the *New English Dictionary*, that the former influenced the latter, so that while innovative, the *New English Dictionary* was not independent of earlier dictionaries, not quite as new as the title promised. Chapter 10 suggests that Kersey's primary innovation is the careful treatment of compound adjectives and nouns, a further development of Cockeram's interest in morphological complexity, so, Miyoshi believes, tied to the seventeenth century more than leading into the eighteenth.

Each of Miyoshi's chapters looks forensically into a very precise matter of dictionary structure in one or two dictionaries and each has its illuminating moment. But their narrowness is also a limiting factor and sometimes we are misled, much as Miyoshi rightly claims we can be misled by Starnes and Noyes. So, I accept Miyoshi's point about the *Gazophylacium's* influence on Kersey, and I agree that the *New English Dictionary* is textually a seventeenth-century specimen, but the textual matters aren't the only salient aspects of that dictionary. It leads into the eighteenth century, as Allen Walker Read observed, because Kersey was "the first professional lexicographer" (2003, 223). He saw the purpose and art of lexicography differently from his schoolmaster predecessors — the "hard words dictionary" tradition might as aptly be described the "schoolmaster dictionary" tradition — and in that respect he looks forward to Bailey and Johnson, however many lemmata he carried over from the *Gazophylacium* or any other dictionary.

Taken by themselves, then, Miyoshi's chapters, though well connected to one another, lack essential context. Fortunately, Considine's introduction outlines both "The seventeenth-century monolingual English dictionary tradition" (xxiii–xxiv) and "Studying the tradition: before and after Starnes and Noyes" (xxiv–xxviii), and explains Miyoshi's critical intervention in those traditions and the ways in which his work complements and improves upon Starnes and Noyes (xxviii–xxxvii). Considine's knowledge of the subject is deep and wide, but the introduction is brief and appealing — the sort to which only genuine erudition can lead. Miyoshi proves that seventeenth-century dictionaries are textually much more interrelated than we had realized and reiterates what we've known for a while, that lexicographers of the time rather freely borrowed from one another. But why presume originality when the best possible definition has already been written? I find repeating Considine's conclusion similarly irresistible: "*The English Dictionary from Cawdrey to Johnson* will continue, for the time being, to be the authority of first recourse, but after reading what it has to say on a given topic, it will always be wise to ask, 'does Miyoshi have anything to say about that?' and to turn to this book" (xxxvii). Just so.

References

Coleman J. and S. Ogilvie. 2009. Forensic Dictionary Analysis: Principles and Practice. *International Journal of Lexicography* 22.2, 1-22.

Considine J. 2012. Elisha Coles in Context. *Dictionaries* XXXIII, 42-47.

Miyoshi K. 2007. *Johnson's and Webster's Verbal Examples, with Special Reference to Exemplifying Usage in Dictionary Entries*. Lexicographica Series Maior 132. Tübingen: Niemeyer.

Miyoshi K. 2008. *Gazophylacium Anglicanum* (1689), a Turning Point in the History of the General English Dictionary. *Kernerman Dictionary News* 16, 4-8.

Read A. W. 2003. *The Beginnings of English Lexicography*, (ed.), Adams, M. *Dictionaries* XXIV, 187-226.

Starnes D. T. and G. E. Noyes. 1991. *The English Dictionary from Cawdrey to Johnson 1604-1755*. Amsterdam Studies in the Theory and History of Linguistic Science 57, (ed.), Stein, G. Amsterdam and Philadelphia: John Benjamins.

Michael Adams

Edinburgh), Jane Solomon (Dictionary.com), and Ben Zimmer (Wall Street Journal, formerly Vocabulary.com). One more keynote will be held for the first time as part of the Adam Kilgarriff Lecture, in memory of our colleague and friend Adam Kilgarriff, by Paweł Rutkowski (University of Warsaw), winner of the Adam Kilgarriff Prize 2017.

The programme includes traditionally three parallel sessions, software demonstrations, poster sessions, a book and software exhibition, and a social event. On Wednesday 20 September, particpants will gather in beach pavilion Paal 14 on the Dutch coast for a relaxing evening with BBQ while taking in the stunning views of the North Sea.

The conference is preceded by the Final Meeting of COST Action IS1305 European Network of e-Lexicography (http://elexicography.eu/), which will take place on Monday 18 September at the same venue. Two workshops are scheduled on Thursday afternoon, immediately after the conference: a Sketch Engine workshop, where participants can learn about the new interface and upgrade their corpus skills, and a K Dictionaries workshop, on developing and handling human- and machine-driven lexicographic resources.

For more information on the conference, visit the revamped eLex website at https://elex.link/elex2017/.

We hope that you will join us at eLex 2017 and we look forward to welcoming you in Leiden.

Carole Tiberius

on behalf of the eLex 2017 Organising Committee.

Met zoveel woorden. Gids voor trefzeker taalgebruik. Rik Schutz en Ludo Permentier

Met zoveel woorden. Gids voor trefzeker taalgebruik

Rik Schutz en Ludo Permentier

Amsterdam: Amsterdam University Press, 2016
Paperback, 603 pages, €29.95
ISBN 9789462981805
http://en.aup.nl/books/9789462981805-met-zoveel-woorden.html

In 2016 the language lovers and (Dutch) language experts Rik Schutz and Ludo Permentier published *Met zoveel woorden* (Mzw), which may be translated as 'With so many words.' It may be useful to begin with explaining what Mzw is not. Well, it is not a dictionary, though the authors on the inside front cover refer to the book's dictionary section. They argue on p. IX that Mzw differs from traditional dictionaries. So, it's not a traditional dictionary. Nor is it a textbook aimed at extending the users' knowledge of Dutch (VII). The authors explicitly state that they do not pretend to offer their readers anything that they did not already know. Most of the expressions and phrases included in their book are known to most native speakers of Dutch, they say. Mzw does not include proverbs, because they hardly occur in normal Dutch texts. Mzw deliberately does not contain synonyms, the argument being that good collections of synonyms are already available.

Alright then, Mzw is not a (traditional) dictionary, it's not a textbook, and it's not a dictionary of proverbs, or of synonyms. So, what is it? The book's subtitle gives an indication of what the book wants to achieve. It is meant to be a "guide to well-chosen language use". Mzw wants to help its users to improve their Dutch writing. It would seem to aim at a general public of people who want to write in Dutch, to translate into Dutch, or simply to have a good time reading about Dutch (VII). The authors grouped together the wealth of expressions and phrases that are available to express oneself in Dutch in well-chosen, apt terms. The book provides the user with a useful and often illuminating overview of expressions and phrases that belong to a certain concept. Here we see the difference between Mzw and traditional dictionaries. If Mzw were a dictionary, it would be an onomasiological one. The authors seek to assist writers and translators by offering them ways to enforce concepts and to intensify their meaning as well as ways to aptly word a concept. To meet their goals, they provide 2,800 intensifications and 5,700 idiomatic expressions. Next to the dictionary part, there are three registers that should help the user find what he or she is looking for.

The book offers its users what the authors call a *grabbelton*, a lucky bag, or a grab bag, from which the reader may take what fits best in their text (VII). The authors make an appeal to the users' own linguistic feeling and their knowledge of the expressions found. Their appeal would seem to narrow down the target group considerably. You have to be an educated speaker of Dutch to fully appreciate Mzw. Non-native translators into Dutch must have a fairly high command of Dutch. For the latter group Mzw may be a textbook after all. The rich collection of expressions given in the book must be particularly tempting for them.

In the introduction Schutz and Permentier offer detailed instructions of how their book should be used. In addition, the inside front and back covers have a brief outline of the instructions with three sample questions divided into two or three steps. Assuming that most users will refer to the outline, I will put it to the test. The test will show how the book seeks to be a guide to well-chosen language use.

The first search question the authors give is 'how can you express a concept in an evocative way?' Reading all the time about a certain administration, the concept *beïnvloeding* 'influencing' springs to mind. Step 1 makes it necessary for me to figure out the most common word for that concept. Note that this step is completely intuitive and thus subjective. I am afraid that it will often require a more than average command of Dutch. If there is a more common word for *beïnvloeding*, I cannot think of it. Maybe it is the most common word then? Time for step 2, which involves looking up the word in the dictionary section. *Beïnvloeding* is not listed, however, nor is the verb *beïnvloeden* 'influence'. Okay, let's think again, *manipulatie* 'manipulation' is not the most common word for the concept *beïnvloeding*, I would say, but it comes close. *Manipulatie* appears to be not listed either, but *manipuleren* 'manipulate' is. Under *manipuleren* we find the expression *iets naar je hand zetten* 'force/bend something to one's will.' That is an expression one could certainly use to liven up a text. The dictionary section provides for every expression a most welcome example sentence that should make clear in what context it might be used. Other than from the Internet, the authors do not provide the exact sources of their examples. The example given here is probably indeed from the Internet, and

maybe even quoted from a text on a certain administration:

> *De achterliggende visie is dat de wereld er is voor de mens, dat het tot de basisrechten van de mens behoort de wereld naar zijn hand te zetten.*
> 'The underlying vision is that the world is there for man, that it belongs to the basic human rights to force the world to one's will.'

The second search question is: 'how can you express the meaning of a word more strongly?' Step 1 merely involves looking up the word you would like to intensify in the dictionary section. Let's see what we find under *bekend* 'known, familiar' (that is step 2). Above a dotted line we find the intensifications *overbekend*, 'very well-known, widely known', *welbekend* 'well-known, familiar', and *wijd en zijd bekend* 'widely known, known far and wide.' Below the line we find no less than eleven expressions. The two following examples may serve to illustrate the wide semantic range the expressions in Mzw may have: *bekend zijn als de bonte hond* 'have a bad reputation, be notorious' and *publiek geheim* 'open secret.' The first example shows that it's not only words that intensify, phrases can do the same.

The third question for the user to ask the book is 'which words can be intensified by the word ….?' Let's take the adjective *duivels* 'devilish.' Step 1 requires a search in the first register, which goes from Intensifier > Expression > Headword. There we find, and that completes step 2, two collocations: *duivels dilemma* 'diabolic dilemma' and *duivels plezier* 'sinful/wicked pleasure', with references to *dilemma* and *plezier* respectively. It is worthwhile to follow the references, you may find out that you don't even need the word *duivels*.

The fourth search question the authors give is 'what was this expression again with ...?' So, what was this expression again with *woorden* 'words'? For the first step, we need the second register, Sorting word > Expression > Headword. Under the 'sorting word' *woord* 'word', we find *iets met zoveel woorden zeggen* 'say something in so many words' (step 2). It's listed three times, referring to the headwords *duidelijk* 'clear', *expliciet* 'explicit', and *nadrukkelijk* 'express' respectively. The third step involves checking the headwords referred to in the dictionary section. Again, it's worthwhile to take this third step if only for your own pleasure, because that's what Mzw abundantly gives to anyone interested in Dutch expressions.

There is a third, fairly short, register that is not discussed on the inside covers. This register claims that it will enable the user to find any of the over two thousand lemmata used to arrange the 7,500 expressions in the dictionary section. The headwords are grouped in eleven categories ranging from *Aandacht: valt het wel op?* 'Attention: does it attract attention?' to *Verstand: kan het worden begrepen?* '(Power of) reason: can it be understood?' Does the third register help us find the concept *ongeduld* 'impatience?' It does, *ongeduld* is under the category *gevoel* 'feeling', but you don't need the register to check that it's in the dictionary section. So, why this register? The best I can think of is that the different categories may stir up the users' inspiration.

Browsing through Mzw and writing this review I realized that I have got out of the habit of consulting paper reference works. The only paper reference works on my desk are those that are not available electronically. It appears that roughly the same collection of headwords is available on http://onderwoorden.nl/intensiveringen/. The online version is more extensive than the book when it comes to example sentences. On the other hand, online one does not find the many intensifying expressions that the book so generously provides.

Would I use Mzw myself? Yes, I certainly would. Mzw is a very useful, or even indispensible book for writers in Dutch and translators into Dutch who want to liven up their language. The user may occasionally find that the collection of concepts and expressions is not complete, but then the authors do not pretend to be complete. The Introduction and even the brief outline on the inside covers are excellent guides to the book's content.

To the best of my knowledge Mzw currently is the only printed collection of Dutch intensifications and intensifying expressions. Providing access to Dutch idioms through meaning has been done before, though not exactly in the way Schutz and Permentier did. Their approach makes Mzw the only Dutch reference work of its kind.

Anne Dykstra

globaLex

Globalex mid-2017

Following GLOBALEX 2016 Workshop (that was co-located with LREC, http://ailab.ijs.si/globalex/), the new Globalex website (http://globalex.link/) and preparatory board began to operate in June. The board consists of representatives of the five continental lexicography associations (Danie Prinsloo, Afrilex; Edward Finegan, DSNA; Ilan Kernerman, Asialex; Julia Miller, Australex; Lars Trap Jensen, Euralex) and of the eLex conference series (Iztok Kosem and Simon Krek). The members have been holding skype meetings about once a month, usually having to alternately miss someone due to the time difference between Australia and West Coast USA.

The initial outcomes so far were mainly in form of co-funding the website hosting (USD20 annual per association) and initiating mutual greetings and some participation in each other's conferences in 2017. Most practically, the website is meant to function as a hub for the publications of its members and others. The next milestone is submitting a proposal to hold the second GLOBALEX Workshop at LREC 2018 in Mizayaki, Japan (http://lrec2018.lrec-conf.org/en/). This might be held in cooperation with the Global WordNet Association with the main topic of Lexicography and WordNets. Details will be released in Q4 of 2017.

Ilan Kernerman

A brief account of ASIALEX 2017

Hai Xu

The 11th International Conference of the Asian Association for Lexicography (ASIALEX 2017) was organized by the National Key Research Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies and held in Guangzhou, China from June 10 to 12. As the organizer of its first international conference in 1999, we were proud to host the ASIALEX conference again after it had traveled around nine Asian countries and regions.

The year 2017 is a milestone for the Asian Association for Lexicography, celebrating the 20th anniversary of its foundation. For ASIALEX 2017, we received felicitations from the presidents of our global sister associations AFRILEX, AUSTRALEX, DSNA, and EURALEX.

The keynote speakers included:

- Prof. Jianhua Huang of Guangdong University of Foreign Studies, (the First President of ASIALEX)
- Prof. Andrea Abel of EURAC Research, President of EURALEX
- Dr. Julia Miller of Adelaide University, President of AUSTRALEX
- Dr. Michael Rundell, Editor-in-Chief of Macmillan Dictionary

We also organized two workshops, on Sketch Engine and DPS5, run by Miloš Jakubíček, of Lexical Computing, and Holger Hvelplund, of IDM, respectively.

The theme of ASIALEX 2017 was Lexicography in Asia: Challenges, Innovations and Prospects. The time is ripe to recognize Asian achievements in lexicographic research and practice over the past 20 years, and to look ahead to see how we can respond to new challenges of the revolutions in corpus linguistics and digital lexicography. In the keynote speeches, Huang and Abel spoke on the common theme of the dictionary user's orientation/participation in the digital age, and Rundell and Miller discussed extended units of meaning, or phraseology, which lexicographers are increasingly aware of as representing the norm, rather than the exception, in language. The issues addressed by these speakers consist of cutting-edge concerns, and most certainly deserve our closest attention.

The conference was met by high enthusiasm of scholars and publishers from Asia and beyond. As one of the largest conferences in its series, ASILAEX 2017 hosted around 160 participants from 75 institutes over 24 countries and regions in Asia, Europe, Africa and North America. Of the 130 abstracts submitted, the total number of papers accepted was 111. The talks roughly covered the following topics: digital lexicography, general-purpose lexicography, cognitive approaches to lexicography, bilingual lexicography, pedagogical lexicography, specialized lexicography, and historical lexicography. We were truly indebted to the contributors and reviewers for their hard work in bringing together such a remarkable meeting.

While preparations for this grand event were under way, we sadly lost two great lexicographers who were highly influential in both China and abroad: Professor Gusun Lu of Fudan University, who passed away on July 28, 2016, and Professor Boran Zhang of Nanjing University, who passed away on May 26, 2017. They both made enormous contributions to our field. To honour their achievements, we set up a special session in their memory on the topic of unabridged Chinese-English and English-Chinese dictionaries.

Hai Xu
Guangdong University of Foreign Studies
Convener, ASIALEX 2017

ASIALEX
The Asian Association for Lexicography

Executive Board 2017-2019

Rachel Edita O. Roxas • President | Vincent Ooi • Vice-President | Shirley Dita • Secretray | Deny A. Kwary • Treasurer | GAO Yongwei, LI Lan, Yukio Tono • Members | Jirapa Vitayapirak, Mehmet Gürlek • Conveners | Shigeru Yamada, XU Hai • Co-Chief Editors | Ilan Kernerman • Past President

http://asialex.org/#board

LEXICOGRAPHY Journal of ASIALEX

Shigeru Yamada (Waseda University, Japan) and Hai Xu (Guangdong University of Foreign Studies, China) were appointed Co-Chief Editors as of June 2017.

http://asialex.org/#journal

Next Conferences

ASIALEX 2018 will be held in Krabi, Thailand on 8-10 June.

ASIALEX 2019 will be at Istanbul University, Turkey in June 2019.

http://asialex.org/#conferences

Diccionarios electrónicos: perspectivas para el siglo XXI

In the last generation, lexicographers and terminologists have put into practice new methods for creating more dynamic lexical resources. This century traditional lexicography has experienced a veritable transformation, thanks to its entry in the digital era. A complete revolution that has been possible because language industries have benefited from NLP tools, corpus linguistics, cognitive science and cross-cultural studies.

Many innovative projects have shed new light on e-lexicography, such as the LEAD dictionary (Paquot 2012), the ARTES bilingual LSP dictionary (Kübler and Pecman 2012), DiCoInfo (L'Homme et al. 2012), Wiktionary (Meyer and Gurevych 2012), WordNet (Fellbaum 2010), FrameNet (Fillmore et al. 2003), DANTE (Atkins et al. 2010) and EcoLexicon (Faber et al. 2014). These lexical resources take advantage of all the design potential offered by electronic tools and innovative theoretical approaches. In addition, significant efforts were made by a wide range of agencies and organisms to produce powerful terminological databases such as IATE in Europe (http://iate.europa.eu/) or *Le grand dictionnaire terminologique* in Canada (http://granddictionnaire.com/).

Evidently, future generations of lexicographers will need to use NLP tools to describe language more accurately, since such resources allow the lexicographer to research and express the real use of language as reflected in large corpora rather than rely on armchair speculations. In this sense, the technical possibilities offered by the digital medium are a source of endless potential. For instance, since there is a wide variety of profiles that deal with different types of text and knowledge levels (Bowker 2012), tools can now be more easily adapted to different kinds of users with different needs (Bergenholz 2011). Consequently, the information in a lexical resource may vary, depending on whether, for instance, the text that the dictionary user wishes to create will be submitted to a high-impact journal or is intended for popular dissemination of scientific knowledge.

On the other hand, the ergonomic nature of the translator's 'workbench' has also greatly evolved. This is of paramount importance, since some dictionary users (e.g. translators) devote a significant amount of time to acquiring knowledge in order to understand the conceptual architecture of specialized texts.

Finally, the issue does not concern only how to improve existing tools but also how to produce new multifunctional and interactive e-lexicographic tools that would contain general, conceptual and specialized content (Bowker 2012).

With the aim of discussing these various issues, the training course entitled "Diccionarios electrónicos: perspectivas para el siglo XXI" is held as part of the El Escorial Summer School (Universidad Compultense Madrid) during 17-21 July 2017 under the joint organization of LexiCon Research Group (University of Granada) and the Digital Arts Master's Degree (Complutense University of Madrid), with the sponsorship of Cosnautas, Elhuyar, K Dictionaries and Lexical Computing.

The aim of this summer school is to provide a general overview of new trends in the creation of electronic dictionaries and terminological tools, combined with hands-on sessions where participants can obtain practical experience in dictionary design. Participants will learn about the current protocols, software, and practices in this field of the language industry and will thus acquire some of the necessary skills to use these tools effectively in the development of new dictionaries.

Beatriz Sánchez Cárdenas, grupo de investigación LexiCon, http://lexicon.ugr.es/
Amelia Sanz, grupo de investigación LEETHI, https://ucm.es/leethi

References

Atkins, B. T. S., Kilgarriff, A., and Rundell, M. 2010. Database of ANalysed Texts of English (DANTE): the NEID database project. In *Proceedings of the XIV Euralex International Congress*. Ljouwert: Afûk, 549-556.

Bergenholtz, H. 2011. Access to and presentation of needs-adapted data in monofunctional internet dictionaries. In Fuertes-Olivera, P. A., and Bergenholtz, H. (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury (Continuum), 30-53.

Bowker, L. 2012. *Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities*. In Granger, S., and Paquot, M. (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 379-397.

Faber, P., León Araúz, P., and Reimerink, A. 2014. Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era*. Educational Linguistics, 19. Springer, 267-301.

Fellbaum, C. 2010. WordNet. In *Theory and applications of ontology: computer applications*. Springer Netherlands, 231-243.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. 2003. Background to FrameNet. *International Journal of Lexicography* 16.3, 235-250.

Kübler, N., and Pecman, M. 2012. The ARTE bilingual LSP dictionary: From collocation to higher order phraseology. In Granger, S., and Paquot, M. (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 186-208.

L'Homme, M. C., Robichaud, B., and Leroyer, P. 2012. Encoding collocations in DiCoInfo: From formal to user-friendly representations. In Granger, S., and Paquot, M. (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 211-236

Meyer, C. M., and Gurevych, I. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S., and Paquot, M. (eds.), *Electronic Lexicography*. Oxford: Oxford University Press, 259-292.

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

The Corpus of Polish Sign Language and the *Corpus-based Dictionary of Polish Sign Language*

Polish Sign Language (polski język migowy, PJM) is a natural visual-gestural language that has evolved within the Polish Deaf community since around 1817, when the first school for the deaf was established in Warsaw. Until recently, PJM was highly understudied from the linguistic perspective, but founding the Section for Sign Linguistics at the University of Warsaw provided a unique opportunity to analyze PJM on the basis of solid empirical data. Established in 2010, this is the first Polish unit specializing in studies on the communication of the Deaf, with the aim to develop — based on a vast corpus of video recordings — a comprehensive grammatical and lexicographic description of PJM.

The PJM Corpus project aims at documenting the language which, despite very limited interest among the hearing majority of Poles, forms an important part of Polish and European linguistic and cultural heritage. The underlying idea is to create a database of richly annotated videos showing sign language utterances, produced by Deaf users of PJM reacting to more than 20 different elicitation tasks, such as retelling the content of picture stories and video clips presented to them during the recording session, naming objects, talking about themselves and their experiences, and discussing various topics pertaining to the Deaf.

The group of PJM Corpus participants is intended to be representative of the Polish signing community: they come from different parts of Poland and their selection has taken into account key sociological variables including age, gender, etc. The raw video material obtained in the recording sessions is further segmented, glossed (lemmatized), transcribed with the HamNoSys transcription symbols, translated into written Polish, and tagged with respect to various grammatical features using the iLex software developed at the University of Hamburg. The annotation conventions that are employed have been designed explicitly for this project.

This extensive set of data has been used as the empirical basis for the *Corpus-based Dictionary of Polish Sign Language* (2016), which is the first dictionary of PJM prepared in compliance with modern lexicographical standards. The dictionary was edited by Joanna Łacheta, Małgorzata Czajkowska-Kisil, Jadwiga Linde-Usiekniewicz and Paweł Rutkowski, and is an open-access publication available freely at http://slownikpjm.uw.edu.pl/en/.

Containing many hours of recorded material elicited from a range of individuals, the corpus makes it possible to ascertain which PJM signs are used by Deaf signers and how. Thanks to that, the dictionary records and describes real usage. The definitions are written in Polish, akin to the defining style in typical monolingual dictionaries, i.e. providing semantic information that is more extensive and precise than customarily provided in bilingual dictionaries. All sentential examples are drawn from authentic signed utterances found in the PJM Corpus. To standardize their appearance the original utterances were re-recorded by Deaf members of the dictionary team.

Another practical application of the project over the last three years concerns the development of multimedia textbook adaptations for schoolchildren with special educational needs (including Deaf and hard-of-hearing), commissioned by the Ministry of Education. These have the form of computer programs offering access to thousands of video files with PJM translations of all texts included in the original textbooks. Such attempts to ensure that the Deaf have equal opportunities to communicate, the right to full participation in social life and appropriate educational opportunities, would have not been possible without the solid linguistic foundations of the PJM Corpus and its derived dictionary. This is why the study of PJM grammar and lexicon is inextricably related to the issue of full linguistic rights of the Deaf minority, which is of particular importance in states such as Poland, where the full-fledged nature of sign language communication has been questioned for decades.

Paweł Rutkowski

Paweł Rutkowski, aged 39, is the founder and head of the Section for Sign Linguistics at the University of Warsaw, a general linguist and specialist in natural language syntax, and author of more than a hundred academic publications and textbooks. An alumnus of the University of Warsaw and Adam Mickiewicz University in Poznań, he has been awarded prizes, grants and scholarships, made research visits to leading European and American universities, and is member of the Polish Council for Sign Language at the Ministry of Family, Labor and Social Policy. Dr Rutkowski is winner of the Adam Kilgarriff Prize, 2017.
p.rutkowski@uw.edu.pl