

**Electronic lexicography in the 21st century:
linking lexical data in the digital age**

Proceedings of the eLex 2015 conference

Edited by

Iztok Kosem, Miloš Jakubiček, Jelena Kallas, Simon Krek

<https://elex.link/elex2015/>

11-13 August 2015

Herstmonceux Castle, United Kingdom

**Electronic lexicography in the 21st century:
linking lexical data in the digital age**

**Proceedings of the eLex 2015 conference, 11-13 August 2015,
Herstmonceux Castle, United Kingdom**

Edited by Iztok Kosem, Miloš Jakubíček, Jelena Kallas, Simon Krek

Published by Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia)
Lexical Computing Ltd. (Brighton, United Kingdom)

Creative Commons Attribution ShareAlike 4.0 International License

Ljubljana/Brighton, August 2015

CIP – cataloguing-in-publication
National and university library, Ljubljana, Slovenia

81'374:004.9(082)(0.034.2)

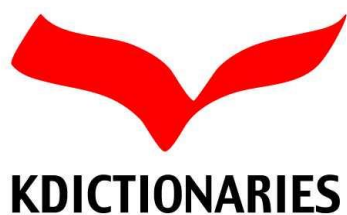
ELEX Conference (2015 ; Herstmonceux)
Electronic lexicography in the 21st century [Elektronski vir] : linking lexical
data in the digital age : proceedings of eLex 2015 Conference, 11-13 August
2015, Herstmonceux Castle, United Kingdom / editors Iztok Kosem ... [et al.]. –
El. knjiga. – Ljubljana: Trojina, Institute for Applied Slovene Studies; Brighton:
Lexical Computing, 2015

Način dostopa (URL): <https://elex.link/elex2015/proceedings/>

ISBN 978-961-93594-3-3 (Trojina, pdf)
1. Gl. stv. nasl. 2. Kosem, Iztok
280722944

Acknowledgements

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



ALPHARY

Words mean the world.



SLOVENIAN RESEARCH AGENCY

CONFERENCE COMMITTEES

Organising Committee

Iztok Kosem, chair
Miloš Jakubiček
Jelena Kallas
Simon Krek
Terka Olšanova

Scientific Committee

Andrea Abel	Robert Lew
Špela Arhar Holdt	Nikola Ljubešić
Nicoletta Calzolari	Henrik Lorentzen
Frantisek Čermak	Amalia Mendes
Patrick Drouin	Rosamund Moon
Darja Fišer	Christine Möhrs
Thierry Fontenelle	Carolin Müller-Spitzer
Polona Gantar	Hilary Nesi
Alexander Geyken	Vincent Ooi
Patrick Hanks	Magali Paquot
Ulrich Heid	Balint Sass
Kris Heylen	Kristina Štrkalj Despot
Ilan Kernerman	Arvi Tavast
Adam Kilgarriff	Carole Tiberius
Annette Klosa	Yukio Tono
Svetla Koeva	Lars Trap Jensen
Iztok Kosem	Agnes Tutin
Simon Krek	Tamas Varadi
Margit Langemets	Serge Verlinde
Lothar Lemnitzer	Piotr Zmigrodzki

TABLE OF CONTENTS

Automatic generation of the Estonian Collocations Dictionary database <i>Jelena KALLAS, Adam KILGARRIFF, Kristina KOPPEL, Elgar KUDRITSKI, Margit LANGEMETS, Jan MICHELFEIT, Maria TUULIK, Ülle VIKS</i>	1
Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German <i>Lothar LEMNITZER, Christian PÖLITZ, Jörg DIDAKOWSKI, Alexander GEYKEN</i>	21
Making a dictionary app from a lexical database: the case of the Contemporary Dictionary of the Swedish Academy <i>Louise HOLMER, Monica VON MARTENS, Emma SKÖLDBERG</i>	32
Towards an Electronic Specialized Dictionary for Learners <i>Marjan ALIPOUR, Benoît ROBICHAUD, Marie-Claude L'HOMME</i>	51
The role of crowdsourcing in lexicography <i>Jaka ČIBEJ, Darja FIŠER, Iztok KOSEM</i>	70
Mobile Lexicography: Let's Do it Right This Time! <i>Henrik KOHLER SIMONSEN</i>	84
What can a social network profile be used for in monolingual lexicography? Examples, strategies, desiderata <i>Monika BIESAGA</i>	105
The Construction of Online Health TermFinder and its English–Chinese Bilingualization <i>Jun DING, Pam PETERS, Adam SMITH</i>	123
Towards the enrichment of terminological resources by scientific corpora analysis <i>Izabella THOMAS, Iana ATANASSOVA</i>	136
medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin <i>Krzysztof NOWAK, Bruno BON</i>	152
Discovering hidden collocations in a bilingual Spanish–English dictionary <i>Margarita ALONSO RAMOS</i>	170
Management and exploitation of conceptual data and information in technical termbases: the electrotechnical vocabulary <i>Laura GIACOMINI</i>	186

Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography <i>Peter MEYER</i>	198
Using machine learning for semi-automatic expansion of the Historical Thesaurus of the Oxford English Dictionary <i>James MCCRACKEN</i>	211
What is a Target Language in an Electronic Dictionary? <i>Anna Helga HANNESDÓTTIR</i>	236
From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography <i>Ana ZWITTER VITEZ, Darja FIŠER</i>	250
Editing an automatically-generated index with K Index Editing Tool <i>Kseniya EGOROVA</i>	268
A study of the users of an online sign language dictionary <i>Mireille VALE</i>	281
Using a Maximum Entropy Classifier to link “good” corpus examples to dictionary senses <i>Alexander GEYKEN, Christian PÖLITZ, Thomas BARTZ</i>	304
Multilingual lexicography for adult immigrant groups: bringing strange bedfellows together <i>Anna VACALOPOULOU, Eleni EFTHIMIOU</i>	315
Overwriting knowledge: analyzing the dynamics of Wikipedia articles <i>Nathalie MEDERAKE</i>	327
Towards a Pan European Lexicography by Means of Linked (Open) Data <i>Thierry DECLERCK, Eveline WANDL-VOGT, Karlheinz MÖRTH</i>	342
Spell-checking on the fly? On the use of a Swedish dictionary app <i>Louise HOLMER, Ann-Kristin HULT, Emma SKÖLDBERG</i>	356
A multilingual trilogy: Developing three multi-language lexicographic datasets <i>Ilan KERNERMAN</i>	372
Multiple Access Paths for Digital Collections of Lexicographic Paper Slips <i>Toma TASOVAC, Snežana PETROVIĆ</i>	384
Longest–commonest Match <i>Adam KILGARRIFF, Vít BAISA, Pavel RYCHLÝ, Miloš JAKUBÍČEK</i>	397

GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary <i>Franck SAJOUS, Nabil HATHOUT</i>	405
Using machine learning for language and structure annotation in an 18th century dictionary <i>Petra BAGO, Nikola LJUBEŠIĆ</i>	427
DEBWrite: Free Customizable Web-based Dictionary Writing System <i>Adam RAMBOUSEK, Aleš HORÁK</i>	443
Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information <i>Pascale RENDERS, Gérard DETHIER, Esther BAIWIR</i>	452
Improving the use of electronic collocation resources by visual analytics techniques <i>Roberto CARLINI, Joan CODINA-FILBA, Leo WANNER</i>	461
Predicting corpus example quality via supervised machine learning <i>Nikola LJUBEŠIĆ, Mario PERONJA</i>	477
Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries <i>Ina RÖSIGER, Johannes SCHÄFER, Tanja GEORGE, Simon TANNERT, Ulrich HEID, Michael DORNA</i>	486
Linked Terminologies: Applying Linked Data Principles to Terminological Resources <i>Philipp CIMIANO, John P. MCCRAE, Víctor RODRÍGUEZ-DONCEL, Tatiana GORNOSTAY, Asunción GÓMEZ-PÉREZ, Benjamin SIEMONEIT, Andis LAGZDINS</i>	504

Automatic generation of the Estonian Collocations Dictionary database

**Jelena Kallas¹, Adam Kilgarriff², Kristina Koppel¹,
Elgar Kudritski¹, Margit Langemets¹, Jan Michelfeit²,
Maria Tuulik¹, Ülle Viks¹**

¹Institute of the Estonian Language, Tallinn, Estonia

²Lexical Computing Ltd., Brighton, England

E-mail: jelena.kallas@eki.ee, kristina.koppel@eki.ee, elgar.kudritski@eki.ee,
margit.langemets@eki.ee, jan.michelfeit@sketchengine.co.uk, maria.tuulik@eki.ee,
ylle.viks@eki.ee

Abstract

This paper reports on the process of the automatic generation of the Estonian Collocations Dictionary (ECD) database. The database has been compiled by the Institute of the Estonian Language in collaboration with Lexical Computing Ltd. The ECD is a monolingual online scholarly dictionary aimed at learners of Estonian as a foreign or second language at the upper intermediate and advanced levels. The dictionary contains about 10,000 headwords, including single and multi-word lexical items. The collocates within each headword are grouped according to the lexico-grammatical structure formed by the collocational phrase, and for collocations example sentences are provided.

For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al., 2004) functions Word List, Word Sketch and Good Dictionary Example (GDEX) were used. The data were automatically extracted in an XML format from the 463-million-word Estonian National Corpus and imported into the XML-based EELEX dictionary writing system. To make the importing of automatically extracted data from Sketch Engine into EELEX possible, the XML structure for extracted data was matched with the XML structure of ECD in EELEX. The ECD project started in 2014 and the dictionary is scheduled to be published in 2018.

Keywords: Corpus Lexicography; Collocations Dictionary; Corpus Query System; Dictionary Writing System; Estonian language

1. Introduction

Due to corpus lexicography development, the automatic generation of lexicographic databases has become a more and more common practise in e-lexicography. Adam Kilgarriff (2013: 78) points out that a corpus can support many aspects of dictionary creation: headword list development; the writing of individual entries, discovering word senses and other lexical units (fixed phrases, compounds, etc.); identifying the salient features of each lexical unit, their syntactic behaviour, the collocations they participate in, and any preferences they have for particular text-types or domains; and providing examples and translations.

As the focus of this article is on collocations, we will discuss the methods that are used for compiling collocations dictionaries and generating collocations databases. Based on the corpus analysis, two main approaches are implemented: automatic and semi-automatic. In the automatic approach, collocational information is automatically extracted from the corpus query system, users get direct access to non-edited collocation patterns and corpora example sentences through web interface, and no editorial work is done in terms of selecting and editing collocations. In the semi-automatic approach, collocational information is automatically extracted from the corpus query system and editorial work is done in order to clean and supplement the database, to reorder the collocates, to edit example sentences, etc.

Examples of the first approach include the projects SkELL (Baisa & Suchomel, 2014) and Wortprofil 2012 (Didakowski & Geyken, 2013). For the SkELL project, the Sketch Engine (Kilgarriff et al., 2004) function Word Sketch was used to discover collocates. By clicking on a collocate, a concordance with highlighted headwords and collocates is shown to users. SkELL uses a large text collection – SkELL corpus – specially gathered for the purpose of English language learning. There are more than 60 million sentences in the SkELL corpus and more than one billion words in total. This amount of textual information provides sufficient coverage of the everyday, standard, formal and professional English language. Wortprofil 2012 provides separated co-occurrence lists for 12 different grammatical relations and links them to their corpus contexts, where the node word and its collocate co-occur. The co-occurrence lists and their ordering are based on statistical computations over a fully automatic annotated German corpus containing about 1.8 billion tokens.

The second approach was implemented, for example, by Kosem et al. (2013). The corpus data (grammatical relations, collocations, examples and grammatical labels) were automatically extracted from the 1.18-billion-word Gigafida corpus of Slovene. After the data were extracted, they were post-processed by lexicographers. Analytical and editorial tasks were undertaken.

From the user's point of view, both approaches have their advantages. Providing users with edited, proofread material follows the classical conception of academic dictionary publication. The editorial team has full control over the outcome on each level of the dictionary micro-structure (headwords, collocations, example sentences, etc.). Providing users with direct access to the non-edited corpus data also has benefits. New users are often familiar with such software systems as web search engines and they consciously or unconsciously consider the post-processing of outcomes to be a natural task. In addition, direct access to the full set of non-edited corpora examples gives learners a broader overview of a collocation's behaviour in different contexts.

In this paper, we introduce the general concept of the dictionary and describe the approach that we used for the creation of the ECD database (see also Kallas et al., 2015). The data were automatically extracted from the corpus query system Sketch

Engine¹ (Kilgarriff et al., 2004), imported into the dictionary writing system EELEX² (Langemets et al., 2006; Jürviste et al., 2013) and will be post-processed by lexicographers. We have chosen the semi-automatic method for the following reasons. Firstly, the aim of the project was to compile an academic collocations dictionary with edited content. Secondly, the newest and the biggest Estonian National Corpus (EstonianNC)³ does not completely fulfil the criteria for a learners' dictionary. The corpus is not balanced; mostly it consists of periodicals, forums and blogs. This means that non-standard language (e.g. slang) is presented and needs to be removed manually. In addition, as the corpus includes field-specific science journals, terminological collocations need to be analysed separately and some removed in order to provide users with general language content only. Also, the output depends on the quality of the lemmatizer, the part-of-speech tagger and the morphological analysis. In terms of the Estonian National Corpus, there are still a lot of mistakes in tagging and as a result of insufficient disambiguation. This influences the quality of the outcome. The previously conducted evaluation of the Estonian Word Sketches revealed that two-thirds or more of the collocations were assessed by lexicographers as relevant and almost one-third were assessed as irrelevant (Kallas, 2013).

2. Estonian Collocations Dictionary

The Estonian Collocations Dictionary is a monolingual online, corpus-driven, scholarly dictionary aimed at learners of Estonian as a foreign language or second language at the upper intermediate and advanced levels (B2 to C1) according to the Common European Framework of Reference for Languages. The dictionary contains about 10,000 headwords, including single lexical items and multi-word lexical items (mostly multi-word verbs).

The primary source of the dictionary database is the recently compiled Estonian National Corpus (463 million words). The corpus consists of the Estonian Reference Corpus (contains texts written up to 2008) and the Estonian Web Corpus etTenTen13 (350 million tokens). etTenTen13 was compiled by Lexical Computing Ltd. It was crawled by SpiderLing (Pomikalek & Suchomel, 2012), encoded in UTF-8, cleaned and de-duplicated. The corpus was annotated morphologically, lemmatized, partially disambiguated and annotated by clauses by FiloSoft LLC, and installed into Sketch Engine software.

The Estonian National Corpus has 12 subcorpora (see Figure 1).

¹ <https://the.sketchengine.co.uk/auth/corpora/> (20.05.15).

² <http://eelex.eki.ee/> (20.05.15).

³ ske.li/estonian_national_corpus (20.05.15).

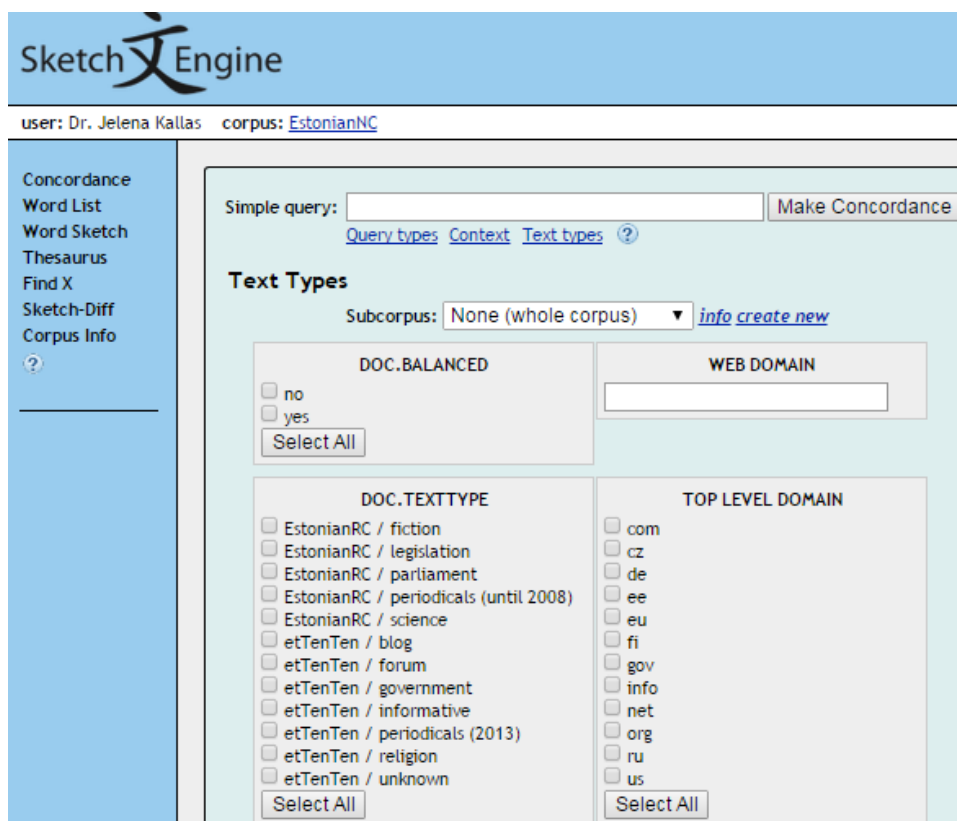


Figure 1: Subcorpora types of the Estonian National Corpus

Periodicals form 29% of the corpus, forums and blogs form 23%, informative texts 9%, parliament and religion subcorpora 4%, and unknown texts 35%. For text-type identification, FiloSoft LCC used 1) domain classification made by the Institute of the Estonian Language (e.g. periodicals and religion), 2) information in web addresses, and 3) the internal structure of the text (e.g. if a text contained a date, time or the word *vasta* 'answer-PRS-2SG', it was classified as a forum)⁴. During the mark-up of the corpus, text-type was added as metadata to the corpus.

In Estonian lexicography, the ECD project is the first dictionary focused exclusively on presenting collocational information in a systematic way. The analysis of Estonian dictionaries (Langemets et al., 2005; Kallas & Tuulik, 2011) determined that traditionally in Estonian dictionaries collocations are presented implicitly on the level of examples. The first attempt to present collocations explicitly was made in the Basic Estonian Dictionary⁵ (BED) project (Kallas et al., 2014). The dictionary contains 5,000 headwords, which correspond to B1-level vocabulary. On the first level, collocations were grouped according to the lexico-grammatical structure formed by the collocational phrase, e.g. Adj+N (adjective+noun) or Adv+V (adverb+verb). All together there were 13 types of collocation patterns in BED. On the second level, noun-verb collocations were sub-grouped according to the syntactical function of

⁴ <http://www2.keeleeveeb.ee/dict/corpus/ettenten/about.html> (19.05.15).

⁵ <http://www.eki.ee/dict/psv/> (19.05.15).

nouns (subject, object or adverbial), whereas other collocations were divided into semantically-motivated subgroups.

The ECD methodological conception follows the principles that were elaborated for the Basic Estonian Dictionary. The main difference is that the ECD, as a specialized dictionary, focuses on collocation patterns only; definitions are provided only for polysemous words, and there are no restrictions on vocabulary (in the BED, only words that were given as headwords in the dictionary could be used as parts of collocations). The advantage of the ECD compared to the BED is that we are able to give relevant collocations even if the frequency of one of the collocates is very low, e.g. *konn krooksub* 'frog croaks'. Often these collocations are particularly useful for learners.

For this project we define collocations as semantically transparent, meaningful and statistically significant combinations of content words with other lexical units. The typology of collocation patterns was elaborated for the ECD (see Table 1). Roth (2013: 155) indicates that in collocation lexicography one can distinguish two concepts: *node* and *collocate* (Sinclair, 1966) vs. *base* and *collocator* (Hausmann, 1985). In the ECD, we follow the concept of node and collocate, which means that each component of a collocation can be either a node or collocate, depending on the perspective. We have chosen this approach as we consider it to be more user-friendly. Our aim is for the user to find all frequent collocations connected to the headword in its entry while eliminating the need to navigate between entries. For example, if the user would like to see which nouns in Estonian collocate with the adjective *avar* 'spacious, wide, extensive', as it has a specific range of use, this can be performed within the entry of the adjective.

Noun patterns	
adjective + noun	ilus laul 'beautiful song'
noun (in genitive case) + noun	ekspertide hinnang 'expert opinion' koosoleku otsus 'the decision of the meeting'
noun (in partitive case) + noun	viil leiba 'slice of bread' viil juustu 'slice of cheese'
noun (in adverbial cases) + noun	kullast ehted 'gold jewellery'
noun (as subject) + verb	hobune hirnub 'horse neighs' palavik tõuseb, palavik langeb 'temperature rises, temperature falls'
noun (as object) + verb	arvutit sisse lülitama, arvutit välja lülitama 'turn on a computer / turn off a computer'
noun (as adverbial) + verb	aktsiatesse investeerima 'invest in stocks' arutlusele tulema 'enter into discussion'
noun+adpositional phrase	lepingu kohaselt 'according to a contract'
adverb + noun	raagus puud 'bare trees' omaette tuba 'separate room'
noun + verb in <i>ma-</i> or <i>da-</i> infinitive	meister valetama 'master to lie' soov laulda 'a wish to sing'
coordinating construction comparison constructions	päike ja tuul 'sun and wind' elu kui kabaree 'life as a cabaret'

Adjective patterns	
adjective + noun	raske otsus ‘hard decision’
noun (in adverbial cases) + adjective	rõõmsates toonides ‘in bright colours’ rõõmsal häälel ‘in a cheerful voice’
adverb + adjective	väga aeglane ‘very slow’ silmatorkavalt hea ‘strikingly good’
adjective (in translative case) + verb	rikkaks saama ‘get rich’
adjective (in essive case) + verb	rikkana tunduma ‘seem wealthy’
adjective + verb in <i>ma-</i> või <i>da-</i> infinitive	ilus vaadata ‘nice to look at’ raske mõista ‘hard to understand’
adjective + adjective	igavene suur ‘enormously big’
coordinating constructions comparison constructions	rikas ja ilus ‘rich and beautiful’ valge kui lumi ‘white as snow’ must nagu süsi ‘black as coal’
Adverb patterns	
adverb + adverb	aina rohkem ‘more and more’ väga kiiresti ‘very fast’
adverb + adjective	väga aeglane ‘very slow’
adverb + verb	kiiresti jooksmas ‘run fast’
noun + adverb	ideid täis ‘full of ideas’
coordinating construction comparison constructions	hästi ja kiiresti ‘well and fast’ kergelt kui õhk ‘lighter than air’
Verb patterns	
adverb + verb	kiiresti jooksmas ‘run fast’
noun (as subject) + verb	hobune hirnub ‘horse neighs’ palavik tõuseb, palavik langeb ‘temperature rises / temperature falls’
noun (as object) + verb	arvutit sisse lülitama, arvutit välja lülitama ‘turn on a computer / turn off a computer’
noun (as adverbial) + verb	aktsiatesse investeerima ‘invest in stocks’
adjective (in translative) + verb	täiskasvanuks saama ‘to become an adult’
adjective (in essive) + verb	rikkana tunduma ‘seem wealthy’
infinite verb + finite verb	ajab nutma ‘makes me cry’ jätab maksmata ‘leaves unpaid’
coordinating construction	kirjutama ja lugema ‘to write and read’

Table 1: Collocation patterns in ECD

Components of collocations are presented as lemmas (e.g. *hea laul* (good-ADJ-SG-NOM song-SG-NOM) ‘good song’, *omaette tuba* (separate-ADV room-SG-NOM) ‘separate room’) or in particular inflectional word forms (e.g. *viil leiba* (‘slice-SG-NOM bread-SG-PART) ‘slice of bread’, *rõõmsates toonides* (bright-ADJ-PL-INE colour-PL-INE) ‘in bright colours’). In this way, learners acquire additional grammatical information, which makes it easier for them to put the collocation into use.

For the grouping of collocations, we use morphosyntactic and syntactic criteria. At the first level, we group collocates according to their word class (with nouns, with adjectives, with adverbs and with verbs). Coordinating and comparison constructions are shown as separate units. At the second level, noun–noun, adjective–noun and

adjective–verb collocates are sub-grouped according to the inflectional word form (case) of the collocate, and noun–verb collocations are sub-grouped according to the syntactical function of the nouns (subject, object or adverbial). For sorting, we rely on raw frequency information and list collocates accordingly.

All collocation patterns are illustrated with example sentences, which were extracted automatically from the EstonianNC and will be post-processed by lexicographers. Where possible, we chose authentic examples, but if needed (e.g. very long sentences, specific vocabulary, slang or rare words) the sentences are shortened and edited.

3. Automatic generation of the database

For the automatic generation of the ECD database, we implemented the methodology proposed by Kosem et al. (2013: 35–36). The information was extracted from Sketch Engine (Kilgarriff et al., 2004) in an XML-format and imported into the EELEX dictionary writing system (Langemets et al., 2006; Jürviste et al., 2013). The procedure required the following: a selection of lemmas, fine-grained Sketch Grammar, GDEX (Kilgarriff et al., 2008) configuration, settings for extraction and the API script to extract data from Word Sketch.

3.1 Headword list development

The headword list of ECD contains 10,000 headwords. Only content words are presented as headwords: nouns, adjectives, verbs and adverbs. As Kilgarriff et al. (2014: 547) note, collocation dictionaries concern the core of the vocabulary: they are not for very rare words or grammatical words, but for common nouns, verbs and adjectives, which make up 99% of the headword list in a standard dictionary. In the ECD, nouns form 68%, adjectives 14%, verbs 15% and adverbs 3% of the headword list. Only manner adverbs are included in the headword list, e.g. *kergesti* 'easily' and *pehmelt* 'gently'.

For the creation of the headword list, the Sketch Engine function Word List was used.

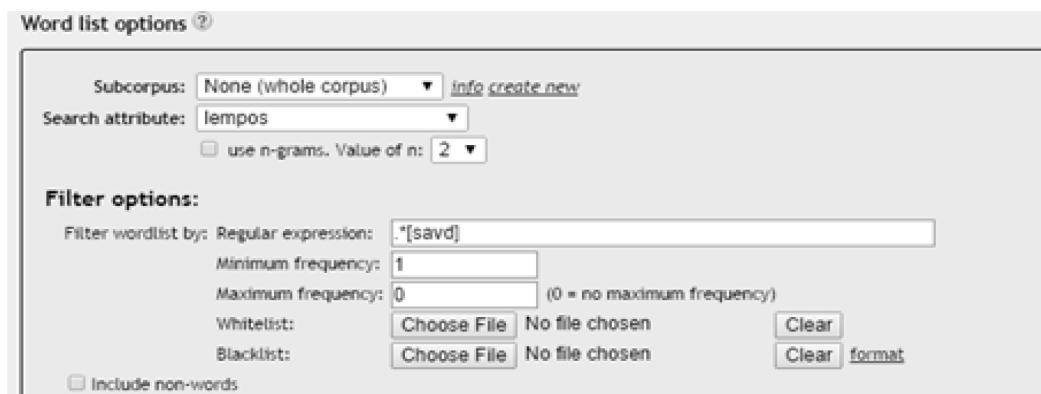


Figure 2: Word List function in Sketch Engine

Figure 2 illustrates the general parameters that were used for the headword list generation: the whole corpus is searched; the search attribute is *lempos*; regular expression is used to identify only words that are tagged as nouns, adjectives, verbs or adverbs; the minimal frequency of the lemma is 1; there is no maximum frequency.

As a basis for the ECD headword list, we took the first 10,500 frequent words, which needed to be checked manually. This was necessary to eliminate “noise” derived from mistakes in tagging and from insufficient disambiguation. Some headwords had to be removed, for example headwords with two kinds of spelling (e.g. *mänedžer* vs. *mänedzher* ‘manager’, *šokk* vs. *shokk* ‘shock’, *režiim* vs. *rezhiim* ‘regime’), abbreviations (e.g. *EEK*, *EUR* and *TOIM*), proper nouns and various terms (e.g. *süsinikdioksiid* ‘carbon dioxide’).

In parallel with corpus data analysis, we also used already existing lists of multi-word verbs. These lexical units were added manually.

After the headword list was developed, it was divided into two frequency classes: for Class I the most frequent 5,000 words, with a minimum frequency in EstonianNC of 5057; and for Class II the 5,000 mid-frequency words, with a minimum frequency in EstonianNC of 1057. Different settings for extraction were elaborated for different frequency classes (see section 3.4).

3.2 Sketch Grammar

For the detection of collocations, the Sketch Engine function Word Sketch was used. A word sketch is a summary of a word's grammatical and collocational behaviour (Kilgarriff et al., 2004).

Estonian Word Sketch Grammar is geared towards the specification of the Estonian National Corpus and relies on lists of syntagmatic relations of Estonian nouns, adjectives, adverbs and verbs, formed on the basis of traditional and formal grammar descriptions (Kallas, 2013). Word Sketch Grammar version 1.5 for Estonian was completed in 2013 and contained 85 rules. In 2014 the new version of Sketch Grammar was elaborated. Version 1.6 has 109 rules, including 16 *unary*-type rules (which make it possible to analyse the usage of inflectional forms of nouns and adjectives), four *symmetric*-type rules (which detect coordinating and comparison constructions, for example *päike ja tuul* ‘sun and wind’, *ilus ja noor* ‘beautiful and young’, and *hoolima ja hoolitsem* ‘to care and to take care’); 16 *dual*-type rules (which make it possible to search for co-occurrences of two lemmas, for example *päike + paistma* ‘sun + shine’), and 73 *colloc*-type rules (which make it possible to detect three-word collocations, for example *hoolitsema laste eest* ‘to take care of the kids’, and make it possible to present two-word collocations in a way that one component is presented as a lemma and the other in the particular inflectional form, for example *kari lambaid* (flock-SG-NOM sheep-PL-PART) ‘flock of sheep’, *rääkima aktsendita* (talk-INF accent-SG-ABE) ‘talk

without an accent', and *suhtuma lugupidamisega* (treat-INF respect-SG-COM) 'to treat with respect+'⁶.

Colloc-type rules proved to be very efficient for Estonian Sketch Grammar. Estonian has a rich morphological system: the nouns decline in 14 cases both in singular and plural; and verbs are inflected for tense, person, mood and voice (Liin et al., 2012). For that reason, presenting collocates as lemmas makes the whole collocation very opaque. *Colloc*-rules are particularly useful in the case of homonyms. Figure 3 displays a selection of grammatical relations for the homonyms *koor_1* (choir-SG-NOM): *koori* (choir-SG-GEN) vs. *koor_2* (peel-SG-NOM; cream -SG-NOM): *koore* (peel-SG-GEN; cream-SG-GEN), i.e. 'choir' vs. 'peel; cream': *kooris laulma* (choir-SG-INE sing-INF) 'sing in a choir', *kooriga liituma* (choir-SG-COM join-INF) 'join a choir', but *koorega kartulid* (peel-SG-NOM potato-PL-NOM) 'potatoes with peels', *koorega kohv* (cream-SG-COM coffee-SG-NOM) 'coffee with cream', etc.

koor (common noun)			
EstonianNC freq = 27,820 (49.39 per million)			
Constructions	Adj modifier 3,329 1.40	subject of 2,094 2.20	object of 392 1.10
omastav 10,304 1.90	röösk 858 12.21	laulma 245 9.15	riisuma 61 10.93
nimetav 8,750 1.30	tühi 150 5.53	esinema 121 7.14	lisama 19 2.76
kaasaütlev 2,262 3.10	suur 103 1.18	esitama 111 5.94	kasutama 17 2.27
seesütlev 1,801 1.50	õhuke 64 6.49	andma 98 3.75	juhatama 16 6.61
osastav 1,773 0.50	paks 64 5.36	saama 83 2.06	vahustama 13 8.77
alaleütlev 904 1.20			
seestütlev 886 1.10			
alalütlev 386 0.30			
omastav modifier 3,372 1.30	omastav modifies 4,553 1.80	participle modifier 607 1.80	ja/või 1,395 1.80
koguduse_koor 261 11.10	koori_dirigent 181 10.23	riivitud 107 10.76	piim 208 6.98
kiriku_koor 88 9.67	koori_liige 126 9.75	vahustatud 43 10.68	orkester 162 8.13
puu_koor 78 9.50	koori_repertuaar 95 9.36	osalenud 22 6.26	solist 58 7.20
kooli_koor 62 9.18	koori_peadirigent 79 9.10	laulnud 13 8.89	või 53 7.08
sidruni_koor 59 9.11	koori_laulja 78 9.08	loodud 13 4.34	ansambel 44 5.34
adverbial sisseütlev of 153 2.00	adverbial seesütlev of 556 3.60	adverbial kaasaütlev of 137 2.40	
koori_juhatama 82 13.48	kooris_laulma 189 13.02	kooriga_liituma 7 10.64	
koori_dirigeerima 11 11.10	kooris_hüüdma 26 10.52	kooriga_laulma 7 10.64	
koori_kuuluma 8 10.67	kooris_vastama 25 10.46	koorega_keitma 5 10.17	
koori_asutama 5 10.02	koorides_laulma 18 10.01	kooriga_töötama 5 10.17	
	kooris_karjuma 16 9.84		
kaasaütlev modifies 675 3.50			
keedetud_koorega 66 11.39			
kartulid_koorega 33 10.51			
kohv_koorega 15 9.45			
sibul_koorega 14 9.35			
seotud_kooriga 7 8.38			

Figure 3: Word Sketch for the noun *koor* 'choir; peel; cream' (from etTenTen13)

⁶ For more on directives used in the Sketch Grammar, see <https://www.sketchengine.co.uk/documentation/wiki/SkE/GrammarWriting> (20.05.15).

The new Sketch Grammar version 1.6 includes all of the lexico-grammatical structures that will be presented in the collocations dictionary (see Table 1). After the new version of Estonian Sketch Grammar was elaborated, settings for extraction were developed for nouns, adjectives, adverbs and verbs; we decided on such parameters as the frequency of the grammatical relation, the frequency of the co-occurrence of the collocates and the score of collocation (see section 3.4).

3.3 GDEX configurations

GDEX (Kilgarriff et al., 2008) is a tool that rates the quality of sentences and helps the lexicographer to select the best. GDEX works as a filter: it evaluates syntactic and lexical features of sentences and sorts concordances according to how perfectly they meet all the relevant criteria. As a result, GDEX offers a list of sentences: the better candidates are at the top of the list and the not-so-good ones at the bottom. The theoretical framework for GDEX development is proposed in Kilgarriff et al. (2008) and Kosem et al. (2011) and Kosem et al. (2013).

To clarify the GDEX parameters for Estonian, we used the example sentences of the Basic Estonian Dictionary (BED) and the Dictionary of Estonian (ED) (Langemets et al., 2010, to be published in 2018), and compared them to etTenTen13 web corpora sentences. The BED and ED dictionaries were used as the gold standard for dictionary example sentences. BED example sentences are compiled by lexicographers. They are didactic units and the aim is to show how words are used in context. The target audience of the ED is not language learners but well-educated native speakers. For that reason, the level of lexicographic adaptation of example sentences is much lower. etTenTen13 corpus sentences are fully authentic.

We analysed such parameters as the minimum and maximum number of words in a sentence, sentence length, word length and the number of subordinate clauses. Only sentences with substantives, adjectives, adverbs and verbs were taken into account. For each part of speech we analysed 150 sentences from three sources: 50 sentences from the BED, 50 sentences from the ED, and 50 sentences from etTenTen13. Tables 2 and 3 summarize the results of the analysis.

Quantitative analysis of the parameters clearly showed the peculiarity of sentences. Example sentences in BED, which has teaching purposes, are usually very short (the maximum number of words is 11, the average number of words in a sentence is 4.36–6.44). Sentences in ED are also rather short: the maximum number of words is 13 and the average number of words in a sentence is 4.72–6.42. Authentic sentences in corpora have very different characteristics. The difference is extremely large: the number of words in a sentence extends to 56 and the average number of words in a sentence is 15–16.9.

	Number of words	Average sentence length (words)	Average word length (characters)
Substantives			
BED	3–9	5.08	5.6
ED	3–12	6.42	6.7
etTenTen13	4–40	15.8	5.2
Adjectives			
BED	3–10	5.08	5.3
ED	5–11	6.44	6.7
etTenTen13	3–37	15	5.23
Verbs			
BED	3–7	4.36	6.21
ED	2–10	4.72	5.66
etTenTen13	6–56	16.9	6
Adverbs			
BED	3–11	5.44	4.96
ED	3–13	5.74	6.1
etTenTen13	7–42	16.8	5.64

Table 2: Parameters for BED and ED example sentences and etTenTen13 corpora sentences. Average word length varies only between 4.96 and 6.21 characters. At the same time, words in Estonian can be quite long, e.g. *kiiruisutamismestri võistlused* 'speed skating championships' (30 characters); so it is reasonable to also set maximum word lengths.

	Percentage of subordinate clauses (%)
Substantives	
BED	0%
ED	12%
etTenTen13	18%
Adjectives	
BED	0%
ED	14%
etTenTen13	58%
Verbs	
BED	8%
ED	10%
etTenTen13	76%
Adverbs	
BED	20%
ED	16%
etTenTen13	76%

Table 3: Percentage of subordinate clauses in BED, ED and etTenTen13 corpora sentences

The analysis of subordinate clauses showed that the number of subordinate clauses was rather small in the BED and ED example sentences, while authentic sentences in etTenTen13 web corpora included more subordinate clauses (18% in the case of substantives, 58% in the case of adjectives, and 76% in the case of verbs and adverbs) (see Table 3).

The reason for this might be that the lexicographer thinks of the example sentence as an addition to the definition and chooses not to add information that does not really illustrate a word's use. Sentences in web corpora reflect the desire and the need to provide readers with more context.

It also appeared that all the sentences in BED and ED included a predicate. In corpus sentences, there were a lot of elliptic sentences. Corpus sentences are also characterized by a large number of proper nouns and numbers.

Based on the empirical analysis of the sentences and also on the theoretical framework proposed by Kilgarriff et al. (2008), Kosem et al. (2011) and Kosem et al. (2013), we developed the following classifiers for GDEX for Estonian:

- whole sentences starting with capital letter and ending with (.), (!) or (?);
- sentences longer than five words;
- sentences shorter than 20 words;
- penalize sentences which contain words with a frequency of less than five words;
- penalize sentences with words longer than 20 characters;
- penalize sentences with more than two commas, or with brackets, colons, semicolons, hyphens, quotation marks and dashes;
- penalize sentences with words starting with capital letters. Penalize sentences with H (=Proper noun) and Y (=abbreviation) POS-tags;
- penalize sentences with “bad words”;
- penalize sentences with the pronouns *mina* 'I', *sina* 'you', *tema* 'he/she', *see* 'it' and *too* 'that', and the adverbs *siin* 'here' *seal* 'there';
- sentences should not start with the pronouns *mina* 'I', *sina* 'you' or *tema* 'he/she', or the local adverbs e.g. *siin* 'here' and *seal* 'there';
- penalize sentences which start with punctuation marks (typical informal texts) and with J (=conjunction) POS-tags;
- penalize sentences where lemmas are repeated;

- penalize sentences with tokens containing mixed symbols (e.g. letters and numbers), URLs and email addresses.

One parameter was that a sentence should contain a verb as a predicate; otherwise, the sentence was elliptical. But this parameter would only be possible to implement if the corpus was semantically annotated.

The blacklist is based on a list of words (compiled by FiloSoft LCC⁷) that the Estonian speller should not offer as replacements for unknown words. To supplement the list, we analysed words in the EDE dictionary that were marked as vulgar, pejorative, colloquial or slang. We added such words as *türa* 'dick', *narkots* 'dope', etc. We also added internet acronyms (*omg*, *wtf*, *lol*, *irw*) and curse words in English and Russian (*fuck*, *pohui*) and their adapted variants (*fakk*, *pohh*). The final list contained 446 words.

Figure 4 illustrates the API script written by Jan Michelfeit for the Estonian GDEX configuration.

```
min([word_frequency(w, 250000000) for w in words]) > 5
formula: >
(50 * all(is_whole_sentence(), length > 5, length < 20, max([len(w) for w in words]) < 20, blacklist(words, illegal_chars),
1-match(lemmas[0], adverbs_bad_start), min([word_frequency(w, 250000000) for w in words]) > 5)
+ 50 * optimal_interval(length, 10, 12)
* greylist(words, rare_chars, 0.05) * 1.09
* greylist(lemposs, anaphors, 0.1)
* greylist(lemmas, bad_words, 0.25)
* greylist(tags, abbreviation, 0.5)
* (0.5 + 0.5 * (tags[0] != conjunction))
* (1 - 0.5 * (tags[0] == verb) * match(features[0], verb_nonfinite_suffix))
) / 100

variables:
illegal_chars: ([<|\\|>|\\^|@])
rare_chars: ([A-Z0-9'.,!?:;-])
conjunction: J
abbreviation: Y
anaphors: ^(mina-p|sina-p|tema-p|see-p|too-p|siin-d|seal-d)$
adverbs_bad_start: ^(nagu|siin|siia|siit|seal|sinna|sealt|siis|seejärel)$
verb: V
verb_nonfinite_suffix: ^(mata|mast|mas|maks|des)$
bad_words: ^(loll|jama|kurat...)$
```

Figure 4: GDEX configuration file⁸

As a result, the output of GDEX improved substantially. Figure 5 illustrates that after the GDEX parameters were applied, there were considerably fewer subordinate clauses in the output and sentences were generally shorter.

⁷ The authors thank Heiki-Jaan Kaalep (FiloSoft LCC) for the list.

⁸ The list of 'bad words' is skipped.

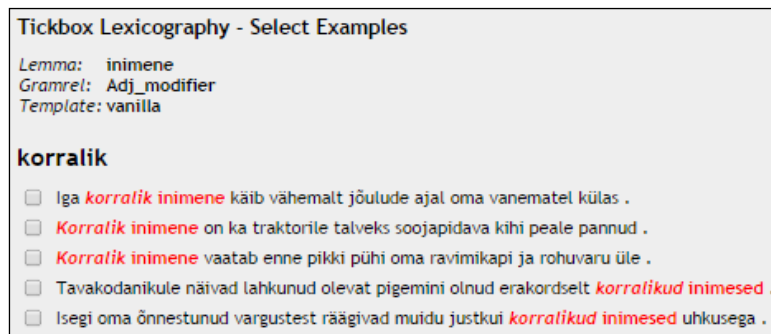


Figure 5: Automatically generated sentences for the collocation *korralik inimene* 'decent person'

For each collocation, we extracted five sentences, but for less frequent collocations there could be fewer than five examples in total. In this case, the program gave all examples without applying the parameters.

For future research testing additional GDEX classifiers proposed by Kosem et al. (2013) could be considered. For example, position of lemma, second collocate (collocate of collocate), or Levenshtein distance could be applied. We could test also different GDEX configurations for each word class.

3.4 Settings for extraction

The parameters used for the extraction of data were the following:

- a list of grammatical relations for nouns, adjectives, verbs and adverbs was elaborated. For nouns, we extracted 23 grammatical relations, for adjectives nine grammatical relations, for verbs 27 grammatical relations and for adverbs five grammatical relations;
- the minimal frequency of a collocate: 10 (for the frequency I class) and five (for the frequency II class);
- the minimal salience of a collocate: positive Dice, except for three grammatical relations (N_PP, Adj_PP and V_PP) we added that the Dice should be at least 2.00 (if less than 2.00 it is mostly noise);
- the minimum frequency of the grammatical relation: 10;
- the minimum salience of the grammatical relation: positive Dice;
- the number of examples sentences for a collocate: five.

We extracted collocates in a fixed order according to grammatical relations, e.g. for nouns first come adjectives, then verbs, then other nouns, then and/or-grammatical

relations. For some grammatical relations we also used stop-lists (e.g. modal verbs as collocates of nouns). Extracted collocates were ranked by frequency.

We also extracted all possible information about the frequency of collocates and grammatical relations:

- general frequency of lemmas;
- overall frequency of grammatical relations;
- overall score of grammatical relations;
- frequency of each collocate;
- score of each collocate.

Also GDEX-score could be extracted to show lexicographers how well the particular sentence corresponds to the parameters.

In perspective, it is possible to use frequency numbers for adding frequency labels ('star rating') to identify high-frequency, mid-frequency and low-frequency words. Also, statistical data can be used for different kinds of visualization of lexical data in the dictionary interface.

The data were extracted from Sketch Engine in XML-format (see Figure 6) and imported into the dictionary writing system EELEX (Langemets et al., 2006; Jürviste et al., 2011) (see Figure 7). To make the importing of automatically extracted data from Sketch Engine into EELEX possible, the XML structure for extracted data was matched with the XML structure of the ECD in EELEX.

```
<?xml version="1.0"?>
<sr>
  - <headword>
    <lemma>auto</lemma>
    <pos>s</pos>
    <freq>304721</freq>
  - <gramrel>
    <grname>Adj_modifier</grname>
    <freq>30618</freq>
    <score>1.240256</score>
  - <collocation>
    <collo>uus</collo>
    <freq>5498</freq>
    <score>6.830433</score>
  - <example>
    Uus
    <b>auto</b>
    ja tundmatu võistlus, sunnivad mehi prognoosides ettevaatlikeks.
  </example>
  - <example>
    Kavatsen soetada uue
    <b>auto</b>
    ja mark oleks kindlalt Škoda Octavia.
  </example>
  - <example>
    Ford nõuab sõitjailt häid tulemusi ning panustab samal ajal uue
    <b>auto</b>
    ehitamisse.
  </example>
  - <example>
    Selle asemel hakatakse käibemaksuga maksustama otseselt uute
    <b>autode</b>
    isiklikku kasutust.
  </example>
  - <example>
    Eesti Raudtee on aga müügiturul siiski pigem vana kui uue
    <b>auto</b>
    seisuses.
  </example>
</collocation>
- <collocation>
```

Figure 6: XML sample of generated database

As a result, we generated a database of ECD which contains 10,939 headwords, 82,678 grammatical relations, 493,971 collocates and 2,469,855 example sentences (five example sentences for each collocate). Additionally, the database includes the part-of-speech and overall frequency number of each headword, the overall frequency of each gramrel and collocate, and the score of each gramrel and collocation.

Currently, the database is being examined, edited and supplemented by lexicographers. The manual inspection and analysis of the collocates that were disregarded in the automatic extraction process are being carried out by lexicographers.

Preliminary observations regarding editing collocations are that deleting is necessary mainly in the case of mistakes in tagging and from insufficient disambiguation; in the case of specific terms that are not part of general Estonian (*analüütiline filosoofia* 'analytical philosophy'); and in the case of very frequent words that do not combine salient collocations with headwords: *mees* 'man', *naine* 'women', *tegema* 'to do', *ajama* 'to make; to drive', etc.

The screenshot displays the EELEX software interface. The left pane shows the XML view of an article for the headword 'kriitiliselt'. The XML structure includes tags for frequency (3449), grammatical relations (e.g., Adv_modifier), and various collocates with their respective frequencies and scores. The right pane shows a dictionary entry preview for 'kriitiliselt' with a frequency of 3449. It lists several grammatical relations and collocates with their frequencies and scores, such as 'Adv_modifier 372 (11.485423)', 'väga kriitiliselt 123 (2.844166)', and 'üsna kriitiliselt 36 (3.538450)'. The interface also shows navigation controls and a search bar.

Figure 7: The presentation of the extracted data in EELEX: editing window in XML view (left) and dictionary entry preview (right).

Regarding example sentences, although the initial idea was to present edited example sentences for each collocation, this proved to be too time-consuming. For one group, this can amount to 20 collocations and for one headword there are several collocational groups, thus leading to more than 200 sentences per entry. Therefore, we decided to give separate example sentences only for each collocation containing a verb and provide at least one example per group for other grammatical relations: adjective–noun, noun–noun, adverb–adjective, etc.

Figure 7 demonstrates the presentation of an outcome in the dictionary writing system EELEX.

4. Conclusions

For the automatic generation of the ECD database, the corpus query system Sketch Engine (Kilgarriff et al., 2004) Word List, Word Sketch and Good Dictionary Example (GDEX) functions were used. The data were automatically extracted in an XML format from the 463-million-word Estonian National Corpus and imported into the XML-based EELEX dictionary writing system (Langemets et al., 2006; Jürviste et al., 2011). To make the importing of automatically extracted data from Sketch Engine into EELEX possible, the XML structure for extracted data was matched with the XML structure of the ECD in EELEX.

We implemented the methodology proposed by Kosem et al. (2013). The procedure required the following: a selection of lemmas, fine-grained Sketch Grammar, GDEX (Kilgarriff et al., 2008) configuration, the API script to extract data from Word Sketch and settings for extraction. The list of lemmas was compiled using the Word List function. The latest Sketch Grammar version 1.6 was developed and improved; it includes all of the lexico-grammatical structures that will be presented in the ECD. The Grammar contains 116 rules in total. For the extraction of dictionary examples, the first version of GDEX for Estonian was developed. Classifiers connected with sentence optimum length, word optimum length, number of punctuation marks, word frequency, lemma repetition, anaphors, tokens with capital letters and symbols, abbreviations and a list of 'bad words' were proposed and implemented. The use of classifiers brought significant improvements to the output.

For automatic extraction, the following parameters were specified: a list of grammatical relations, minimum frequency and salience of grammatical relations, the number of collocates per grammatical relation, the minimum frequency and salience of a collocate, and the number of examples per collocate.

As a result, the database contains 10,939 headwords, 82,678 grammatical relations, 493,971 collocates and 2,469,855 example sentences (five example sentences for each collocate). Additionally, the database includes the part of speech and overall frequency number of each headword, the overall frequency of each gramrel and collocate, and the

score of each gramrel and collocation. Currently, the database is being examined, edited and supplemented by lexicographers.

5. Acknowledgements

The Estonian collocations dictionary project is supported by the National Programme for Estonian Language and Cultural Memory II (2014–2018) and the National Programme for Estonian Language Technology (2011–2017). The authors would also like to thank COST Action “European Network of e-Lexicography” for the opportunity to share knowledge and participate in network activities.

6. References

- Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno: Tribun EU, 2014. pp. 63–70.
- Didakowski, J. & Geyken, A. (2013). From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*. 2nd Work Report of the Academic Network "Internet Lexicography". Mannheim: Institut für Deutsche Sprache. (OPAL - Online publizierte Arbeiten zur Linguistik X/2012), pp. 43–52. Available at: http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf.
- EDE: *Eesti keele seletav sõnaraamat I–VI [The Explanatory Dictionary of Estonian]*. (2009). 2nd edition. M. Langemets, M. Tiits, T. Valdre, L. Veskis, Ü. Viks (eds.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.
- Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In H. Bergenholtz & J. Mugdan (eds.) *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.06.1984*. Tübingen: Niemeyer, pp. 118–129.
- Jürviste, M., Kallas, J., Langemets, M., Tuulik M. & Viks, Ü. (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In I. Kosem & K. Kosem (eds.) *eLexicography in the 21st Century: New Applications for New Users, Proceedings of eLex 2011, Bled, 10–12 November 2011*. Ljubljana: Trojina, Institute for Applied Slovenian Studies, pp. 106–112. Available at: <http://elex2011.trojina.si/Vsebina/proceedings/eLex2011-13.pdf>.
- Kallas, J. & Tuulik, M. (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamisprotsessid. *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]*, 7, pp. 59–75.
- Kallas, J. (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus - ja õppeleksikograafias. [Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography.] Tallinn: Tallinn University. Dissertations

on Humanities Sciences.

- Kallas, J.; Tuulik, M. & Langemets, M. (2014). The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 15–19 July 2014, Bolzano/Bozen. Bolzano/Bozen: European Academy, pp. 1109–1119. Available at: http://euralex2014.eurac.edu/en/callforpapers/Documents/EURALEX_Part_3.pdf.
- Kallas, J.; Koppel, K. & Tuulik, M. (2015). Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. [New possibilities in corpus lexicography based on the example of the Estonian Collocations Dictionary.] *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]*, 11, pp. 75–94.
- Kilgarriff, A.; Rychly, P.; Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In: G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress*. Lorient: Université de Bretagne Sud, pp. 105–116.
- Kilgarriff, A.; Husák, M.; McAdam, K.; Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra*, pp. 425–432.
- Kilgarriff, A. (2013). Using Corpora and the Web as Data Sources for Dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. Bloomsbury, London. Chapter 4.1, pp. 77–96.
- Kilgarriff, A.; Rychlý, P.; Jakubicek, M.; Kovář, V.; Baisa, V. & Kocincová, L. (2014). Extrinsic Corpus Evaluation with a Collocation Dictionary Task. In N. Calzolari, N. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (eds.) *LREC (Language Resources and Evaluation Conference), Reykjavik, Iceland*, pp. 454–552. Available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/52_Paper.pdf.
- Kosem, I.; Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, pp. 151–159. Available at: <http://elex2011.trojina.si/Vsebina/proceedings/eLex2011-19.pdf>.
- Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In: I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia*, pp. 17–19. Available at: http://eki.ee/elex2013/proceedings/eLex2013_03_Kosem+Gantar+Krek.pdf.
- Langemets, M.; Mägedi, M. & Viks, Ü. (2005). Süntaktiline info sõnastikus: probleeme ja väljavaateid. *Eesti Rakenduslingvistika Ühingu aastaraamat [Estonian Papers in Applied Linguistics]*, 1, pp. 71–98.
- Langemets, M.; Loopmann, A. & Viks, Ü. (2006). The IEL dictionary management system of Estonian. In G.-M. De Schryver (ed.) *DWS 2006: Proceedings of the*

- Fourth International Workshop on Dictionary Writing Systems: Pre-EURALEX workshop: Fourth International Workshop on Dictionary Writing System. Turin, 5th September 2006. Turin: University of Turin, pp. 11–16. Available at: <http://nlp.fi.muni.cz/dws06/dws2006.pdf>.
- Langemets, M.; Tiits, M; Valdre, T. & Voll, P. (2010). In spe: ühekõiteline eesti keele sõnaraamat. *Keel ja Kirjandus*, 11, pp. 793–810.
- Liin, K.; Muischnek, K.; Müürisep, K. & Vider, K. (2012). *Eesti keel digiajastul* [*The Estonian Language in the Digital Age*]. Valge raamatu sari [White Paper Series]. G. Rahm ja H. Uszkoreit (eds.). Heidelberg [etc.]: Springer.
- Pomikalek, J. & Suchomel, V. (2012). Efficient web crawling for large text corpora. In A. Kilgarriff & S. Sharoff (eds.) *Proceedings of the 7th Web-as-Corpus workshop, Lyon, France*, pp. 39-43.
- Roth, T. (2013). Going Online with a German Collocations Dictionary. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17–19 October 2013, Tallinn, Estonia, pp. 152–163. Available at: http://eki.ee/elex2013/proceedings/eLex2013_11_Roth.pdf.
- Sinclair, J. (1966). Beginning the Study of Lexis. In C. E. Bazell et al. (eds.) *In Memory of J. R. Firth*. London: Longman, pp. 410–430.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German

Lothar Lemnitzer¹, Christian Pölitz²,
Jörg Didakowski¹, Alexander Geyken¹

¹ Berlin-Brandenburgische Akademie der Wissenschaften, 10117 Berlin, Jägerstr. 22

² Technische Universität Dortmund, Fakultät für Informatik, Otto-Hahn-Str. 14, 44227

Dortmund

E-mail: {lemnitzer,didakowski,geyken}@bbaw.de, poelitz@uni-dortmund.de

Abstract

The work we will present in this paper is part of a dictionary project at the Berlin-Brandenburg Academy of Sciences and Humanities. For a large number of headwords, example sentences for their respective lexicographic descriptions have to be retrieved from a corpus of contemporary German. Lexicographers are typically faced with a huge number of corpus citations. Therefore, a tool that selects only good examples (those which are considered for inclusion into the dictionary) and dismisses the other ones would be time and effort effective. A rule-based good-example extractor proved to offer a good starting point, but the tool still delivers too many unacceptable citations. We have therefore tried to combine this tool with a machine learner that is trained on the decisions of an experienced lexicographer. The learner has been optimized to reject a large share of the example sentences. We present the machine learning results on a test data set with various combinations of linguistic features and quantify the gain in time and effort for the lexicographers. We also discuss the shortcomings of our approach and suggest some measures to counter them.

Keywords: example extraction; machine learning; corpus linguistics; German

1. Introduction and motivation

The work that will be reported in this paper originates from a large dictionary project at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). The task is to update a legacy dictionary of contemporary German (Klein & Geyken, 2010). Approximately 45,000 lexical units that have become part of the German vocabulary during the last 40 years have to be registered and handled lexicographically (cf. Geyken & Lemnitzer, 2012). One of the principles of the work is to illustrate the lexicographical description, in particular concerning the meanings and usages of lexical items, with citations from a large German corpus.

The underlying corpus has been built and continually extended at the BBAW (cf. Geyken, 2007). A large share of it can be consulted and queried through a search engine on the website of the project (www.dwds.de). The corpus currently contains

approximately 3 billion tokens. The sampling of new headwords from this corpus was mainly frequency based – most of the new headwords occur in these corpora with a frequency of >0.3 ppm (cf. Geyken & Lemnitzer, 2012); in absolute numbers: at least one thousand times. It is therefore impossible for a team of currently six lexicographers to read and check all these citations and to select the best three to five of them for inclusion in the dictionary article. Other straightforward alternatives such as sampling of k examples out of n , or just the first k examples, would not be satisfying either. Too many interesting contexts would escape the lexicographers’ attention just because these citations occur further down the list. It has therefore been decided early in the project to work with a “good example extractor”. The number of citations is parametrizable, i.e. the tool delivers for a headword those n citations that are ranked highest according to some qualitative criteria (see section 3 for further details). In the course of the lexicographical work – several hundred entries have currently been edited with the help of this tool – it was revealed that the selection of citations offered by this tool is still far from optimal. In particular, the number of “false positives”, i.e. citations which are ranked high but are rejected by the lexicographers, is still far too high. As little of the lexicographers’ work as possible should be wasted by checking bad corpus citations. To achieve this goal, it has been decided to post-process the output of the good-example extractor by a machine learning approach. The applied method should ideally learn lexicographical quality criteria and thus reduce the number of examples to those which are most likely to be considered by them for inclusion in the dictionary article.

In this paper we will report first results of this approach, i.e. of combining a rule-based good-example extractor with a machine learning component into a processing pipeline. In section 2, we will give an overview of related work. In section 3 we will briefly outline the operation mode of the rule-based extractor. In section 4 we will characterize the data we use for our machine learning experiments. Section 5 will be devoted to a description of our machine learning approach. The results of the experiments will be presented in section 6. We will end with a conclusion and an outline of our further work.

2. Related Work

Activities in the field of good-example extraction are comparatively recent. Of course discussion among lexicographers regarding what counts as good examples and for which purposes have been taking place for a long time. See, for example, Harras (1989) who mentions a list of linguistic criteria that a good lexicographic example should meet. Many of the introductions into (practical) lexicography, e.g. Svensén (2004: 281ff.), Atkins & Rundell (2008: 452ff.) and Engelberg & Lemnitzer (2009: 235ff.) devote at least a section to the function and quality of citations and other examples. However, only the advent of very large corpora that provide large numbers of citations made a (semi-)automatic pre-selection of material necessary. The seminal work in this field is that of Adam Kilgarriff and colleagues (Kilgarriff et al., 2008). They present a

rule-based approach to extracting good examples on the basis of some operationalisable quality criteria. The good example extractor implemented at the BBAW largely follows the approach presented in their paper (see section 3 and Didakowski et al., 2012). However, bringing ML methods into the field of automatic Gdex has recently become more impactful (cf. Rundell, 2014). In February 2015, a workshop of the “European Network of e-Lexicography was devoted exclusively to this topic (<http://www.elexicography.eu/working-groups/working-group-3/wg3-workshops/automatic-extraction-of-good-dictionary-examples>). On this occasion researchers from several European dictionary projects presented their work on that topic. To the best of our knowledge none of the work presented there has been published so far (but cf. Kosem et al., 2011 and Volodina et al., 2012). However, from the slides that are available on the website it can be deduced that some of the projects involve machine learning methods and tools in order to improve the precision of the extraction task.

3. Combining machine-learning with a rule-based approach

In Didakowski et al. (2012) we presented a good-example extractor that serves the lexicographers at the DWDS project by reducing the number of citations to be inspected. The extractor provides only those citations for a headword which are classified as most suitable with regards to a set of predefined rules. The extractor implements hard and soft rules which work on sentence level and global rules which work on a set of citations. The violation of a hard rule leads to immediate rejection of a citation. An example of such a rule is that a citation must be within a predefined range for sentence length. On the other hand, soft rules are used to rank the remaining citations by score. If a citation does not meet a soft rule it receives a lower score than a citation which does. A typical soft rule is that a citation should contain as few free pronouns as possible (for further details, cf. Didakowski et al., 2012). Additionally, the set of citations which is presented to the lexicographers should be well distributed among several text types (newspapers, novels, scientific prose, etc.) as well as over time – the dictionary should cover the period between 1900 and the present. For this purpose, global rules are applied to the ranked citation set making use of bibliographic metadata. In this connection the extractor is parametrizable – the users can decide how many citations are presented to them. The motivation behind using such a tool is not only to save time and effort for the lexicographers – who have more important things to do than reading hundreds of nearly identical and mostly uninteresting citations – but also to provide them with a “starter set” of typical usage types from which they should be able to construct the various senses of the headword. Furthermore, for the dictionary user the examples should be comprehensible without further context.

In the course of the work with that tool it became evident that 15 to 20 examples serve as a good material basis for the lexicographers to obtain an overview of the various uses of most of the lexical items. It also arose that the ratio of good to bad examples was less than optimal. Lexicographers are still confronted with too many examples

which they dismiss for various reasons. For example, many of the dismissed examples a) are structurally too complex to be exposed to the dictionary user; b) contain still too many pronouns and are therefore hard to comprehend without further context; c) are structurally incomplete even if the parser provides a “complete” analysis (list items are typical examples of such incomplete structures) or d) contain spelling or slight grammatical errors. It could thus offer a considerable saving of time (and money) if the lexicographers are provided with a smaller and better sample of citations. Such a task, however, is beyond the capabilities of a rule-based extractor that has to balance internal features, such as linguistic information, and external features, such as the temporal and topical distribution of the citations. For such reasons the idea arose to apply a machine learner to the output of the rule-based example extractor. The learner should be trained on the examples which have been already classified as either appropriate or not appropriate for inclusion into a dictionary article. In the future, the machine learning component should ideally reduce the inappropriate examples and keep the appropriate ones. In the following section we describe the data used for the training and testing of the learner.

4. The data

From the list of headwords that are to be included in the updated dictionary, we selected approximately 1,050 headwords. For each of these headwords, the good example extractor provided 18 examples at most – for some of the headwords only a smaller number of good examples were available. This totaled approximately 13,200 examples. All examples that had passed the rule-based good-example extractor were classified by one of the authors, a trained and experienced lexicographer, into one of two classes: (1) appropriate for inclusion, and (2) not appropriate for inclusion. These classified examples are used as training and test data for the machine learning task. The numbers in the data set are as follows: 5,984 have been labeled as appropriate (= class 1, “good”); 7,328 examples as not appropriate (= class 2, “bad”).

For the machine learning experiment, the set of classified examples was split into two, half for training and half for testing. Assignment to one of the two groups was done randomly. The distribution of good and bad examples over the two sets is shown in Table 1.

Quality Dataset	Good (= class 1)	Bad (= class 2)
Training set	3,607	3,011
Test set	2,377	4,317
Sum total	5,984	7,328

Table 1: Distribution of examples between the training and test sets.

Due to the random sampling, the distribution of class 1 and class 2 examples varies in the training and test sets. However, this difference in distribution does not affect the performance of the machine learning component.

5. The machine-learning approach

Our goal is to further refine the output of the good-example extractor from Didakowski et al. (2012) by combining it with a machine learning approach. We use Support Vector Machines (SVM, cf. Joachims, 1998) to “learn” which classifiers should be able to separate the good examples from the bad ones. The SVM learns a non-linear decision function that maps a set of features extracted from the example citations to a binary variable. We use several distinct representations of the texts in order to extract these features. In particular, we use a bag-of-words representation that encodes frequencies of words in the texts, parts-of-speech representations that assign word classes to the text tokens, and parse trees that encode syntactic structures. The text of each example is transformed into a sequence of these elements according to the different representations: for bag-of-words, the text is represented as sequence of words, for part-of-speech, the text is represented as sequence of the morpho-syntactic classes of the words; for parse trees, we represent the texts as sequences of trees in bracket notation.

For example, the text “I went to Lancaster” is represented as follows. For bag-of-words representation we receive “I went to Lancaster”; for part-of-speech we get “PP VVD TO NP”; and for the parse tree we are given “(S(NP-SBJ(PRD),VP(VVD,PP(TO,NP(NNP))))))”.

Sub-string kernels as proposed by Vishwanathan & Smola (2004) are used to calculate the similarity between examples based on common subsequences in the corresponding representations. All subsequences are used as features, i.e. all of the resulting substrings, sub-trees of the parse trees and sequences of part-of-speech tags. For instance, one feature of the above text “I went to Lancaster” in its part-of-speech representation is how many times the two labels “PR” and “VP” co-occur. Similarities between texts are encoded in a so-called kernel matrix that is used for the SVM. The entries in this matrix can be considered as indicators of the similarity of two texts based on the number of shared features, hence common sub-strings, common sub-graphs in the parse trees or common subsequences of parts-of-speech. Using the kernel matrix, we are able to train the SVM even on large feature sets, since we need only to calculate common subsequences instead of enumerating all possible subsequences of our texts in the corresponding representations. Further details on kernel methods can be found in Hoffmann et al. (2007).

We implemented our method in Java as Plugin in RapidMiner (Mierswa, 2009), a state of the art Data Mining tool. The bag-of-words representation was built by transforming the tokens of the example texts into normalized words (‘lemmas’). The

parts-of-speech and the parse trees were assigned to the texts by the Stanford Parser with a grammar for German (cf. Rafferty et al., 2008). The SVM was learned using the LibSVM library (cf. Chang et al., 2011) which is available in the RapidMiner software. The calculation of the kernel matrix was also implemented in Java as Plugin for RapidMiner. The individual kernel entries were calculated following Vishwanathan & Smola (2004). The implementation uses efficient data structures and hashing mechanisms that facilitate and therefore speed up the calculations. Thus we are able to calculate the kernel matrix for large data sets of many long text examples.

6. Results

The machine learner would be perfect if it would sort out (and remove) all examples from the test set which have been hand-labeled as not appropriate by the human annotator and, on the other hand, accept all examples which have been labeled as appropriate. We know that this is impossible. First, the decisions of the human annotator are arbitrary to some degree and cannot be predicted by even the best machine learner. Second, the training and test set may differ in many regards. Therefore, we can imagine the optimal result as either of the two following strategies: a) the learner tries to keep as many good (= class 1) examples as possible, at the price of also keeping (too) many of the bad (= class 2) examples. In other words, the learner will be optimized for a lower precision and a higher recall. That would be a conservative approach (i.e. one that conserves many examples for further inspection by the lexicographers); b) the learner tries to remove as many bad examples as possible, at the price of removing (too) many good examples along the way. In other words, the learner will be optimized for a higher precision and a lower recall. That would be the more radical approach.

Since our goal is to reduce the lexicographers' time spent reading and considering a surplus of bad examples, and in light of the fact that most headwords are represented by many examples in the corpus, we chose the second, radical, approach for the training strategy of the machine learner.

Subsequently, we will report on the performance of the learner on the test data set with three different sets of features: bag-of-words (or, more correctly, bag-of-lemmas), sequences of parts-of-speech and sub-trees of parse trees as well as combinations thereof. For each of these features we use the sub-sequence kernel described earlier to train a support vector machine such as machine learning. Since the decision is a binary one, i.e. assigning an example to one of two classes, and the performance of the learner is compared to human judgement, the data can be ordered and presented in a four-cell (2x2) contingency table. The four cells contain the number of examples that are a) assigned to class 1 by the human annotator ('ha') and by the machine learner ('ml'), b) assigned to class 2 by ha and class 1 by ml; c) assigned to class 1 by ha and class 2 by ml and d) assigned to class 2 by both ha and ml.

ha ml	class 1 (good)	class 2 (bad)	
class 1	603 (a)	487 (b)	1,090 (e)
class 2	1,774 (c)	3,830 (d)	5,604 (f)
	2,377 (g)	4,317 (h)	6,694 (i)

Table 2: A 2x2 contingency table for an example data set.

We can compute the marginal sums for each of the rows and columns (cells e–h) and the sum total (cell i). In Table 2 we present the full contingency table for one of our experiments. We can derive the following measures from this table:

- recall for class 1 examples = $603 / 2,377 = 25.3\%$ (i.e. approx. one fourth of the class 1 examples according to ha are labeled as such by ml)
- recall for class 2 examples = $3,830 / 4,317 = 88.7\%$
- precision for class 1 examples: $603 / 1,090 = 55.3\%$ (i.e. slightly more than half of the class 1 examples according to ha are accepted by ml, the rest are dismissed)
- precision for class 2 examples: $3,830 / 5,604 = 68.3\%$.

From these figures we further derive the F-score, i.e. the weighted mean of recall and precision as well as the accuracy. Accuracy is defined as the number of correctly-classified examples divided by the sum total of examples (i.e (cell a + cell d) / cell i). For our example:

- the F-score for class 1 examples is 0.34
- the F-score for class 2 examples is 0.76
- the accuracy is 0.66

In Table 3, we list the recall and precision for both class 1 and class 2 examples, which are listed for several feature settings.

Feature representation	Recall class 1	Precision class 1	Recall class 2	Precision class 2
Bag-of-lemmas	0.23	0.55	0.89	0.68
Part-of-speeches	0.30	0.57	0.87	0.69
Parse trees	0.32	0.60	0.88	0.70

Table 3: Recall and precision for both classes and different sets of features

From these values, the F-score for class 1 and class 2 examples, as well as the accuracy value, can be derived, see Table 4.

Feature representation	F-score class 1	F-score class 2	Accuracy
Bag-of-lemmas	0.32	0.76	0.66
Part-of-speeches	0.39	0.78	0.67
Parse trees	0.42	0.78	0.68

Table 4: F-score and accuracy for different sets of features.

The best values achieved are highlighted.

The data in Tables 3 and 4 show that all feature settings work reasonably well, i.e. we achieve a significant reduction of class 2 examples while still preserving a sufficient number of class 1 examples. The differences between the feature configurations are minimal, with the parse tree feature generating the best result.

From the point of view of the lexicographer, two questions are important beyond the measurable performance of the learner: i) how many (good, bad) examples do I get rid of? and ii) do I have to face, at the end of the selection process, a significant share of headwords with no example left at all? Let us look into both questions on the basis of our test data set and the example given in Table 1.

i) From the 6,694 examples that have been selected by the rule-based example extractor, only 1,090, i.e. 16.3%, have been accepted by the learner and therefore are available for the lexicographers' inspection. Of course, the loss of good examples is also considerable. In the example setting 1,774 class 1 citations would be lost, which leads us to the second question.

ii) The test data consist of examples for 438 headwords. For 415 of these, there is at least one example which has been classified into class 1 by the learner. Unfortunately, for only 342 of the headwords there is at least one example that has also been assigned to class 1 by the human annotator. The loss of (really) good examples is therefore considerable and should be remedied somehow.

As we have shown above, the implementation of a machine learning component as a filter is also a matter of choosing a good measure of permeability of such a filter. However, there is no invariant optimal setting for this measure. The optimal setting depends upon the task and the context. In our context, there were a sufficiently large number of citations to draw from, a limited amount of time for the lexicographers to inspect these examples and a rather small number of citations which were eventually selected for inclusion in the dictionary. The optimal setting in such a context equates

to a reduction of as many bad examples as possible, at the price of also removing many good examples. Nevertheless, it is not acceptable that for a larger amount of headwords no example is accepted at all. In the next section we will therefore present some suggestions of how to cope with this ‘collateral damage’.

7. Conclusions and further work

We have learned from our experiment that the machine learner, using the radical approach to removing example sentences from the initial set, also removes a considerable number of examples the lexicographers might want to see and potentially consider for inclusion in their articles. We, therefore, suggest the following strategies to remedy this ‘collateral damage’.

1. The simplest strategy would be to increase the initial data set, i.e. to instruct the rule-based good-example extractor to provide a larger number of example sentences. As a consequence, the number of examples that are accepted by the machine learner is larger but still of a higher quality than the set of examples that is initially delivered by the good-example extractor.
2. A more ambitious approach would be to use more information in order to balance the number of false negatives (= rejected examples we would like to see) against the number of false positives (= accepted examples which we would not like to see). One of the interesting characteristics of the machine learning approach that we have been using is that it does not only deliver a decision but also a confidence level for the decision. The confidence values for all possible decisions add up to 1; therefore, they can be interpreted as the probability that the decision is correct. Currently, the value is set to class 2 if the confidence towards this class is >0.5 . One could try to set a higher confidence level for the rejection of an example sentence. We have not yet looked into this, but an experiment with different thresholds might improve the results.

Another issue which affects all forms of example selection is the polysemy of many headwords. Typically, a polysemous word is very often used in one major sense, and less often or infrequently in its other(s) senses. This kind of distribution of usage examples over sentences makes each kind of sampling prone to the error of missing all examples for the infrequent sense(s). The burden to detect such gaps is again with the lexicographer. Ideally, the example sentences for a headword are initially grouped into clusters that, with more or less precision, represent different senses of the headword and outliers that cannot be easily assigned to any sense. Such an approach to combining good example extraction with word sense induction has been suggested by Rundell et al. (2014). We will in our future research follow the ideas expressed in this paper and apply them to our (German) data.

8. Acknowledgements

This research has been carried out in the context of the BMBF-funded project KobRA (*Korpus-basierte Recherche und Analyse mit Hilfe von Data-Mining*, grant ID 01UG1245) and the “Digitales Wörterbuch der Deutschen Sprache” (DWDS) at the Berlin-Brandenburg Academy of Sciences. We also want to express our gratitude to the developers of the “Rapid Miner” data mining tool.

9. References

- Atkins, S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Chih-Chung C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages. DOI=10.1145/1961189.1961199 <http://doi.acm.org/10.1145/1961189.1961199>
- Didakowski, J. et al. (2012). Automatic example sentence extraction for a contemporary German dictionary. In: Proceedings EURALEX 2012, Oslo, pp. 343-349.
- Engelberg, S.n & Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung. 4. Auflage*. Tübingen: Stauffenburg.
- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, C. (ed.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum, pp. 23-41.
- Geyken, A. & Lemnitzer, L. (2012). Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In *Proceedings of EURALEX 2012, Oslo*, pp. 362-366.
- Harras, G. (1989). Theorie des lexikographischen Beispiels. In F.J. Hausmann, O. Reichmann, H.E. Wiegand & L. Zgusta (eds.) *Wörterbücher Dictionaries Dictionnaires: Ein internationales Handbuch zur Lexikographie*, Berlin/New York: de Gruyter, pp. 1003-1114.
- Hofmann, T., Scholkopf, B. & Smola, A. J. (2007). 'Kernel Methods in Machine Learning', Published online at [arXiv.org \url{http://arxiv.org/abs/math/0701907v2}](http://arxiv.org/abs/math/0701907v2).
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In C. Nedellec & C. Rouveirol (eds.) *Proceedings of the 10th European Conference on Machine Learning (ECML '98)*, London: Springer-Verlag, pp. 137-142.
- Kilgarriff, A. et al. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus, In E. Bernal & J. DeCesaris (eds.) *Proceedings of the Thirteenth EURALEX International Congress*, Barcelona, Spain, pp. 425-432.
- Klein, W. & Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In U. Heid et al. (eds.). *Lexikographica*. Berlin/New York, pp. 79-93.
- Kosem, I., Husak, M. & McCarthy, D. (2011). GDEX for Slovene. Ljubljana. Trojina,

- Institute for Applied Slovene Studies. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex2011, Bled, Slovenia, 10 – 12 November 2011*, Ljubljana; Trojina, Institute for Applied Slovene Studies, pp. 151-159.
- Lodhi, H. et al. (2002). Text classification using string kernels. *J. Mach. Learn. Res.* 2 (March 2002), 419-444. DOI=10.1162/153244302760200687 <http://dx.doi.org/10.1162/153244302760200687>
- Mierswa, I. (2009), 'Non-Convex and Multi-Objective Optimization in Data Mining - Non-Convex and Multi-Objective Optimization for Statistical Learning and Numerical Feature Engineering.', PhD thesis, Universität Dortmund.
- Rafferty, A.N. & Manning, C.D. (2008). Parsing three German treebanks: lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German (PaGe '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 40-46.
- Rundell, M. et al. (2014): Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. In: *Proc. 16th EURALEX International congress, Bolzano, July 2014*, Bolzano: EURAC, pp. 319-329.
- Svensén, B. (2004). *A Handbook of Lexicography*. Cambridge: Cambridge University Press.
- Vishwanathan, S. V. N. & Smola, A. J. (2004). Fast Kernels for String and Tree Matching. In K. Tsuda, B. Schölkopf & J. Vert (eds.) 'Kernels and Bioinformatics' , MIT Press, Cambridge, MA, USA.
- Volodina, E. et al. (2012): Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation. Workshop on NLP in Computer-Assisted Language Learning. Proceedings of the SLTC 2012 workshop on NLP for CALL. Linköping Electronic Conference Proceedings 80: pp. 59–70. Accessed online: http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/SLTC2012_hitex_reviewed.pdf

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Making a dictionary app from a lexical database: the case of the Contemporary Dictionary of the Swedish Academy

Louise Holmer, Monica von Martens, Emma Sköldberg

Department of Swedish, University of Gothenburg

PO Box 200, SE-405 30 Gothenburg, Sweden

E-mail: louise.holmer@svenska.gu.se, monica.martens@gu.se, emma.skoeldberg@svenska.gu.se

Abstract

Developing an app version of a printed dictionary is a new challenge faced by lexicographers. Lexicographers involved in the app development process must consider fundamental lexicographic aspects as well as learn to understand technological and usage issues inherent to the new media. An inevitable question is how closely the content and layout can be made to match the printed dictionary while still offering ‘digital’ functionality such as linking, collapsed sections, audio, etc. Only a few reports discussing these issues have so far been published.

The aim of our paper is to further advance the exchange of knowledge and experience by sharing our observations made during the development of a new app corresponding to the comprehensive printed dictionary, *Svensk ordbok utgiven av Svenska Akademien* (the Contemporary Dictionary of the Swedish Academy, 2009). The app is the result of close cooperation between the financier The Swedish Academy, lexicographers and system developers at the Department of Swedish, University of Gothenburg and Isolve AB, a Stockholm-based app development agency specializing in dictionary apps.

Keywords: dictionary app; electronic dictionary; dictionary application; lexical database

1. Introduction

As the extent of use of smartphones and tablets increases, so does the development and use of dictionary apps. Gao (2013: 213) describes the present lexicographic situation as follows:

"In order to tap the potential of the vast global mobile market, dictionary publishers, large and small, have jumped on the appification bandwagon and launched their respective dictionary apps with the same zeal displayed a couple of years ago when they rolled out their online dictionaries."

Developing an app version of a printed or digital dictionary that meets the needs of both old and new user groups, however, is not that simple. During the development process, dictionary app developers must consider several fundamental lexicographic aspects. Surprisingly – and unfortunately for those who want to build on the experience amassed by other lexicographers – reports on the ideas and decisions behind each dictionary app are few in number (Holmer & Sköldberg, 2014; cf. e.g. Gao, 2013; Rundell, 2013; Simonsen, 2014a, b). Of course, all decisions made by

lexicographers and app developers are based on ideas about target users and their specific needs. In that sense, the prerequisites and conditions for apps are diverse. Nevertheless, we believe there is a need for a wide-ranging discussion of the considerations that go into dictionary apps.

The principal aim of our paper is to further the increased exchange of views on, and experiences of, dictionary app development and usage. We will present the ideas, principles and the lexical database of the new app corresponding to the comprehensive printed dictionary, *Svensk ordbok utgiven av Svenska Akademien*, or SO (the Contemporary Dictionary of the Swedish Academy, 2009). SO contains about 65,000 headwords with thorough definitions. It includes, among other items, exhaustive pronunciation information and morphological information such as word parts, word formation and derivatives. The dictionary also provides about 25,000 etymologies, and the year of its first occurrence in Swedish is given for every numbered sense. There are also about 1,000 well-known literary citations and about 400 elaborated usage guidelines (see Malmgren, 2009; Malmgren & Sköldberg, 2013). The SO app is the result of close cooperation between the Swedish Academy, lexicographers and system developers at the Department of Swedish, University of Gothenburg (where the authors of this paper are employed), and Isolve AB, one of the leading dictionary app development agencies in Sweden.

In the next section we discuss dictionary app development and the results of studies on dictionary app use. Section 3 presents the lexical database and the IT environment at the Department of Swedish, University of Gothenburg. We discuss the SO app in section 4, with a focus on its content, design and search functions. Section 5 contains our final remarks.

2. Development and use of dictionary apps

A lexicographic team faces many issues when developing a dictionary app for smartphones and tablets. One important question concerns the app content in relation to the lexicographic data in the corresponding printed or online dictionary; that is, must the content be identical? Another key issue is how to display and make the most of the content while taking advantage of the inherent functionality of each platform. As Rundell (2013: 5) points out, a dictionary accessed on a computer or a mobile device has considerable advantages over its analogue predecessors. One obvious benefit is related to space. Gao (2013: 215) points out that “unlimited space offers the lexicographers a variety of choices, such as the addition of many entries, the multimedia content, the listing of related words, and the inclusion of more than one language in the dictionary, etc.” However, according to Lew (in press) it is very important to make a distinction between *storage space* and *presentation space* in a lexicographic resource. Due to the size of a smartphone or tablet interface, the presentation space of dictionary apps is very limited and this must always be kept in mind when considering possibilities and preparing the data. But Lew also discusses a

third category of space, which he calls *perceptual space*. This concerns the capacity of the dictionary user to perceive and process data. In other words, compared with storage and presentation space, perceptual space is not a property of the dictionary or the medium, but rather of the user. Lew states that presenting an overly rich microstructure can lead to information overload. He writes: “As a result, users find it difficult to extract the relevant information and may be less willing to proceed beyond the initial sense(s) of an entry” (see also Tarp, 2012 on problems related to information overload in dictionaries). Lew (in press) concludes that user research is needed to first establish what content should be immediately displayed on the screen, and what content should be deferred.

The issue of space has yet another aspect that is less often discussed by lexicographers. The possibility of accessing related content via hyperlinks in the text does in fact save storage space – memory – in any well-structured electronic dictionary because the need to duplicate information is more or less eradicated. In a printed dictionary, redundancy is necessary to avoid forcing the reader to shift focus from one entry to another; in an electronic dictionary, it is not only unnecessary but highly inadvisable. Duplication of information is the mother of inconsistency and should be avoided as far as possible, especially since the users of digital media often expect more frequent content updates, which serves to dramatically increase the problem of data integrity if the same information is stored in multiple locations.

There are also semi-technical decisions to be made when developing a dictionary app, such as whether the app is going to work online, offline, or perhaps be a hybrid of the two types. Among the dictionary apps developed in the Nordic countries, a clear majority seem to work offline, i.e., the entire dictionary content is downloaded to the phone/tablet upon installation. This applies for instance to the apps developed by Norstedts, the leading commercial dictionary publisher in Sweden. The dictionary apps developed by the Society for Danish Language and Literature, on the other hand, are online apps, which means that the mobile device must be connected to the internet to work. Merriam-Webster Dictionary apps can be classified as hybrids. No internet connection is required to view definitions and transliterations of pronunciation, however, users do need network access to hear audio pronunciations, study the illustrations and use the voice search feature. Generally, it can be regarded as a disadvantage if a mobile app requires network access since the connection might be slow, unstable, non-existent or expensive (Rundell, 2013: 5). However, the online format also has very clear advantages, not least in view of the possibility of linking to an online version of the dictionary, updating the content, and presenting an up-to-date *Word of the Day* (see Holmer & Sköldberg, 2014).

Other issues arise when considering which mobile devices and operating systems to focus on targeting (iPhone, Android, etc.), and similarities and differences between operating systems, smartphone models and tablet versions (cf. Winestock & Jeong, 2014: 112).

Finally, a dictionary app producer must make decisions about pricing. Since the mid-2000s, dictionary sales have fallen sharply in Sweden. As Törnqvist (2010: 485–486) points out, many Swedish users now expect linguistic information to be available free of charge. Rundell (2013: 11) reports on the situation of English dictionaries, which appears more positive. However, he also states that the only digital products dictionary users seem to want to spend money on are those that can be installed on their own device, e.g. as an app.

Additionally, there is growing awareness of the importance of open access to data and software produced with the help of government grants, at least in the Nordic countries. For example, the Ministry of Education in Finland has demanded that any dictionary produced by the Institute for the Languages of Finland must be accessible to the public free of charge. This means that previously existing partnerships between academic institutions, commercial publishers and software producers must be reconsidered.

Depending on the commercial market for funding, dictionary producers have to think outside the box. Rundell (2013: 11) states that “What we need is a new entrepreneurship to create new products for new users, doing what we have always done: helping people to write, learn and understand language, working closely together with scientists and programmers to finally step into the digital future.” Dictionaries are a fundamental component of the communicative and cultural infrastructure of a language community – the question is not whether dictionaries will exist in the future, but rather whether lexicographers will be a part of producing those dictionaries.

Very little information is available regarding the underlying ideas, visions and the actual development process behind the dictionary apps available today. Consequently, the usual procedure of benchmarking before embarking on a development project in order to grasp the state of the art is no easy task. There are few reviews of dictionary apps (see e.g. Holmer, 2011; Hoel, 2012; Svarverud, 2014 with a Nordic perspective). Holmer & Sköldberg (2014) examine four Danish apps developed by the Society for Danish Language and Literature. They conclude that these monolingual apps differ widely in terms of functionality and presentation of dictionary content. They also raise the issue of whether the app format is suitable for all kinds of dictionaries. The modern Danish Dictionary app (DDO) is perceived as a dynamic and very well-functioning lexicographic product that can serve as a model for other apps. The legacy dictionary apps (e.g. such as for the *Dictionary of the Danish Language*), are much simpler in terms of both dictionary content and functionality, and serve mainly as advertisements of the online version of the same dictionary.

Holmer & Sköldberg (2014) also question to what extent dictionary app development is – or should be – based on the results of user studies. Surveys of dictionary app usage are still few in number. Marelló (2014) compares the usage of three different

versions of a bilingual dictionary (printed, online and app) among high school students. Simonsen (2014a, b) studies the usage of the dictionary app Medicin.dk, a knowledge-based medical resource used by most health care professionals in Denmark. Based on his data, Simonsen (2014b: 259–260) states that the typical mobile user is on the move and accesses information on the go, typically performing simple searches. This makes the mobile user impatient, imprecise and preoccupied with other things. The mobile user's situation primarily supports simple, precise, communicative lexicographic functions, and is not suited to support complex, cognitive lexicographic functions. Simonsen (2014a, b) also points out that the mobile user navigates both the physical world and the user interface of the mobile device at the same time. Finally, the size of the user interface and the typical user's situation mean that complex data and long text segments do not constitute optimal mobile data.

A user study that is highly relevant in relation to the development of the SO app is that carried out by the editors of the *Swedish Academy Glossary* (SAOL) (Holmer, Hult & Sköldböck, 2015). The SAOL is a monolingual Swedish glossary that contains about 125,000 headwords. It provides information primarily on spelling and inflection and explains the meaning of words to a minor extent. The app version of the SAOL was released in 2011 and is based on the 13th edition of the glossary, published in 2006. The app reflects the printed version and provides the full inflectional pattern for each lemma.

The app user study was performed in the early spring of 2015 in the form of a web survey. The questions concerned user behaviour and situations, opinions on the design and layout of the app, suggestions for a forthcoming version and background information about the respondents. The study resulted in 264 submitted questionnaires with a very low internal dropout rate.

The results of the study show that the app is mainly used for spelling, meaning, inflection and checking whether a particular word is included in the glossary. The respondents were fairly well-educated and often use the app in situations related to work. Overall, the respondents were very satisfied with the app and always or almost always find what they are looking for. When it comes to pricing and willingness to pay for a future version (the current version is free), older users were a bit more willing to pay, and most respondents said they would not want to pay more than 50 Swedish kronor (a bit more than 5 euros). In a forthcoming version, many respondents would like to have a wildcard search (a feature that is lacking in the current version), and cross-referencing via hyperlinks. By asking about their latest lookup, the editors discovered that the respondents tend to search for words that do not belong to the core vocabulary, and in their comments they explained that they were looking for object forms and variant forms of words, correct spelling, etc. Many of them also expressed a need for more detailed definitions, and some suggested an app version of the more comprehensive SO dictionary, i.e. the forthcoming app presented in this paper.

Before going into detail regarding the SO app, in the next section we give an overview of the lexical database and the related IT environment at the Centre for Lexicology and Lexicography, Department of Swedish, University of Gothenburg.

3. The Lexical Database and IT environment

3.1 Background

The origin of the database we are using to produce our app dates back to the mid-1960s when Sture Allén¹, a pioneer of computer-based lexicography at the University of Gothenburg, started gathering frequency-based data on contemporary language (see Malmgren & Sköldberg, 2013). These data evolved into a highly structured lexical database, designed mainly by Christian Sjögreen, a leading systems engineer at the subsequently formed Department of Computational Linguistics.

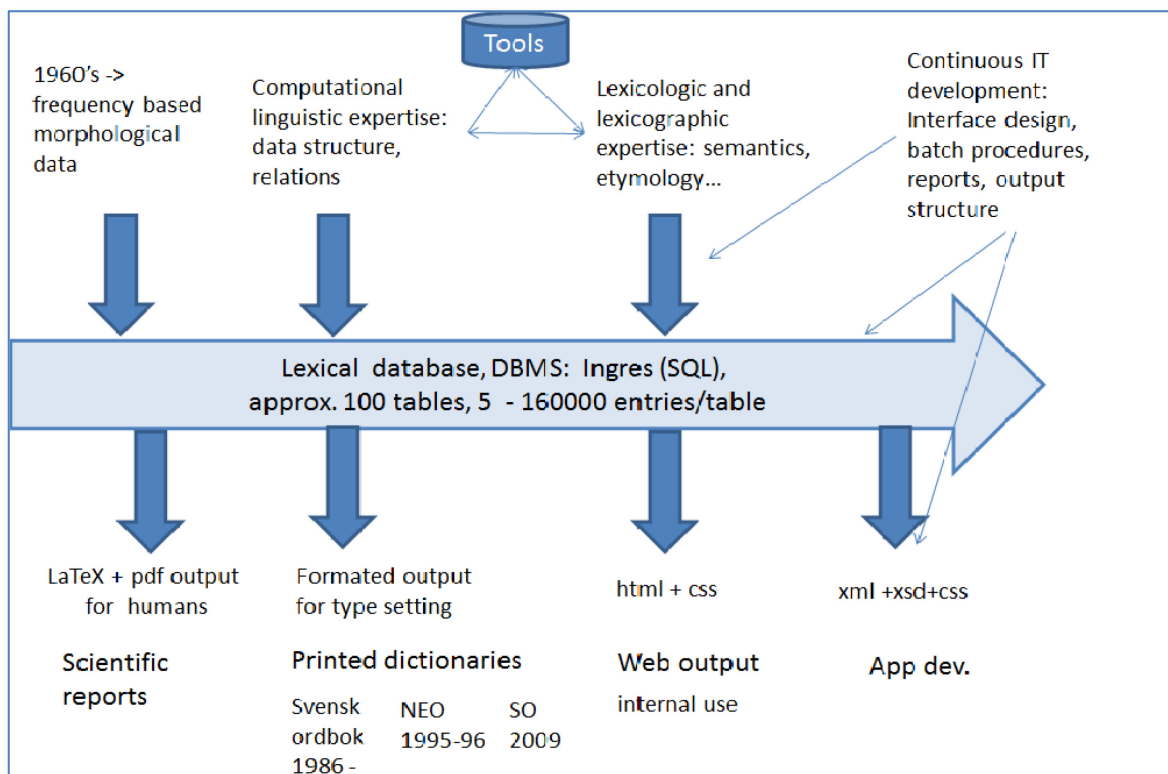


Figure 1: The Lexical Database, overview of processes, input and output

This database has since been continuously augmented and updated and used to produce a number of printed dictionaries in collaboration with different publishers (cf. overview in Figure 1). The latest printed dictionary produced was the SO, published in 2009. The database is currently owned by the Swedish Academy and maintained by the Department of Swedish at the University of Gothenburg. The Swedish Academy guarantees long-term funding of the work done by lexicographers, system administrators and developers at the university.

¹ See http://www.svenskaakademien.se/en/the_academy/members/938e01b1-b318-4c23-ba05-954127697d2a.

3.2 Infrastructure and traditional input/output

All data are stored in a dedicated relational database using the Ingres DBMS. With some exceptions, each information type has its own table (cf. Figure 2, where for example, information on pronunciation is stored in its own table with separate columns for primary and secondary pronunciation). Each main item has a unique number which acts as the key when joining tables, i.e. the classic relational database architecture.

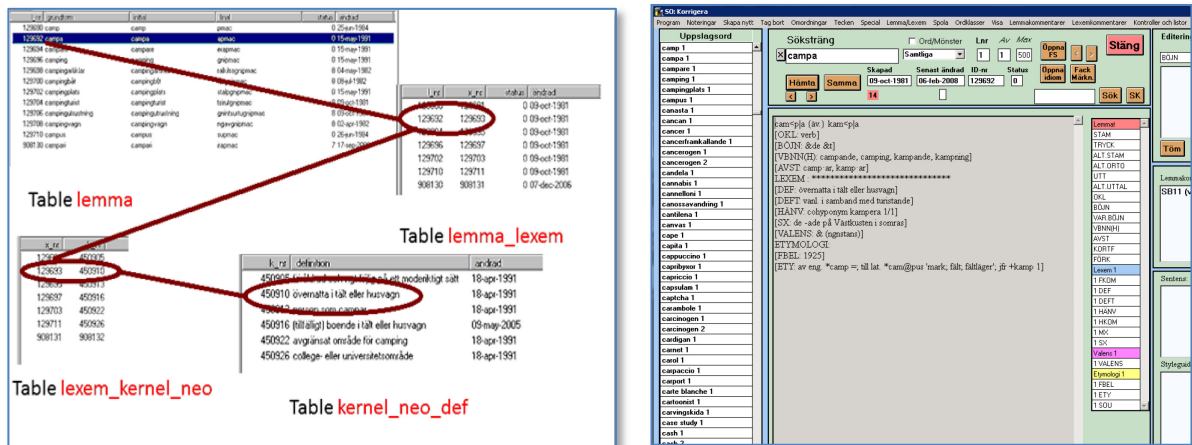


Figure 2: Ingres tables and editing frame

The front end used for editing is designed in OpenRoad and users seldom need to be concerned with the underlying structure (see Figure 2). Editing is performed by scholars and PhD students specialized in lexicography. A number of in-house systems, such as corpora and morphological databases, have been developed in order to serve as resources for the lexicographic team.

Traditionally, output from the system has been through C programs producing LaTeX output converted either to PDF for human eyes or a format suitable for typesetting. Figure 3 shows an entry from the most recent printed dictionary (SO, 2009).

cam'pa äv. kam'pa verb ~de ~t ♦ camp-ar, kamp-ar
 SUBST.: **campande, camping, kampande, kampning**
 • övernatta i tält eller husvagn vanl. i samband med
 turistande {JFR kampera}: *de ~de på Västkusten i somras*
 KONSTR.: ~ (ngnstans)
 HIST.: sedan 1925; av eng. *camp* med samma bet.; till lat.
cam'pus 'mark; fält; fältläger'; jfr ¹kamp

Figure 3: Dictionary entry for the word **camp** ('to camp') in SO (2009)

3.3 Redesigned for new media

In 2010 there were plans to create a Swedish language website presenting online versions of the dictionaries and other language resources owned by the Swedish Academy. Development addressing this goal resulted in a PHP/HTML application that was subsequently used as an in-house tool.

The main issues addressed during this phase were:

- Converting typesetting instructions into tag attributes + CSS style sheets
- Identifying and converting special characters into standard UTF-8
- Creating functions for identifying and linking referenced words
- Adapting the dictionary article layout to web browser functionality, e.g. using morphological information from another lexical resource² to produce flexible line breaks (soft hyphens)
- Amending inconsistencies and fixing referencing errors in the database
- Identifying parts of the text structure suitable for collapsing (see Figures 4 and 5)



Figures 4 and 5: HTML prototype showing collapsed text

² Svensk Morfologisk Databas (SMDB), an in-house tool for handling morphological data.

The decision was made to proceed with an app instead of a browser-based online version of the printed dictionary in 2013. App production was contracted to Isolve AB. The content of the app was produced using a modified version of the existing PHP/HTML application. The XML file exchange format was used for enhanced verifiability and the structure was formalized in an XSD schema file (cf. Figures 6 and 7). Audio files were also added.

```

<artikel><lemma id="lnr129692"><homonr></homonr><grundform>cam`pa</grundform><ljudfil
filnamn="129692_1.mp3" /><alt><altkom>äv.
</altkom><fm>kam`pa</fm><ortoref><hvtag>lnr202472</hvtag><hvord><hvhomo></hvhomo>kam`pa</hvord></ortor
ef></alt>
<ordklass>verb</ordklass><bojning> ~de ~t</bojning><avstav>camp.ar, kamp.ar</avstav><lexem
id="xnr129693"><kernel id="knr450910" /><punkt></punkt><def>över-natta i tält eller hus-vagn</def>
<def>vanligen i sam-band med turistande</def> <fack>sport.</fack>
<hy><hvrgrupp><hvtyp>2FR</hvtyp><hvtyp2>cohyponym</hvtyp2>
<hvtag>xnr202494</hvtag><hvord><hvhomo></hvhomo>kamper`a</hvord></hvrgrupp></hv>
<detaljer id="detaljer129693"><exempelblock><sbblock1x><syntex>de campade på västkusten i
somras</syntex></sbblock1x></exempelblock>
<valens><vt>ampa (ngnstans) </vt></valens><etymologiblock><etymologi><fb>sedan 1925 </fb><et>av eng.
<besl>camp</besl> med samma betydelse; till lat. <besl>cam`pus</besl> 'mark; fält; fält-läger'; jfr
<hvtag>lnr202468</hvtag><hvord><hvhomo>1</hvhomo>kamp</hvord> </et></etymologi></etymologiblock>
</detaljer>
</lexem>
<ordbilddn><ordbtxt>Subst. :</ordbtxt><ordbilddn><hvid>vb1d-
129692</hvid><hvord><hvhomo></hvhomo>campand</hvord>,
</ordbilddn><ordbilddn><hvtag>lnr129696</hvtag><hvord><hvhomo></hvhomo>camping</hvord>,
</ordbilddn><ordbilddn><hvid>vb1d3-129692</hvid><hvord><hvhomo></hvhomo>kampand</hvord>,
</ordbilddn><ordbilddn><hvid>vb1d4-
129692</hvid><hvord><hvhomo></hvhomo>kampning</hvord></ordbilddn></ordbilddn>
</lemma>
</artikel>

```

Figure 6: XML file sent to app development firm

```

SO_litteraturbanken_ver_1.18.xsd - Anteckningar
Aktiv Redigera Format Visa Hjälp
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" >
<!--
<xs:element name="so">
<xs:complexType>
<xs:sequence>
<xs:element ref="artikel" maxOccurs="unbounded"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="artikel">
<xs:complexType>
<xs:sequence>
<xs:element ref="lemma" maxOccurs="unbounded"/>
</xs:sequence>
</xs:complexType>
</xs:element>
<xs:element name="lemma">
<xs:complexType>
<xs:sequence>
<xs:element name="homonr" type="xs:token" minOccurs="0" maxOccurs="1" />
<xs:element name="grundform" type="xs:string" minOccurs="1" maxOccurs="1" />
<xs:element ref="ljudfil" minOccurs="0" maxOccurs="9" />
<xs:choice>
<xs:group ref="EjAlternativuttalEjBojning" minOccurs="0" maxOccurs="1" />
<xs:group ref="OjAlternativuttalBojning" minOccurs="0" maxOccurs="1" />
<xs:group ref="UttalFinsochAltuttalFinsEllerEjAlternativ" minOccurs="0" maxOccurs="1" />
<xs:group ref="AlternativFinsEjEllerEfterBoj" minOccurs="0" maxOccurs="1" />
<xs:element ref="hv0" minOccurs="0" maxOccurs="unbounded" />
</xs:choice>
<xs:element ref="kortform" minOccurs="0" maxOccurs="1" />
<xs:element ref="forkort" minOccurs="0" maxOccurs="1" />
<xs:element ref="lexem" minOccurs="0" maxOccurs="unbounded" />
<xs:element ref="ordbilddn" minOccurs="0" maxOccurs="1" />
<xs:element ref="sentens" minOccurs="0" maxOccurs="1" />
<xs:element ref="stilruta" minOccurs="0" maxOccurs="1" />
</xs:sequence>
<xs:attribute name="id" type="xs:ID" use="required"/>
<xs:attribute name="lnr" type="xs:NMTOKEN" use="optional"/>
</xs:complexType>
</xs:element>
<xs:group name="AlternativFinsEjEllerEfterBoj">
<xs:sequence>
<xs:element ref="ordklass" minOccurs="1" maxOccurs="1"/>
<xs:element ref="alt" minOccurs="0" maxOccurs="unbounded"/>
</xs:sequence>
</xs:group>
<xs:group name="UttalFinsochAltuttalFinsEllerEjAlternativ">
<xs:sequence>
<xs:element ref="uttal" minOccurs="1" maxOccurs="1"/>
<xs:choice>
<xs:group ref="EjAlternativuttalEjBojning" minOccurs="0" maxOccurs="1" />
<xs:group ref="OjAlternativuttalBojning" minOccurs="0" maxOccurs="1" />
<xs:group ref="AlternativuttalochBojning" minOccurs="0" maxOccurs="1" />
</xs:choice>
</xs:sequence>
</xs:group>
<xs:element name="alt">
<xs:complexType>
<xs:sequence>
<xs:element name="altkom" minOccurs="0" maxOccurs="1"/>
<xs:group ref="Fagrp" minOccurs="1" maxOccurs="2"/>

```

Figure 7: XML schema description

4. The SO dictionary app

It is a challenge to develop a dictionary app that not only accurately reflects the comprehensive printed dictionary from 2009 but that can also be regarded as an independent lexicographical resource. Release of the app is planned for late summer of 2015. As we have indicated, the app will be the result of close collaboration and a

working method characterized by flexibility, especially between the lexicographers and system developers in Gothenburg and the app developers at I solve AB in Stockholm. During the process, errors in the database and extraction programs have been identified and fixed and different app versions have been examined and tested by an extensive test group consisting mainly of lexicographers at the University of Gothenburg and employees at the Swedish Academy. The app has not yet been subjected to a user study. However, the editors and system developers have been able to draw some conclusions about app user behaviour thanks to the recently performed user study on the related project, the *Swedish Academy Glossary*, or SAOL (see section 2).

The primary target user groups of the printed version are native Swedish speakers and advanced learners. The dictionary is polyfunctional; it supports both receptive and productive user situations, while also fulfilling a documentary function. It has not yet undergone a user survey, but according to letters to the editorial board, it is mainly used by translators, proofreaders, language teachers, etc., i.e. people who work with the Swedish language on a professional level. However, as a result of the app format, the dictionary has the potential to reach a wider user group, for example younger and less experienced dictionary users. Taking this into account, it is very important to avoid information overload; the lexical data must be presented in a way the users can handle, otherwise it might hinder the retrieval of information needed (cf. Tarp, 2012: 255 and Lew, in section 2).

4.1. Platforms

The SO app is designed for iOS and Android. This decision was based on the dominant position of the iPhone and Android operating systems in the Swedish mobile market. Furthermore, statistics on the number of downloads of the dictionary app of the SAOL glossary show that those platforms are the most common among Swedish users. The results of the survey of active SAOL app users also support this idea.

The SO app is a hybrid of an online and offline resource. Although the app allows users to look up words and to view definitions and examples offline, a network connection is required to access audio pronunciations. The main content of the app is static, but some of the information can be updated as soon as an internet connection is available, e.g. the sections concerning the Swedish Academy, related apps, etc.

4.2 User interface – layout

The user interface described here is the current version of the iPhone GUI. The iPad GUI is similar to the iPhone GUI, but we take advantage of the larger screen size. For example, the list of entries and the current entry are shown on the same screen on iPads. When this paper was written, development of the Android version of the app had only just begun.

The main screen of the app contains the status bar, a navigation control, a search bar, a few standard icons and a text field (the content of which depends on the current activity). The user can choose between searching for a word (*Sök*), bookmarks (*Bokmärken*), search history (*Historik*), usage guidelines for particular words (*Stilrutor*), Word of the Day (*Dagens ord*), news (*Aktuellt*), general information about the app, the Swedish Academy, help, etc. and giving feedback (*Återkoppling*).

The default mode is *Search*, and the user is presented with a list of dictionary entries beginning with **A** when the app opens.

Compared with other dictionary apps, such as those developed by the Society for Danish Language and Literature, a considerable amount of information is provided concerning the dictionary content. The user instructions are also relatively comprehensive. According to Svensén (2009: 459), “it is a truth universally acknowledged in lexicographic circles that user’s guides are very seldom consulted”, but the user survey on the SAOL app shows that a relatively high number of the respondents consulted this information in the glossary app. With the aim of developing the SO app to be as independent as possible from its analogue predecessor, we find it important for users to be able to get all of the information on lexicographic content in the app without being forced to consult the book.



Figure 8: Drawer menu in the SO dictionary app

The *Stilrutor* (‘style boxes’) screen shows a list of all dictionary entries that include usage guidelines. For example, in the list you can find the entry *genitiv* (‘genitive’) followed by instructions on correct usage of genitive apostrophes in Swedish. In comparison with the printed dictionary, these guidelines have enhanced visibility and are more easily accessed in the app, since they can be found not only when you happen to look up a particular word, but also through the action item *Stilrutor*.

In the *Historik* screen (‘search history’), shortcuts to recent lookups are listed chronologically. The *Bokmärken* (‘bookmarks’) screen contains a similar list of shortcuts to bookmarked dictionary entries.

The intention behind introducing headwords as *Dagens ord* (‘Word of the Day’) is to provide a sample of the comprehensive content of SO. It is hoped that the selected entries will serve as illustrative examples of entries and pique the user’s interest in delving more deeply into the dictionary content. They may also stimulate vocabulary building and support language acquisition, especially among learners of Swedish. In some dictionary apps, such as Dictionary.com, the *Word of the Day* seems to be selected at random. In the online working Danish DDO app, the headword is continuously refreshed, resulting in a dynamic, fresh “look” (see Holmer & Sköldbberg, 2014).

In the SO app, the *Word of the Day* display does not require internet access. The selection of words is made in advance by the lexicographers. A list of entries is prepared for a period of one year and the entries are set for specific dates. Some of the selected entries are closely connected to a certain season (e.g. *krokus*, ‘crocus’) or a Swedish feast day (e.g. *påskägg*, ‘Easter egg’). Furthermore, some words on the list derive from Old Swedish, e.g. the noun *dag* (‘day’), which dates from the 9th century, and others are relatively new loanwords (e.g. *sudoku* ‘sudoku’). Many are also conspicuously metaphorical (e.g. *flaskhals*, ‘bottleneck’, *bokslukare*, ‘book swallower’, i.e. a voracious reader). With this feature, we thus hope the *Word of the Day* in the SO app will increase interest in the Swedish language and its vocabulary in general.

The Swedish verb *campa* (‘to camp’) is presented in Figure 9.

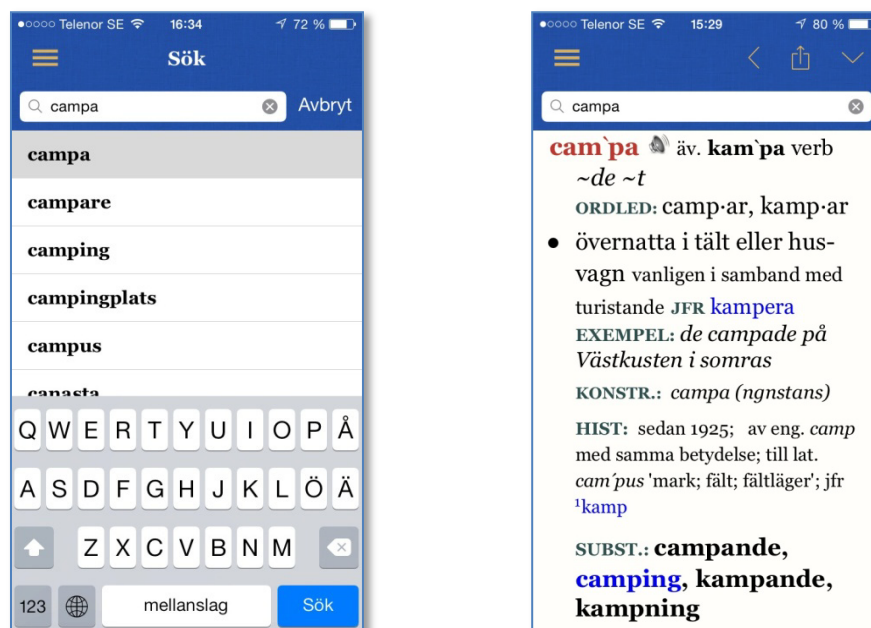


Figure 9: The list of entries as shown when a word is typed into the search bar and the content of the entry **campa** (‘to camp’) as shown when selected by the user

As shown in Figure 9, the user interface of the app is mainly blue, white and black. The headwords are presented in red to make them stand out. We intend for the interface design to be perceived as stylistically clean and aesthetically pleasing. The aim of the design is also to connect the app to the deep blue cover of the printed version of the SO and thus make the app seem familiar to people who have previously used the printed dictionary. At the same time, the connection with the related app, the SAOL (also financed by the Swedish Academy), must be evident. In light of these facts, the icon, which is blue and includes the classic Academy emblem – a laurel wreath surrounding the Academy motto “Snille och smak” (‘Talent and Taste’) – was chosen.

The default text size in the SO app is comparable to the text size in other dictionary apps. Hopefully users will successfully hit the touch zones on the screen and the keyboard even when on the move, which is not unusual in app usage (see Simonsen in section 2). In order to meet the needs of different users and user situations, the display of the article content is adapted to a smaller or bigger font size in response to standard pinch-in/pinch-out zooming gestures. This is made possible in the SO app thanks to built-in soft hyphens, which allow for dynamic word wrapping (see section 3).

An additional aspect of the app design is that users can switch between portrait and landscape modes simply by turning the mobile or tablet (cf. for example, dictionary apps by Longman and Merriam-Webster in this respect). Even though users have personal preferences for portrait or landscape, their choices are also affected by their situations. For example, people may tend to watch video in landscape mode and read in portrait. It is important for users to be able to make their own choice and individualize their usage. In relation to this function, the use of soft hyphens is also important; without them, the right margin of entries could appear ragged.

4.3. Content and search functions

The entire lexicographic content of the printed version of the SO is included in the app. In this regard, the SO app is quite different from the Danish DDO app, which only provides a sample of the content found in the web version (see Holmer & Sköldbberg, 2014). If users wish to see everything in that lexicographic resource, they are obliged to consult the dictionary site ordnet.dk (easily accessed via links in the app). It could be argued that the links in the DDO encourage users to go to the online version of the dictionary. As a result, the app could be regarded as a spin-off of the web version. But, as previously mentioned, the DDO app is a highly effective lexicographic product, so this is not the case. A similar model in the SO app with external links to an online version is not possible because the Swedish Academy decided to prioritize the dictionary app rather than an online version (see section 3).

An addition to the lexicographical content of the SO is the inclusion of approximately

65,000 human-read audio files. These files constitute an important aid for users, especially learners of Swedish. Access to audio pronunciation will probably also increase interest in the dictionary among native Swedish speakers. However, the integration of audio pronunciation also raises new issues. When users can listen to pronunciations of the headwords, will phonetic transcription – information that many users have difficulties interpreting without consulting the pronunciation key – still be necessary? (cf. Svensén, 2009: 383). In such an online resource there is plenty of presentation space; but presentation space is very restricted on a mobile screen (see Lew in section 2). We have decided to keep the phonetic transcriptions in the dictionary app for two main reasons. As Lew (in press) points out, learners of Swedish may not be able to hear phonemic distinctions since their perception is filtered through the phonological system of their native language. Moreover, as mentioned, audio pronunciation is only accessible when the user is connected to an internet network.

A common, simple search is performed by starting to type the sought-for word. The list of matching entries adjusts as the user types. The headword is shown at the top of the list followed by the rest of the dictionary headwords (compared to, for example, the DDO app, which only shows the next 29 headwords). Like in most dictionary apps, it is possible to scroll up and down in the lemma list. This function is essential for people who want to gain an understanding of nearby headwords, something that is of course simple in a book. By clicking a lemma in the list, the whole entry is shown. The search algorithm also supports searches for inflected forms (algorithm developed by Isolve AB).

Users can additionally perform phrase searches. The algorithm is the same as for a simple search. The search string “kalla fötter” (‘cold feet’) generates the idiom variants *få kalla fötter* (‘get cold feet’) and *ge ngn kalla fötter* (‘give someone cold feet’) (see Figure 10). But the search string also generates other results from other entries in the dictionary containing the word forms *cold* and *foot*, for example, a syntactical example in the entry *doppa* (‘dip’): “The water was so cold she only dipped her feet”. The word forms in the search string are distinguished in the hits with bold typeface. In each example, information on the entry (in blue) and the information category (in grey) is given. In this particular case, users are informed that the idiom *få kalla fötter* is placed under the noun *foot* (‘foot’). They can also easily check the entry in question, as it constitutes a cross-reference. The lexicographers decided which types of phrases were to be indexed and used for this search. In short, idioms and other fixed phrases that include two or more word forms given in the search string are presented at the top of the list because it is reasonable to assume that this is the multi-lexical unit that users want in most cases.



Figure 10: Result of a phrase search in the SO dictionary app

There is also a spell-check function developed by Isolve AB.

The SO app also supports wildcard search. A search string like “*boll*” (*ball*) generates hits such as *bollhav* (‘ball pit’), *fotboll* (‘football’) and *snöbollseffekt* (‘snowball effect’). These kinds of searches may appeal to scholars. Considering that there is no online version of the SO, the app may be used to perform different kinds of lexicological studies. This function may also very well appeal to users interested in solving e.g. crosswords.

When it comes to article microstructure, the users of the printed SO will probably find the layout familiar. The italics and different type sizes are still there. However, the printed version of the SO, like many other dictionaries, is characterized by compression (cf. Lew, in press). With the aim of making the entries and information more accessible to users, more headings are included and many of the abbreviations are dissolved and shown in full text, for example, for the part of speech, the tilde used to mark the lemma in the entry text is replaced with the lemma, etc. Even though the display area on a mobile device is very limited, we find this an important consideration by the users, especially learners of Swedish.

4.4. Collapsing and expanding

To make extensive dictionary articles clearer and the dictionary content easier to grasp, longer entries in the SO app are shown in a collapsed form. See Figure 11 for examples of a collapsed and an expanded version of the noun *harmon*i (‘harmony’).

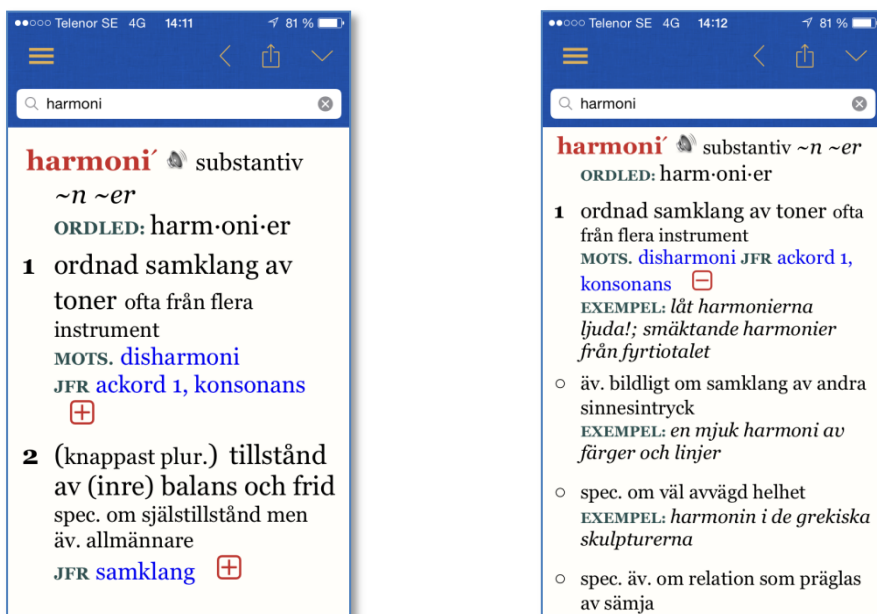


Figure 11: Two versions of the entry **harmoni** (‘harmony’), collapsed version (left) and extended version (right).

A relevant question is what is considered to be a “long” or “short” entry (cf. Trap-Jensen, 2010). We have chosen to collapse dictionary entries that display over more than one screen size (iPhone 5). In the development process we have been experimenting with the optimal amount of data presented by default using the HTML prototype (see Figures 4 and 5). Tarp (2012) states that the problem concerning individual entries on the screen is not only how much *can* be presented at a given time to a dictionary user, but also how much *should* be presented. In the present app version, details that belong to separate core meanings are hidden if all data cannot fit onto one screen so that the user gets a clear overview of the semantic structure. We present the following information categories in the collapsed view: headword, pronunciation, part of speech, inflected forms, definitions (of core meanings) and related words (like synonyms and antonyms). By touching the expansion symbol (the plus sign), users can access subordinated meanings, idioms, information on valency, etymology, etc. Lew (in press) concludes that user research is needed to establish what content should be displayed immediately on the screen, and what content should be deferred. We hope to perform such a study when the app has been on the market for a while.

4.5 Cross references and hyperlinks

The printed dictionary contains a considerable number of references to specific meanings of related words. For comparison, Figure 11 shows links in blue. These references have been implemented as hyperlinks in the app by means of supplementing XML tags with ID/IDREF attributes. During this process, a number of faults in the database were brought to light, such as references to words that had been excluded in print. Finding these kinds of errors is common to any IT

development project and must be taken into account when embarking on an appification project.

5. Final remarks

In this paper we present the ideas behind a new Swedish dictionary app, which we hope will reflect the comprehensive *Contemporary Dictionary of the Swedish Academy* (the SO), 2009. We present the lexical database that has been evolving since the mid-1960s and which has resulted in numerous scientific reports, printed dictionaries, internal web interfaces and finally a dictionary app. We also highlight strategic considerations for optimising the layout and presentation of the database content so that it fits the app display while retaining as much as possible the look and feel of the physical book.

The SO app will cost 49 Swedish kronor (about 5 euros), which is competitive compared to the printed dictionary, which costs about 500 Swedish kronor (a bit more than 50 euros). The price of dictionary apps on the Swedish market, such as Norstedts, range from 49 Swedish kronor for the smaller ones to 390 Swedish kronor (40 euros) for the most comprehensive bilingual Swedish/English dictionary. Thus, the SO app can be considered heavily subsidized. It is well-known to lexicographers that dictionary projects are expensive, take a long time and are never really finished. But as the SAOL app user study has shown, many users are not really willing to pay for dictionary apps, and even if they are, they are not prepared to pay very much. Even though the Swedish Academy could in theory give away the app for free, taking the decision to charge a small amount reflects a desired position concerning high quality lexicographical products.

This article aimed to participate in a broader discussion of experiences involved with producing dictionary apps, app development and app user behaviour. In this paper, we focussed on the mobile phone app since the tablet app is somewhat different. The app presented here is planned to be released in late summer of 2015. This app has been tested by an extended test group, but has not yet been the object of a user study *per se*. So far, the SO printed dictionary has not been researched from a user's perspective either. However, the editors and system developers were able to draw some conclusions regarding app user behaviour as a result of a user study on a related project, the *Swedish Academy Glossary* (SAOL) that was just carried out in March 2015.

It is possible to approach the use of online dictionary apps with log files and statistics. App developers who want to gain insight into user behaviour with offline dictionary apps may be supported by mobile app measurement and advertising platforms like Flurry Analytics from Yahoo! (<http://www.flurry.com/>). By implementing Flurry in the SO app, the lexicographic team and app developers can gain deeper understanding of app user behaviour through the analysis of usage data, such as lookups, session duration, operative systems, device models, etc. Flurry will

be implemented in the SO app and knowledge about how the app is actually used will be invaluable when preparing updates and improving future versions.

6. References

- Dictionary.com*. Dictionary.com, LCC. App version 5.2.1. (Accessed 25 May 2015)
- DDO: *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab. App version 2.0.11 (iOS). (Accessed 25 May 2015)
- Gao, Y. (2013). The Appification of Dictionaries: From a Chinese Perspective. In Kosem et al. (eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*. Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 213–224.
- Hoel, J. (2012). Appsolutt fingerferdig! En anmeldelse av ordbokappene RO og SAOL. *LexicoNordica* 19, pp. 255–271.
- Holmer, L. (2011). Norstedts ordboksappar. *LexicoNordica* 18, pp. 307–322.
- Holmer, L, Hult, A.-K. & Sköldberg, E. (2015). Spell-checking on the fly? On the use of a Swedish dictionary app. Proceedings of eLex conference 11-13 Aug. 2015.
- Holmer, L. & Sköldberg, E. (2014). Appifiering till allas lycka? Om danska ordboksappar med särskilt fokus på DDO. *LexicoNordica* 21, pp. 235–252.
- Lew, R. (in press). Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries. In P. Bański & B. Wójtowicz (eds.) *Issues in Modern Lexicography*. München: Lincom Europa.
- Malmgren, S.-G. (2009). On production-oriented information in Swedish monolingual defining dictionaries. In: S. Nielsen & S. Tarp (eds.) *Lexicography in the 21st century. In honour of Henning Bergenholtz*. Amsterdam/Philadelphia: John Benjamins, pp. 93–102.
- Malmgren, S.-G. & Sköldberg, E. (2013). The lexicography of Swedish and other Scandinavian languages. *International Journal of Lexicography* 26(2), pp. 117–134.
- Marello, C. (2014). Using Mobile Bilingual Dictionaries in an EFL Class. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014*. Bolzano/Bozen, pp. 63–83.
- Merriam-Webster: *Merriam-Webster Dictionary*. Merriam-Webster Inc. App version 3.2 (iOS). (Accessed 25 May 2015)
- Rundell, M. (2013). Redefining the dictionary: From print to digital. In *Kernerman Dictionary News* 21. Available at: <http://kictionaries.com/kdn/kdn21.pdf>.
- Simonsen, H. Køhler (2014a). Brugerne er allerede mobile! In R. Vatvedt Fjeld & M. Hovdenak (eds.) *Nordiske studier i leksikografi* 12. Oslo: Novus, pp. 416–429.
- Simonsen, H. Køhler (2014b). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014*. Bolzano/Bozen, pp. 249–261.
- SAOL13: *The Swedish Academy Glossary*. (2006). 13th edition. Svenska Akademien &

- Isolve AB. App version 1.1.8 (iOS). (Accessed 25 May 2015)
- SO: *Svensk ordbok utgiven av Svenska Akademien*. (2009). Stockholm: Norstedts.
- Svarverud, R. (2014). Nye kvalitetsverktøy for brukere av kinesisk i Skandinavia. *LexicoNordica* 21, pp. 341–356.
- Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- Trap-Jensen, L. (2010). One, Two, Many: Customization and User Profiles in Internet Dictionaries. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Leeuwarden: Fryske Akademy, pp. 1133–1143.
- Tarp, S. (2012). Online dictionaries: today and tomorrow. *Lexicographica* 28, pp. 253–267.
- Törnqvist, L. (2010). Ordböcker på Internet och Internet som ordbok. In H. Lönnroth & K. Nikula (eds.) *Nordiska studier i lexicografi* 10. Tammerfors, pp. 484–493.
- Winestock, C. & Y.-k. Jeong (2014). An analysis of the smartphone dictionary app market. *Lexicography Journal of ASIALEX* (2014(1)), pp. 109–119.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Towards an Electronic Specialized Dictionary for Learners

Marjan Alipour, Benoît Robichaud, Marie-Claude L'Homme

Observatoire de linguistique Sens-Texte (OLST), Département de linguistique et de traduction,
Université de Montréal, Montréal (Québec), Canada

Email: marjan.alipour@umontreal.ca, benoit.robichaud@umontreal.ca, mc.lhomme@umontreal.ca

Abstract

This paper describes the strategies devised in order to convert the *DiCoInfo*, *Dictionnaire fondamental de l'informatique et de l'Internet*, a specialized lexical database, into a learners' dictionary. Our main goal is to obtain a user-oriented dictionary (i.e. that meets specific user needs). Firstly, we defined the types of users towards which our dictionary is targeted: translation students are our first intended users. Then we determined the use situations and the functions of our dictionary: it should provide assistance in communicative and cognitive situations (Tarp, 2008). We made several changes to adapt the data categories of the DiCoInfo to these functions and user needs. In addition, we simplified the presentation: layout, display of data categories, access to data and addition of multimedia. In this user-oriented version, the data is presented in such a way that users who do not have a background in linguistics can easily interpret the contents of the data categories. Finally, different technologies were integrated in the process and hopefully contribute to make the new version even more accessible.

Keywords: electronic dictionary; learners' dictionary; specialized dictionary; dictionary functions; user needs

1. Introduction¹

Many studies on online general learners' dictionaries contribute to better understanding the needs of users and to design more efficient reference tools (Dziemianko, 2010; Lew, 2012; Lew & de Schryver, 2014). However, little research has focused on specialized electronic dictionaries and few specialized dictionaries for learners have been published up to now (a few notable exceptions are Pyne & Tuck, 1996 and Binon et al., 2000). We believe that students studying translation and technical writing require dictionaries to help them in vocabulary acquisition, but also to assist them when reading, translating or producing specialized texts. However, many questions remain unanswered: What are the properties of a specialized learners' dictionary? What should a specialized learners' dictionary look like in order to meet specific user needs?

This paper describes the method developed in order to convert an existing specialized lexical database into a learners' dictionary taking into account specific categories of users

¹ This work was supported by the Fonds Société et Culture of the Government of Québec. The authors would like to thank the reviewers whose comments helped clarify some parts of the paper.

and predefined use situations. Furthermore, we devised different strategies to present the data in a more user-friendly and simple way.

The lexical database for which this work was undertaken is the *DiCoInfo*, *Dictionnaire fondamental de l'informatique et de l'Internet* (hereafter DiCoInfo), a multilingual database that contains basic terms from the fields of computing and the internet. In previous work, user-friendly displays and access routes were designed for specific data categories (collocations, Jousse et al., 2011; actantial structures, L'Homme, 2014b). However, this work affected only parts of the articles. The new interface described herein is based on a work carried out by Marjan Alipour in her Master's dissertation (Alipour, 2014) who analyzed the entire structure of the DiCoInfo and devised a user-oriented dictionary based on the theory of lexicographical functions (Bergenholtz & Tarp, 2003; Tarp, 2008). We also took the opportunity to explore the potential of using new technologies to ensure that our user-oriented and user-friendliness objectives were met.

The paper is organized as follows. Section 2 gives a brief description of the contents of the DiCoInfo. Section 3 gives more details about the types of users we target and the cognitive and communicative situations that the DiCoInfo is now designed to meet, and describes the rationale behind each change made to the original interface for creating a user-oriented version.

2. The DiCoInfo

The DiCoInfo is an online specialized resource that contains English, French, and Spanish terms related to computing and the internet [En. *browse*, *configuration*; Fr. *naviguer*, *configuration*; Es. *navegar*, *configuración*]. It describes terms that belong to various parts of speech: nouns [*email*, *printer*], verbs [*download*, *print*], adjectives [*dynamic*, *virtual*] and adverbs [*dynamically*, *online*]. Currently, the DiCoInfo contains approximately 1,100 entries in French, 850 entries in English, and the Spanish version is under development. The content data is encoded in XML files (stored in an eXist database) and converted using customized XSLT stylesheets into HTML pages so that it can be published on the Internet (Jousse et al., 2011).

Articles that are completed have the following data categories (L'Homme, 2014a, b):

- *Headword*: The lemma associated with a sense number.
- *Grammatical information*: The part of speech, along with gender (for nouns in Spanish and French) and government pattern (for verbs).
- *Status*: The degree of completion of the entry, the editing is completed or still ongoing.
- *Actantial structure (AS)*: The actants and their semantic role are defined.

- *Definition*: A statement of the meaning of the headword, where actants (labeled with semantic roles) are highlighted with different colors.
- *Synonyms and variants*.
- *Contexts*: Three sentences are displayed to show how the term is used in specialized texts. In some entries, up to 20 contexts are annotated and users can access them on demand.
- *Lexical relations*: A list of terms that share paradigmatic relations (antonyms, other parts of speech, derivatives, etc.) and some syntagmatic relations (those that are described in the category labeled *Types of*).
- *Combinations*: A list of terms that share syntagmatic relations with the headword (mostly verbal collocates).

The DiCoInfo is original when compared with other specialized dictionaries since most of them are conceptual in nature and give encyclopaedic information (for instance, the *Dicofr.com* provides definitions and, in some cases, additional explanatory notes). Resources seldom provide information on syntagmatic and paradigmatic relations between terms of the domain. Unlike these resources, the DiCoInfo provides a complete description of the lexico-semantic properties of terms. In addition to providing definitions², the DiCoInfo supplies information about their linguistic behaviour, such as a statement of the actantial structure in which the semantic actants are labeled with a system of semantic roles (Agent, Patient, etc.) and typical terms (L’Homme, 2010; 2014a, b).

Example: Actantial structure³ for *keyboard*

a keyboard: ~ used by **user**_{1{Agent}} to act on **command**_{1{Patient}}, **data**_{1{Patient}}

In addition, as was mentioned above, the DiCoInfo describes the multiple relationships between terms, which can be paradigmatic (e.g. synonyms or near-synonyms [Ex. *browse: surf*], antonyms [Ex. *download: upload*], word families [Ex. *boot: bootable, reboot*], etc.) or syntagmatic (i.e. collocations [Ex. *document: save a ~; attach a ~*]) (L’Homme, 2010). These relationships are encoded with lexical functions (LFs) based on Explanatory and Combinatorial Lexicology (Mel’čuk et al., 1984–1999; 1995) and further described with a natural language explanation.

² Several French entries contain definitions. In English, this data category is available only for approximately 100 terms for the time being.

³ In the DiCoInfo, two systems are used to label the actants. First, a typical term is supposed to be indicative of the kinds of terms that can be used to instantiate an actant. Then, semantic roles (such as Agent, Patient, Destination, Instrument) indicate the relationship between the actant and the term. When users hover the mouse over a typical term (e.g. *user*₁) in the definition or the actantial structure, a tooltip pops up to show its role (e.g. Agent).

Finally, a set of sentences (up to 20) are extracted from specialized corpora and added to entries. These sentences are annotated based on the methodology developed in FrameNet (Ruppenhofer et al., 2010). Annotated contexts allow users to visualize how headwords combine with actants (and also non-obligatory participants) in real texts.

Originally, the DiCoInfo was designed as a research tool for exploring the potential of lexical semantics frameworks to account for the linguistic properties of terms. Little effort had been made to adapt it to user needs. Later on, work was carried out to simplify the presentation of specific data categories, namely collocations and actantial structures (Jousse et al., 2011; L’Homme, 2014b). This previous work showed that we could take advantage of the contents of the entries while presenting parts of them in a more user-friendly way. In addition, we could change the way the data is presented without affecting the initial structure of the database entries or the encoding methodology followed by lexicographers. However, we did realize that much more could be done to simplify the presentation of entries (change the overall display of data categories, keep the linguistic metalanguage in the background, take advantage of new technologies, etc.).

All these characteristics certainly contribute to making the former version of the DiCoInfo a rich resource. First, terminologists and lexicographers browse it to explore the linguistic properties of terms and use it as a means of formalizing hypotheses on them. We also believe it could prove useful for other users, such as translators, whose work often requires access information on the behaviour of terms in specialized texts (L’Homme, 2014a). But is all the information supplied in the DiCoInfo relevant for non-expert users who do not necessarily have a background in linguistics or in lexicography? Is the presentation of the data adapted to their needs? In fact, we think that the data contained in the DiCoInfo can be useful for students in translation and technical writing since it describes the functioning of terms in texts. However, we also believe that some data should be presented in a different way in order to facilitate their understanding and increase their usability. Next, what about the metalanguage used in the DiCoInfo? In fact, this metalanguage can be quite opaque for users such as translators. For example, lexical functions (LFs) are represented with labels that can be difficult to decipher for anyone who is unfamiliar with them, not mentioning the fact that some labels may be very complex (e.g. **IncepReal**₁: “to start using”; **FinReal**₁: “to stop using”; **Caus**₁**Able**₁**Func**₀: “cause something to be able to occur”). An alternative solution was required for this metalanguage in order to make the dictionary more user-friendly and efficient for our types of users. The strategies devised for this purpose are described in the next section.

3. Strategies Developed for the Conversion of the DiCoInfo

To develop our conversion method, we first determined the types of users and use situations of our dictionary based on the system of lexicographical functions (Bergenholtz & Tarp, 2003; Tarp, 2008; Fuertes-Olivera et al., 2012). We then explored different ways to adapt the data categories of the DiCoInfo to these functions. One of our objectives was to use all the information available in the resource, but present it in such a way that would readily meet the needs of specific users. Finally, we used various available technologies (mostly from the *jQuery UI* framework, Sarrion, 2012) to implement these changes in the new version interface that we think is now more dynamic and responsive.

3.1 Types of Users and Use Situations

The dictionary is intended for French, English and Spanish users who are not experts in the domain of computing and the internet. More specifically, the main targeted users are, on the one hand, translation students and translators who have little experience in this field; and terminologists or terminographers, on the other hand. Other users such as proofreaders and technical writers are also targeted.

We aimed to design a learners' dictionary that could provide help for understanding, producing, or translating specialized texts: these situations are related to communicative situations as defined in Tarp (2008). In addition, the dictionary should be helpful for acquiring knowledge about factual or linguistic matters related to the lexicon of computing. This latter situation corresponds to cognitive situations as defined in Fuertes-Olivera and Nielsen (2012). These functions (presented in more detail below) are based on previous work by Leroyer (2013) who defined lexicographical functions for the former version of the DiCoInfo.

1. Communicative Functions and Use Situations

- Translation of texts: In this situation, the dictionary should assist with translating technical terms and collocations. For example, users who want to translate a text about browsers from English into French may look up the entry *browser*. Then, not only does the DiCoInfo provide a French equivalent, i.e. *navigateur*, it also provides translations for the word family [Ex. *to browse*: *naviguer*; *browsing*: *navigation*], and for different kinds of browsers [Ex. *user-friendly browser*: *navigateur convivial*]. In addition, the information helps users to correctly handle collocations [Ex. *run a browser*: *lancer un navigateur*].
- Reception of texts: In this situation, the dictionary should help users solve problems related to the understanding of terms and expressions while reading texts.

For example, while reading a text on *cables*, users might have to distinguish between a *female connector* and a *male connector*.

- Production of specialized texts: In this situation, the dictionary assists users in solving problems while producing texts. Thus, they can learn how to express an idea correctly by using the exact collocation. For example, they will learn how to produce a phrase with a specific verb and select the right preposition (Ex. *connect a computer to the internet with a cable*).
- Editing and proofreading texts: In this situation, the dictionary can help solve problems that arise while editing or proofreading a text. Users, for example, may identify an erroneous usage of a word or a collocation with the help of information supplied by the dictionary. Thus, if the collocation *disconnect from the Internet* is translated into French as *déconnecter de l'Internet* (that contains errors in the verb usage and the structure of the collocation), they will be able to correct it to *se déconnecter d'Internet*.

2. Cognitive Functions and Use Situations

- Learning terminology of computing: In this situation, users can browse the dictionary in order to acquire knowledge about linguistic matters related to the field of computing.
- Systematic study of the field of computing: In this situation, users can consult the dictionary in order to meet occasional information needs, for preparing a translation for example.

3.2 Changes Made in the DiCoInfo

Once the functions of the dictionary were determined, we then compiled a list of changes to be made to obtain a user-oriented version. The modifications were suggested according to two parameters: simplifying the presentation and ensuring that the functions of the dictionary were fulfilled. After analyzing the former version of the DiCoInfo, we identified two broad categories: 1. Information that already meets the targeted user needs as defined in Subsection 3.1, and thus that should be kept as is; and 2. Information that should be used but displayed in a different way or placed in the background. Modifications were made at several levels: to the interface and its layout, to the data categories, and to the organization of data inside data categories. It is worth mentioning that all the changes mentioned in this paper apply without distinction to all language content, but that some data category contents in English and Spanish have not yet undergone all the changes. Hence the examples given are in French; English translations are provided when possible.

3.2.1 Changes Made According to the First Parameter: Simplification of the Presentation

a. The Homepage

Since the DiCoInfo is designed as an online dictionary, we were able to take advantage of various electronic media for presenting and organizing the data in a clearer and more user-friendly way. The interface of the former DiCoInfo was basic; therefore, efforts were made to improve the attractiveness, simplicity and conciseness of the new version (Figure 1).

b. The Search Interface

In the new version, a much simpler search field than that of the former version was implemented (as can be seen in Figure 1).



Figure 1: Homepage of the former version (above) and new version (below)

An auto-completion search field was added: when two characters are entered, a list of suggestions corresponding to the terms of the DiCoInfo is displayed. Users then select the term they are looking for and the system retrieves the corresponding entry. The interface still provides the possibility to filter the search results by means of options, but in the new version, icons are used to group and present them. Therefore, users can narrow down the search results according to the language, the search mode (a term, a lexical relation, etc.), or the precision level (exact term, term beginning with a specific substring, expression containing substring, etc.). A simple click on the corresponding icon is required to display the options (Figure 2).

c. The Content Layout

In the new version of the DiCoInfo, the interface was adapted to make it more intuitive; data categories are now presented on tabs, a mode that appears to be preferred by users (Müller-Spitzer et al., 2012). These tabs are organized according to data categories along ribbons (Figure 3). Users can navigate easily from one tab to the other to obtain the information they need according to specific use situations. In addition, to allow users to readily visualize what information is contained in each tab, we changed some of the tab names that were rather technical and could be confusing. For example, *Autres parties du discours et dérivés* (En. Other parts of speech and derivatives), was changed into *Famille de mots* (En. Word family).

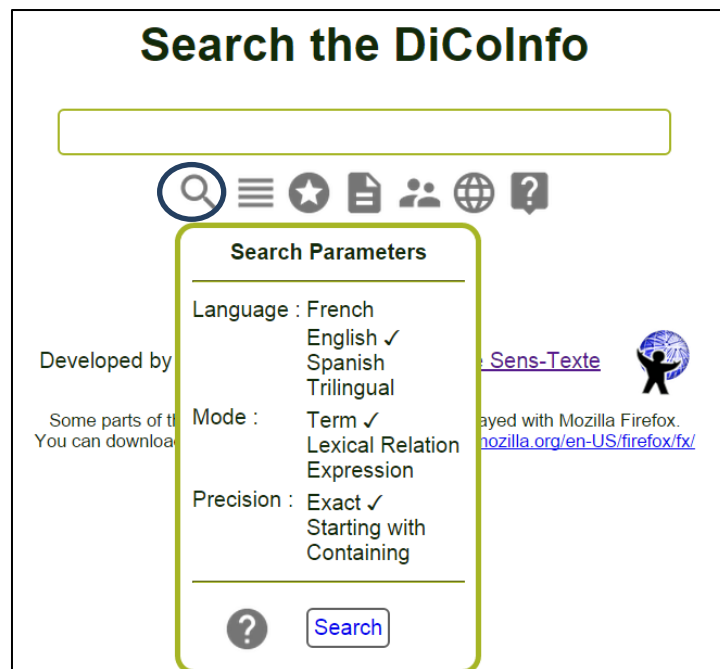




Figure 2: Search options

d. Other Features

In order to ease the search process, we implemented a help dialogue explaining all the options by means of a help icon  (Figure 2). In addition, to make the content data more readable and understandable, we added information dialogues on each data categories ribbon. A simple click on the corresponding information icon  displays an explanation about the specific data category (Figure 3).

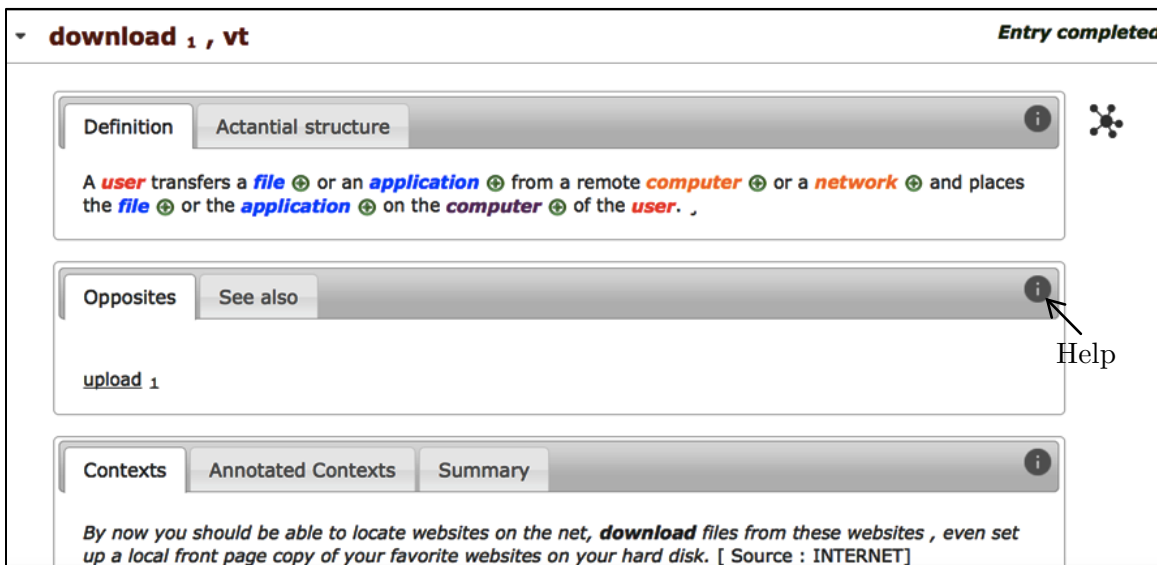


Figure 3: System of tabs in the new version

3.2.2 Changes Made According to the Second Parameter: the Functions of the Dictionary

In this section, the changes made according to the lexicographical functions of the DiCoInfo (described in Subsection 3.1) are explained.

a. Modifications in the Presentation of Data Categories

Since we wanted users to find answers to different problems related to communicative or cognitive situations quickly and efficiently, the presentation of certain data categories was revised.

- *Headword*

The presentation of the headword has changed (Figure 4). The information that is considered essential in communicative situations is summarized when entries are first retrieved. Figure 4 shows the summary given after the search for *Web*.

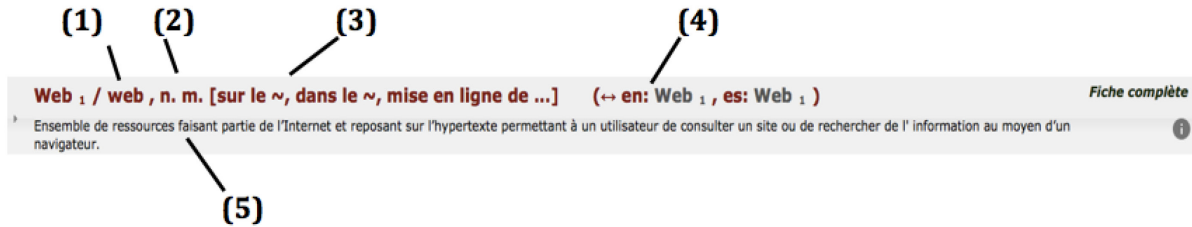


Figure 4: Summarized headword display

As seen in Figure 4, the following information is provided: variant forms of the headword [*Web/web*] (Figure 4:1), grammatical information (Figure 4:2), constructions with prepositions [*Web: on the ~*] (Figure 4:3), and translation equivalents (Figure 4:4). In fact, the variant forms, the constructions with prepositions, and the equivalents are presented immediately so users do not spend time searching for them inside the articles. The constructions with prepositions information, for example, allow translators to see immediately what are the typical prepositions to use with a specific term. The definition (or the actantial structure when no definition is yet available) is also shown to present the meaning of the term (Figure 4:5).

Former version →

clavier ₁, n. m.

un clavier : ~ utilisé par *utilisateur* ₁ pour intervenir sur *commande* ₂, *données* ₁ ⚙

Définition : Périphérique utilisé par un *utilisateur* pour entrer des *données* ⚙ dans un ordinateur ou envoyer des *commandes* ⚙ à l'ordinateur.

Synonyme(s) : clavier d'ordinateur, clavier informatique

Contextes

Liens lexicaux

Combinatoire lexicale

New version →

clavier ₁, n. m. [au ~]

Définition Structure actancielle ⓘ

Périphérique utilisé par un *utilisateur* pour entrer des *données* ⚙ dans un ordinateur ou envoyer des *commandes* ⚙ à l'ordinateur .

Synonyme(s) Voir aussi ⓘ

clavier d'ordinateur, clavier informatique

Contextes ⓘ

Pratiquement tous les portables disposent de connecteurs qui permettent de recevoir un moniteur et un clavier en externe . [Source : PCEX96]

*Si les paramétrages font ici généralement appel à des combinaisons de touches et si le **clavier** apparaît un peu mou à notre goût, l'ensemble se distingue par une ergonomie poussée [Source : PCEX96]*

*La partie la plus à gauche est le bloc des touches de fonctions . Celles-ci sont numérotées de {F1} à {F10}. À l'autre extrémité on retrouve le **clavier** numérique ou le **clavier** curseur (touches de déplacement du curseur). [Source : MSPDOS1]*

Famille de mots Sortes de Autres liens ⓘ

Utiliser / Ne pas utiliser ⓘ

Figure 5: Data category display for *clavier* (En. *keyboard*) in the former version (above) and the new version (below)

- *Data category display*

We also considered the way data categories should be displayed on the page. Once again, decisions were made according to the usability of data in communicative and cognitive situations, and users' profiles. Thus some changes were made in the new version (Figure 5).

The data categories *Definition*, *Synonyms/Opposites* and *Context* are opened by default. The reason for this is to provide some assistance to users who might be unsure about which term to use in a specific context (e.g. in case of synonymy), and how to use it. These data can help them to understand, produce or translate a text (communicative situations). They can also become familiarized with the meaning of terms (cognitive situation). In addition, these data categories do not contain a lot of information, which would otherwise overload the page. Thus we decided to make them appear opened by default.

Some questions arose about the way that the *Actantial structure* data category was presented in the former version (see Section 2): the actants being already available in the definitions, this data category became somehow redundant. In addition, as our users are not expert in linguistics, the way the statement was displayed (with actantial roles and isolated typical terms) might be confusing. After careful consideration, we opted to keep it as a formal alternative to the definition. Thus the tab presenting the actantial structure is placed on the same ribbon as the *Definition* data category and is presented opened to users only if no definition is available. Otherwise, it is placed in the background, so users are able to access it if necessary⁴ (Figure 6).



Figure 6: *Definition* and *Actantial structure* display for *blogue* (En. *blog*) in the former version (above) and new version (below)

⁴ For terms that do not yet have a definition, the *Actantial structure* data category is displayed automatically.

The *Lexical relations* data category (word families, hypernyms and collocations) contains information that should provide help for understanding, producing and translating texts (communicative situations), as well as in mastering the computing terminology (cognitive situations). However, this section contains a considerable amount of information that could also overload the content presentation. Thus the tabs that contain these data categories are not displayed on demand. The organization of lexical relations will be described in the next section. Furthermore, we changed the title of some data categories (*Related meaning* to *See also*) or simply removed them (e.g. *Lexical relations*); again to avoid confusing users with technical metalanguage.

Former version

Explanation	Lexical function	Related term
Related Meanings		
≈	QSyn	document ₁
Other Parts of Speech and Derivatives		
Exploite un ensemble de f.	[Meta]	métafichier
Dans un f.	Loc-in	dans un ~
Types of		
Que l' <i>utilisateur</i> vient de créer	Hypo - Situation	nouveau ~
Qui a une durée limitée	Hypo - Durée	~ temporaire
Qui est de grande taille	MagnTaille	gros ~ ~ lourd
Qui est de petite taille	AntiMagnTaille	petit ~
Qui peut être utilisé	Able2Real@	~ lisible ~ éditable _{1a}
Qui ne peut pas être utilisé	AntiAble2Real@	~ illisible
Un f. qui n'est plus en mesure de fonctionner adéquatement	A2Caus@Degrad	~ corrompu ₁

New version

Explanation	Related term
Que l' <i>utilisateur</i> vient de créer	nouveau ~
Qui a une durée limitée	~ temporaire
Qui est de grande taille	gros ~ ~ lourd
Qui est de petite taille	petit ~
Qui peut être utilisé	~ lisible ~ éditable _{1a}
Qui ne peut pas être utilisé	~ illisible
Un f. qui n'est plus en mesure de fonctionner adéquatement	~ corrompu ₁ ~ endommagé ₁

Figure 7: Lexical relations displayed for *fichier* (En. *file*) in the former version (above) and the new version (below)

As mentioned in Section 2, the relationships between terms are encoded by means of LFs and the labels used to do so can be quite opaque. We thought about the relevance of this information for the targeted users. We consider that while LFs are useful for describing and organizing lexical relations, their labels are difficult to decipher. So we still use them during the encoding process, but hide them in the online version. Therefore, users can find assistance to translate a collocation or a phrase correctly (communicative situations) without being confused by abstract formulae (Figure 7).

- *Data organization*

As mentioned in Section 2, the DiCoInfo lists the numerous lexical relations that exist between the headword and other terms. Related terms are listed in a table (Figure 7). Explanations of the relationships are presented in the left column that describes the LFs (Mel'čuk et al., 1984–1999; 1995). We decided to reorganize the lexical relations, i.e. the collocations and the *Types of data category* (e.g. *key: backspace; Enter ~*).

The procedures for organizing both these data categories are similar. Concerning collocations, previous work had been carried out for classifying them (L'Homme & Leroyer, 2009; Jousse et al., 2011). The solution implemented for collocations consisted of a system of classes in which specific collocations were classified according to their general meaning. For instance, all verbal and deverbal collocates expressing typical uses of an object denoted by a term are placed in a general class called UTILISER/NE PAS UTILISER (En. USE/NOT TO USE). Instead of having all collocates presented at once, users can select the class that is closest to the meaning they wish to express (USE, CREATE, MOVE, and so on).

We used the same general principles to classify the different items appearing under the *Types of data category*. In the previous version of the DiCoInfo, the list of terms was very long, and without a specific organization scheme. In order to facilitate the accessibility of these data, we classified the related terms according to a system of classes defined in L'Homme & Jia (2015). The LFs are used to define our system of classes, and again they are not displayed in the online version: users only have access to the explanation in natural language. First, we group the related terms into *intermediate classes* (IC); then *generic classes* (GC) are defined in which we group the intermediate ones (Figure 8). It should be noted that for the time being these changes have been applied only to the French version, thus the examples are given in French.

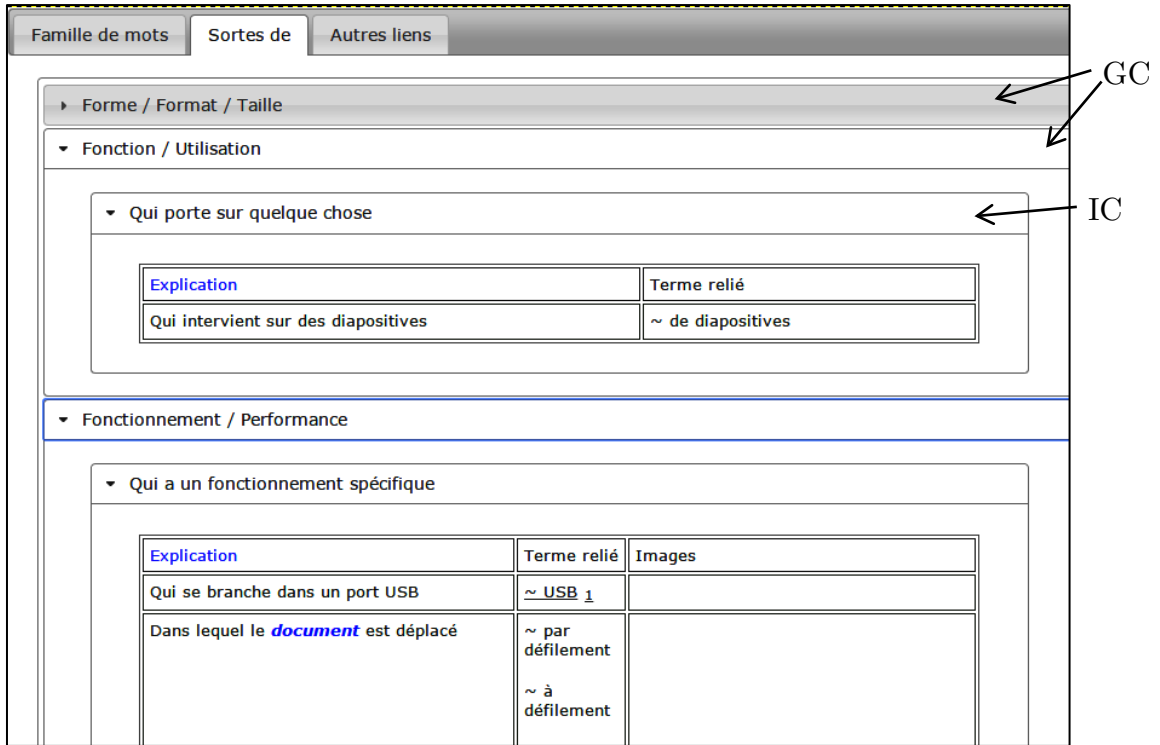


Figure 8: Generic (GC) and intermediate classes (IC) for *numériseur* (En. *scanner*)

As shown in Figure 8, in the new version, we set up a system of accordions that consists of collapsible content panels for presenting the semantic classes. Thus, nested accordions are shown according to the lexical links found in an entry. At the top level, accordions corresponding to the *generic classes* are listed. When expanded, each accordion panel shows in turn inner accordions that correspond to the *intermediate classes*.

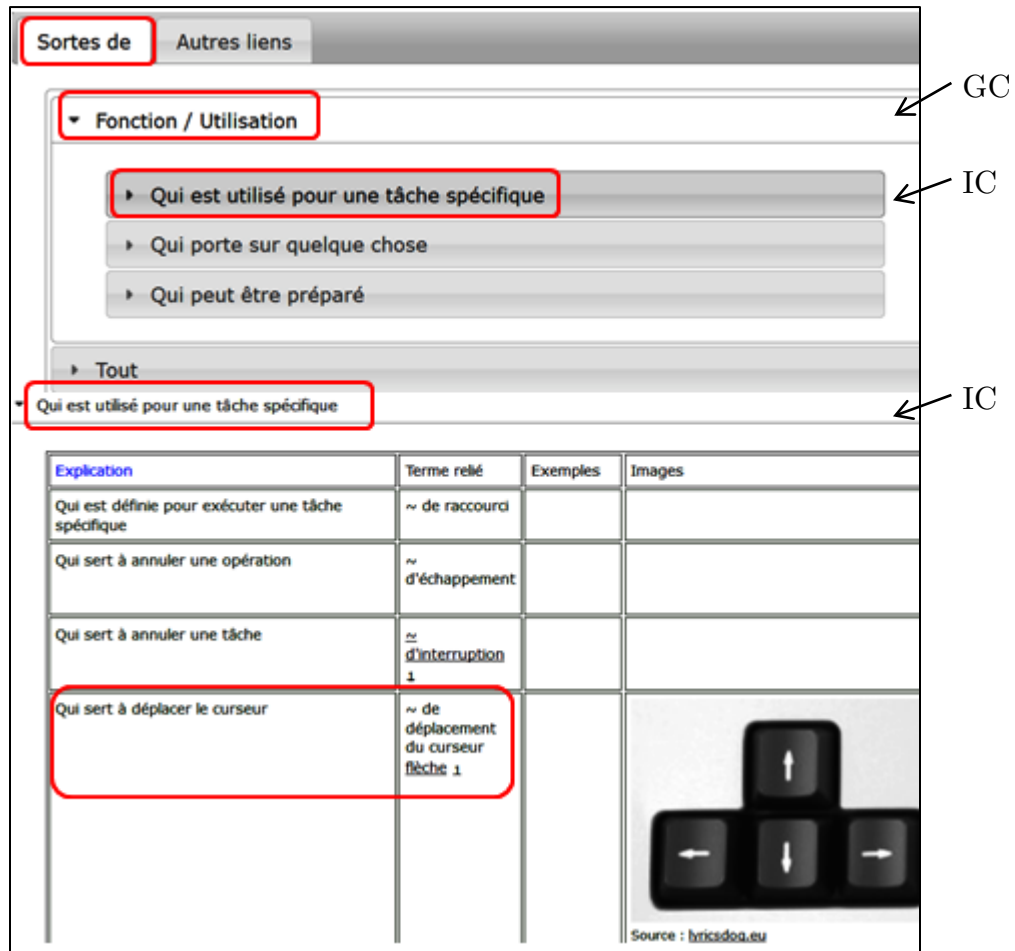


Figure 9: Navigation through the *Types of data* category

In this way, users may look up a related term by considering its meaning, e.g. FONCTION/UTILISATION (En. FUNCTION/USE); FORME/FORMAT/TAILLE (En. FORM/FORMAT/SIZE); MODE DE FONCTIONNEMENT (En. FUNCTIONING MODE), etc. We will illustrate the way users can access a related term with the example *touche* (En. *key*). In this example, it is assumed that a given user wishes to find the French translation of *arrow key* and that he has to go through these four steps (Figure 9):

1. Activate the SORTES DE (*Types of*) tab in the *touche* (En. *key*) entry.
2. Expand the accordion corresponding to the generic class FONCTION/UTILISATION (En. FUNCTION/USE).
3. The accordion containing the intermediate class UTILISÉ POUR UNE TÂCHE SPÉCIFIQUE (En. USED FOR A SPECIFIC TASK) is already opened (i.e. not collapsed) since there is just one item to display.
4. By means of the explanation “Qui sert à déplacer le curseur” (En. “That is used

to move the cursor”), the user accesses the right expression *touche de déplacement de curseur* followed by its synonym *flèche*.

b. Addition of Multimedia

Since “images enhance textual comprehension and complement the linguistic information provided in other data fields” (Faber et al., 2006: 757), pictures were added to some entries (Figure 10). In addition it has been demonstrated that images have a positive effect on vocabulary acquisition (Lew, 2012), and become very useful in cognitive situations. The terms for which they were added represent concrete objects (i.e. *keyboard*, *mouse*, *printer*, etc.). Some pictures were also added within the entries and associated with some related terms in the *Types of data* category (*Key: arrow ~*) (Figure 9).

The screenshot shows a dictionary entry for 'numériseur' (scanner). It includes a definition, synonyms (scanner, scanneur), and a context. An image of a scanner is shown on the right. The source is cited as 'appstorm.net'.

Figure 10: Image for *numériseur* (En. *scanner*)

Relatif à l'utilisateur		
Explication	Terme relié	Exemples
Dont l' utilisateur n'est pas connu	~ anonyme 1b	Si aucun nom d'utilisateur est donné il s'agit alors d'une connexion anonyme , le login sera alors anonymous et la coutume veut que le mot de passe d'une session anonyme soit son adresse e-mail.

Figure 11: Example given in *Types of data* category for *connexion* (En. *connection*): *connexion anonyme* (En. *anonymous login*)

c. Addition of Examples

In order to assist users in communicative situations, we chose to associate some examples with related terms in the *Types of* data category (Figure 11), so that users can see the way related terms are used in specialized contexts. This strategy was also adapted in the DAFA (Binon et al., 2000).

4. Conclusion

In this paper we presented various strategies to convert a specialized lexical database into a learners' dictionary. We defined our learners' dictionary as one that meets specific user needs in specific situations based on the principles of functional lexicography (Tarp, 2008). We redesigned its presentation and layout using technologies that allowed us to take these needs into account in the online version. The targeted users are first and foremost translation students and translators with little experience and whose specific needs are both communicative and cognitive.

The database we adapted is the *DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet* and its transformation raised a certain number of challenges. The database contained technical metalanguage that needed to be placed in the background or hidden altogether. In addition, each entry contained various data categories whose presentation required simplifying. Decisions were made about which modifications were necessary and how they should be carried out. Our objective was to preserve most of the information already provided in the DiCoInfo while presenting it in such a way that it would meet the defined user needs. Finally, these changes were made according to two parameters: simplification of the presentation, and the newly implemented lexicographical functions of the DiCoInfo. Modifications have been made in the interface and its layout. In addition, the presentation of data categories was completely revised; multimedia was also added.

However, there is still some room for improvement. We are currently exploring the possibility of adding images in entries for verbs (*download, write*), as well as in other entries describing terms that denote activities (*compilation*). We are also aware that some explanations for lexical relations should be revised in order to improve their readability. In addition, up to now we have focused on improving the presentation of the DiCoInfo, however additional work could be carried out on the accessibility of the information contained in other data categories in order to make the information spotting simpler and faster. Finally, it would be interesting to collect user feedback on the changes we have made to date and compare the reactions of professional translators with those of translation students.

5. References

- Alipour, M. (2014). *Méthodologie de conversion de dictionnaires spécialisés en dictionnaires d'apprentissage: application au domaine de l'informatique*. Masters Dissertation. Université de Montréal.
- Bergenholtz, H., & Tarp, S. (2003). Two opposing theories: On the H.E. Wiegand's recent discovery of lexicographic functions. *Hermes*, 31, pp. 171-196.
- Binon, J., Verlinde, S., Van Dyck, J., & Bertels, A. (2000). *Dictionnaire d'apprentissage du français des affaires : Dictionnaire de compréhension et de production de la langue des affaires*. Paris: Didier.
- Clark, J. (1999). XSL Transformations (XSLT), Version 1.0. W3C Recommendation 1999. *World Wide Web Consortium*.
- DiCoInfo. *Dictionnaire fondamental de l'informatique et de l'Internet*. Accessed at: <http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi>
- Dziemianko, A. (2010). Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography*, 23(3), pp. 257-273.
- Faber, P. F., Arauz, P. L., Velasco, J. A. P. & Reimerink, A. (2006). Linking images and words: the description of specialized concepts. In E. Corino, C. Marelllo & C. Onesti. (eds.) *Atti del XII Congresso Internazionale di Lessicografia*: Torino, pp. 751-763.
- Fuertes-Olivera, P. A. & Nielsen, S. (2012). Online Dictionaries for Assisting Translators of LSP Texts: the Accounting Dictionaries. *International Journal of Lexicography*, 25(2), pp. 191-215.
- Goguy, E. (1999-2014). *DicoFR: Dictionnaire de l'informatique et d'internet*. Accessed at: <http://www.dicofr.com>.
- Jousse, A. L., L'Homme, M. C., Leroyer, P. & Robichaud, B. (2011). Presenting collocates in a dictionary of computing and the Internet according to user needs. In I. Boguslavsky & L. Wanner (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*. Barcelona, pp. 134-144.
- Leroyer, P. (2013). *Projet de recherche terminographique. Évaluation du DiCoInfo – Tests utilisateurs 2013 – Université de Montréal. Descriptif du test et résultats*. Report.
- Lew, R. (2012). How can we make electronic dictionaries more effective? In S. Granger & M. Paquot (Eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 343-361.
- Lew, R. & de Schryver, G. M. (2014). Dictionary users in the Digital Revolution. *International Journal of Lexicography*, 27(4), pp. 341-359.
- L'Homme, M.-C. (2010). Designing terminological dictionaries for learners based on lexical semantics: The representation of actants. In P.A. Fuertes-Olivera (ed.) *Specialized dictionaries for learners*. Berlin/New York: De Gruyter, pp. 141-153.

- L'Homme, M.-C. (2014a). *Manuel de DiCoInfo*. Accessed at: <http://olst.ling.umontreal.ca/dicoinfo/manuel-DiCoInfo.pdf>.
- L'Homme, M.-C. (2014b). Why Lexical Semantics is Important for E-Lexicography and Why it is Equally Important to Hide its Formal Representations from Users of Dictionaries. *International Journal of Lexicography*, 27 (4), pp. 360-377.
- L'Homme, M.-C. & Jia, Z (2015). Combinaisons lexicales spécialisées à base nominale dans un dictionnaire d'informatique. *Cahiers de lexicologie*, 106, pp. 228-251.
- L'Homme, M.-C., & Leroyer, P. (2009). Combining the semantics of collocations with situation-driven search paths in specialized dictionaries. *Terminology*, 15(2), pp. 258-283.
- Meier, W. et al. (2011). *eXist : an Open Source Native XML Database*. Accessed at: <http://exist-db.org/exist/apps/homepage/index.html>.
- Mel'čuk, I. A., Clas, A. P., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.
- Mel'čuk, I. A. et al. (1984 - 1999). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques*. Volumes I. II. III. IV. Montréal: Presses de l'Université de Montréal.
- Müller-Spitzer, C., Koplenig, A., & Topel, A. (2012). Online dictionary use: Key findings from an empirical research project. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 425-457.
- Pyne, S., & Tuck, A. (1996). *Oxford dictionary of computing for learners of English*. Oxford: Oxford University Press.
- Ruppenhofer, J., Ellsworth M., Petruck M., Johnson, C., & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. Accessed at: <http://framenet.icsi.berkeley.edu>.
- Sarrion, E. (2012). *jQuery UI* (1st ed.). Sebastopol (CA): O'Reilly Media.
- Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge general lexicographical theory with particular focus on learner's lexicography*. Tübingen: Max Niemeyer Verlag.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The role of crowdsourcing in lexicography

Jaka Čibej¹, Darja Fišer¹, Iztok Kosem^{2,3}

¹ Department of Translation, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

² Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia

³ Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

E-mail: jaka.cibej@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, iztok.kosem@trojina.si

Abstract

In the past decade, crowdsourcing has been used with great success in specialized lexicographic tasks, such as collecting candidate lexemes for dictionary updates or validating automatically identified synonyms. However, professional lexicography is only now starting to explore crowdsourcing as an integral part of the workflow, thereby opening a number of important questions that could have lasting consequences on the nature of lexicographic work, its management and financing, as well as the perception, use and life-cycle of the lexicographic product. In this paper, we address these questions through the perspective of a proposal for a new monolingual dictionary of Slovene, in which crowdsourcing will play an integral role at a number of stages of dictionary construction – from headword list creation to dealing with stylistic issues.

Keywords: crowdsourcing; microtask design; crowd motivation; quality control; legal and ethical aspects of crowdsourcing

1. Introduction

Crowdsourcing is a term first introduced in 2006 to signify a process that involves a group of people (also called a crowd) that contribute towards achieving a goal by distributing the overall workload among the individuals in the group (Howe, 2008). The crowd does not necessarily consist of experts in the relevant field. In fact, a number of crowdsourcing projects have shown that even groups of non-expert individuals are talented, creative and productive enough to solve complicated tasks that in the past were solely the domain of experts. Today, due to modern technology and the global spread of the internet, channelling the potential of the crowd is becoming increasingly simple, more affordable and effective.

Although crowdsourcing is discussed with increasing frequency in lexicography, it has not yet been tested in large-scale, diverse and comprehensive settings. As shown by Abel & Meyer (2013), user contributions to dictionaries are currently limited to collaborative lexicographic projects or dictionary correction after publication. At the same time, lexicographers are facing increasing time constraints and amounts of data. What is more, the increasing (semi-)automatization of lexicographic work is turning some stages of dictionary creation into routine processes, for which lexicographers are overqualified. This calls for the introduction of crowdsourcing and user contributions

in dictionary creation. If established, it could have lasting consequences on the nature of lexicographic work, its management and financing, as well as the perception, use and life-cycle of the lexicographic product.

In this paper, we propose to integrate crowdsourcing into the overall workflow of lexicographic projects. We also address a number of important questions that arise in the process, such as the importance of appropriate microtask design, crowd motivation, quality control as well as legal and ethical aspects of crowd payment, all through the perspective of a proposal for a new monolingual dictionary of Slovene, in which crowdsourcing will play an integral role in a number of stages of dictionary construction – from headword list creation to dealing with stylistic issues.

2. The crowd and lexicography

One of the earliest examples of obtaining active participation of the general public in dictionary production was the creation of the Oxford English Dictionary (OED) in the late 19th century, when the OED editorial board encouraged volunteers to send in their contributions containing words and examples of use (Lanxon, 2011).

In the last decade, crowdsourcing has already been used successfully in a number of linguistic projects. For example, when evaluating Puzzle Racer, an annotation game, Jurgens & Navigli (2014) find it to be equally effective compared to annotation by experts, with the costs being 73% lower. Using the CrowdFlower platform, Fossati et al. (2013) crowdsourced the annotation of FrameNet, a lexical database of English, and found the crowdsourcing method to be both faster and more accurate than conventional annotation methods. Using sloWCrowd, a custom developed open-source crowdsourcing tool for lexicographic tasks, Fišer et al. (2014) corrected errors in the automatically developed WordNet for Slovene, and found the annotators' average accuracy to be 80.12%, which is high for complex lexical semantic tasks. When annotating a silver standard corpus of Croatian, Klubička & Ljubešić (2014) find the accuracy of a single worker to be approximately 90%, and the accuracy of the majority answer of three workers to be approximately 97%.

All this suggests that crowdsourcing could also be used in lexicography to great effect – not as a final or main phase of dictionary creation, but as a method to filter and process data before its implementation in actual dictionary creation by lexicographers. However, in order to ensure the effectiveness of the crowdsourcing method, several factors must be taken into consideration: crowd motivation, microtask design, quality control, choice of crowdsourcing platforms, and legal or financial issues. An overview of these aspects is provided in the following sections.

2.1 Crowd motivation

Motivated contributors are crucial for the success of any crowdsourcing project, even more so with languages of limited diffusion, which cannot rely on a large pool of crowdsourcers. According to Lew (2013), the motivation provided by the project initiator can be psychological, social or economic.

Psychological motivation is based on the fact that many internet users find participating in crowdsourcing projects or contributing user-generated content psychologically satisfying or personally fulfilling, either as an act of altruism, a way of expressing their identity or simply because they find it entertaining. This motivational aspect was the basis for the development of games with a purpose (GWAP) – applications that enable individuals to solve tasks while playing a game. Examples include *Phrase Detectives*, an online game for anaphora resolution (Chamberlain et al., 2008); *Verbosity*, a game for collecting common-sense facts (von Ahn et al., 2006); *Puzzle Racer* and *KaBoom!*, both annotation games (Jurgens & Navigli, 2014); and *JeuxDeMots*, a game aimed at building a large-scale lexical network for French (Joubert & Lafourcade, 2012).

With social motivation, individuals are driven by their urge to interact with others who share similar interests. Such a group is willing to contribute to a project that will benefit their community, perhaps by resulting in a useful product or by providing a chance for the individuals to improve their skills or to express their enthusiasm for a particular topic. A subcategory of social motivation is educational motivation (e.g. students solving tasks either as part of their academic obligations or as an extra-credit activity). Another aspect of social motivation involves the recognition a contributor receives for their work and effort in a community; for instance, an esteemed title (e.g. Wikipedia Editor) or credit on a hall-of-fame list. Successful projects involving social motivation include a number of well-known collaborative projects, such as *Wiktionary* and *Urban Dictionary* or its Slovene counterpart *Razvezani jezik*¹.

When crowdsourcing is used for large-scale or commercial projects where a substantial input or long-term involvement is expected from crowdsourcers, researchers typically resort to economic motivation by offering *micropayments*, i.e. small remuneration paid to the contributor for every successfully completed task (cf. Rumshisky, 2011; Akkaya et al., 2010; Fossati et al., 2013). Other types of economic motivation include prizes and vouchers (cf. El-Haj et al., 2014; Fišer et al., 2014). If using economic motivation, it is important to bear in mind the ethical aspects of recruiting and paying the crowdsourcers relative to the difficulty level and time spent on the task, cost of living in their country of residence, easy access to the earnings, etc. (cf. Sabou et al., 2014).

¹ <http://razvezanijezik.org>

2.2 Microtask design

Since microtasks are often undertaken by non-experts, they need to be simple to process both mentally and logistically. They should not be too time-consuming nor should they require a high degree of expertise or too much introductory training. As pointed out by Rumshisky (2011) and Biemann & Nygaard (2010), crowdsourcing tasks should be kept simple (with clear, short instructions) and designed to enable maximum effectiveness by splitting complex annotation into simpler steps. The importance of well-designed microtasks is also pointed out by Kosem et al. (2013), who showed that complex, multi-dimensional questions, or those that require subjective evaluations, do not yield satisfactory results.

2.3 Quality control

There are a number of ways to control the quality and consistency of crowdsourcing results. The first method is the *gold standard*, a dataset which contains a number of microtasks that have been pre-annotated (already answered correctly) by experts. These tasks are offered to crowdsourcers at various points during their work in order to test their reliability. If an individual fails to pass a threshold, his or her answers are deemed unreliable and are excluded from the final results (Rumshisky, 2011).

Another way of controlling quality is to observe *inter-annotator agreement*. This is achieved by offering different crowdsourcers the same task, thus obtaining multiple answers for each task. The final decision is achieved by taking into consideration the *majority vote*, i.e. the answer chosen by the most annotators. Based on the distribution of the multiple answers, a *confidence score* per microtask or per crowdsourcer can be computed (Oyama et al., 2013). However, it is important to consider that an optimal balance must be achieved between multiple annotations for the same task and new annotations, as multiple annotation is costly.

The (borderline or difficult) cases with insufficient consensus among crowdsourcers may then be manually annotated by an expert. This process is called *refereeing*. If the microtasks were designed properly and the annotation process successful, the expert is only required to evaluate a small number of ambiguous examples, while the bulk of the work is still crowdsourced. If, on the other hand, the annotators disagree in a significant number of cases, it might indicate that the microtasks were not designed efficiently, were not assigned to the appropriate target group, or that the annotation guidelines need to be further refined to provide clearer instructions (Fossati et al., 2013).

The last approach to quality control is observing *intra-annotator agreement*, which measures the consistency of a single crowdsourcer in answering the same microtasks at various points of their engagement (Gut & Bayerl, 2004). This allows for the exclusion

of unreliable annotators who are either ‘spam workers’, not knowledgeable enough or not confident enough to provide consistent answers. This process, however, is also costly. The more common the iteration of previous questions, the smaller the number of new annotations that will ultimately be available. Also, iteration should not be noticed by crowdsourcers as this may affect their motivation.

2.4 Legal and financial issues

When using crowdsourcing for lexical resource development, a number of legal and financial issues arise. Although these depend heavily on local legislation and project funding, we provide a general overview of the key issues that need to be taken into consideration. Although they are not central to the content and quality of lexicographic projects, they often act as a significant barrier to lexicography embracing crowdsourcing since most lexicographic teams, especially in academic settings, are unfamiliar with the legislation restrictions in this area and rarely get sufficient support from legal experts in the field.

Dataset availability – If the datasets used in crowdsourcing are to be made available to the public, a suitable license needs to be selected in accordance with local legislation on copyright and personal data protection.

Disclaimer – Before contributing to the project by solving tasks, crowdsourcers should agree to a disclaimer that informs them on how the results of their work will be used.

Crowdsourcer acknowledgement – Because crowdsourcers typically contribute a sizeable amount of work to the project, it needs to be determined if and how they should be credited on the final product in accordance with local copyright legislation.

Recruitment restrictions – Local legislation may impose restrictions on crowdsourcer recruitment. This is especially true in the case of under-aged workers.

Payment restrictions – Another matter to consider is potential payment restrictions, e.g. how local tax legislation treats micropayments or prizes for participating in crowdsourcing projects.

2.5 Crowdsourcing platforms

In this section, we provide an overview of the platforms that have either already been used for crowdsourcing in linguistics or show potential in lexicography. Both commercial and open-source crowdsourcing platforms exist.

The most widely known and used crowdsourcing platform is **Amazon Mechanical Turk**² (cf. Rumshisky, 2011; Rumshisky et al., 2012; Biemann & Nygaard, 2010; Snow et al., 2008). Campaign management, quality control measures and payment support are already integrated in the administrator’s interface, and a substantial crowdsourcing community has already been recruited, at least for the bigger languages. Similar examples are **CrowdFlower**³ and **Clickworker**⁴ which offer a number of applications, ranging from data categorisation to sentiment analysis. Microtasks can be uploaded using CML, CSS or Javascript. Crowdsourcers can be filtered according to age, expertise or geographic location.

Among open-source platforms, the most notable is **Crowdcrafting**⁵, which is based on **PyBossa**⁶, a Python-based open-source framework for creating crowdsourcing projects that can be installed locally and is available under the Creative Commons License BY-SA 4.0. Another open-source tool is **slowCrowd**⁷ (Tavčar et al., 2012), which is PHP/MySQL-based and was originally developed for correcting mistakes in automatically generated semantic lexicons (such as Wordnet), but has been upgraded to allow for project-specific task specifications.

3. Crowdsourcing workflow for lexicography

In this section, we provide an overview of proposals to utilise crowdsourcing methods in the various stages of corpus-based dictionary construction projects. We propose a modular approach that can be adapted to the specific nature of the project at hand and the budget available. Not all stages need to be followed. Their order can be changed and some can be done in parallel, but it is important to at least consider the recommended phases and address the issues raised in each of them, as crowdsourcing is a complex, time-consuming and potentially costly procedure that cannot yield useful results without careful planning and task design.

Before deciding on a crowdsourcing campaign, an estimate of the required investment should be made with respect to time, money and personnel, as the campaign should not take up more time and financial and/or human resources than conventional annotation methods. However, if crowdsourcing is integrated into dictionary construction from the very beginning, different crowdsourcing tasks at all dictionary construction levels can be designed according to the same principles and use the same pre- and post-processing chains and crowdsourcing platform, making the effort of setting up a viable crowdsourcing environment all the more worthwhile.

² <https://www.mturk.com>

³ <http://www.crowdflower.com/>

⁴ <http://www.clickworker.com/en/>

⁵ <http://crowdcrafting.org/>

⁶ <http://pybossa.com/>

⁷ <http://nl.ijs.si/slowcrowd/about.php?project=slowcrowdmain>

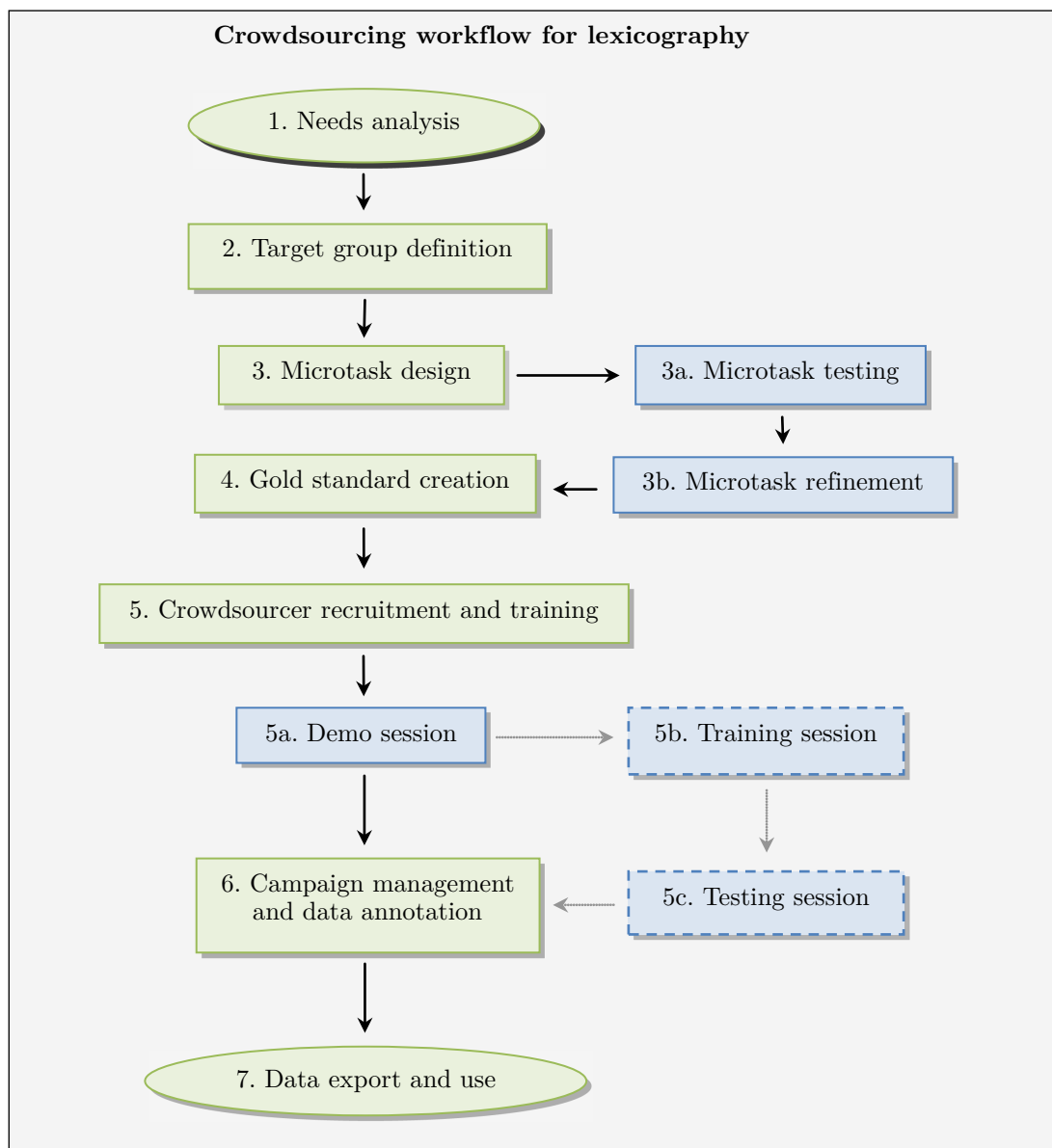


Figure 1: Crowdsourcing workflow for lexicography. Green-coloured boxes represent main phases and blue-coloured ones subphases. Dashed boxes and arrows represent optional phases which can be omitted in small-scale, low-budget campaigns

Phase 1: Needs analysis – The first step of each crowdsourcing campaign requires a thorough needs analysis. Apart from the goal and expectations of the campaign (i.e. what can be expected in terms of volume and usability of the obtained results), it is also necessary to determine the type, amount, availability and format of the data required.

Phase 2: Target group definition – Once the needs have been analysed, it is necessary to determine the required crowdsourcer profile to ensure results of a suitable quality. The problem at hand may be suitable for the general public without any specialized linguistic or lexicographic knowledge or may require a certain degree of expertise and can only be solved effectively by e.g. language students or even expert lexicographers.

Phase 3: Microtask design, testing and refinement – The most important and

difficult part of crowdsourcing is microtask design. As already mentioned, microtasks should be one-dimensional questions with short, clear instructions, suited to the knowledge prerequisites of the target crowdsourcer profile. In addition, solving microtasks should be carried out through a user-friendly interface. No tasks should be included that do not benefit from this method and are likely to provide unreliable results. The designed microtasks need to be tested in a pilot study so that any identified incongruences and inconsistencies can be removed and any unclear, confusing or too complex microtasks refined.

Phase 4: Gold standard creation – A certain number of microtasks needs to be annotated by experts to create a gold standard that is later used to ensure the accuracy of crowdsourcing results, i.e. to filter out unreliable crowdsourceers or answers. The dataset should be as representative of the entire set of microtasks as possible, especially in terms of difficulty and complexity (e.g. it should not include only simple, transparent examples nor should it contain too many borderline examples to make it impossible for the annotators to achieve a sufficient degree of accuracy).

Phase 5: Crowdsourcer recruitment and training – Crowdsourceers need to be recruited and trained. Usually, a *demo session* (e.g. a presentation or a video) is held to introduce the crowdsourceers to the annotation process. The demo session is then followed by a *training session*, which either consists of a live annotation session supervised by an expert who offers advice and additional information to the crowdsourceers should they require it (e.g. with ambiguous borderline examples) or an online annotation session where automated feedback is provided with each answer. The next step is the *testing session*, which is used to determine whether the crowdsourceer has achieved a sufficient degree of accuracy to be recruited. In low-budget scenarios, the training and testing sessions are often skipped.

Phase 6: Data annotation and campaign management – In this step, the recruited crowdsourceers solve the microtasks provided by the initiator. The initiator needs to monitor the campaign and decide whether any additional fine-tuning is necessary, e.g. if the set of microtasks needs to be expanded, if the crowdsourceers are motivated enough to provide a consistent flow of answers, if the results meet the expectations of the project, etc.

Phase 7: Data export and use – The final phase involves exporting the crowdsourceed data into an appropriate format for further use in the project (e.g. algorithm training or inclusion in a dictionary). The crowdsourcing platform should allow the data to be exported at any point of the crowdsourcing campaign for preliminary analyses.

This crowdsourcing workflow will play an integral role in the creation of a new monolingual dictionary of Slovene, the plans for which are presented in the following section.

4. Crowdsourcing for the new Slovene dictionary

Slovar sodobnega slovenskega jezika (SSSJ) is a new monolingual dictionary of Slovene planned by the Centre for Language Resources and Technologies of the University of Ljubljana (CJVT UL)⁸. The goal of the proposed project is to construct a comprehensive corpus-based dictionary of Slovene that will reflect contemporary language use and will be built in accordance with modern lexicographic trends and the increasingly digital and online nature of lexicographic products. The project envisions the creation of an open-source database that will ultimately serve not only as the basis for a new monolingual dictionary of Slovene, but will also enable the development and improvement of both existing and new language technologies for Slovene, as well as the creation of a number of specialised Slovene dictionaries for different user profiles (e.g. linguists, students, learners of Slovene as a foreign language).

The initial proposal by Krek et al. (2013), based on which the plans for SSSJ and related resources are currently being made, envisioned that the SSSJ database would be completed in five years. Judging by experience from similar projects, such as the *Algemeen Nederlands Woordenboek* (Tiberius & Schoonheim, 2014) and the *Great Dictionary of Polish* (Żmigrodzki, 2014), this is a rather short period to create a database of any language from scratch, which is why the proposal includes an important innovation in lexicography: initial automatic extraction of corpus data. This method has already been tested on Slovene by Kosem et al. (2013) and is currently being used for the purposes of the *Estonian Collocation Dictionary* (Kallas et al., forthcoming). However, automatically extracted data requires a great deal of post-processing, including many routine and trivial tasks for lexicographers; this has led to the decision to make crowdsourcing an integral part of the SSSJ database creation, based on numerous good practice examples from abroad (Klubička & Ljubešić, 2014; Jurgens & Navigli, 2014; Fossati et al., 2013; inter alia) and the successful implementation of crowdsourcing in other Slovene projects (Kosem et al., 2013; Fišer et al, 2014).

4.1 SSSJ crowdsourcing scenarios

An example of a crowdsourcing task is distributing automatically extracted examples into different senses and subsenses. During the analysis, a lexicographer first makes a rough draft of sense division with one or more short glosses or an indicator for each sense, and then distributes the (automatically extracted) examples, collocates and grammatical relations, deleting any irrelevant or incorrect information in the process. To a large extent, the distribution of information can be carried out by crowdsourcers with a microtask in which they are asked to assign the extracted corpus examples to

⁸ <http://www.cjvt.si/projekti/sssaj/>

the relevant (sub)sense. In addition to the available senses and subsenses, crowdsourcers may also categorise examples as *None of the above senses*, when the example cannot be attributed to any (sub)sense offered; or as *Unclear example*, when the provided context is insufficient for the crowdsourcer to select one of the (sub)senses. The final decisions are then achieved through a majority vote or, if the majority vote is not unanimous or sufficiently clear (according to a predetermined threshold), through refereeing by a lexicographer.

While the only task for crowdsourcers is the distribution of examples, the results have many other uses. For instance, crowdsourcers indirectly distribute collocates attested in the examples as well as the grammatical relations under which the collocates are provided. Moreover, the examples marked as unclear are candidates for removal from the database or at least for omission from the dictionary entry. If a significant number of examples for a particular collocate is marked as unclear, the collocate itself will also need to be inspected. While a similar approach can be used for the examples categorised as *None of the above senses*, those examples carry two other potentially valuable pieces of information as they can alert the lexicographer to an overly coarse sense division or even to an overlooked (sub)sense.

Crowdsourcing can also be implemented in a number of other aspects of dictionary compilation and language resources (both new and existing); the improvement or development of which is an integral part of a dictionary project, in our case SSSJ. We provide a number of preliminary suggestions in the following paragraphs, but many more can and will be explored within the framework of the SSSJ project, depending on the budget available.

Lexicon – Microtasks concerning the creation of the SSSJ lexicon could involve determining the standard declension paradigm of headwords, the relation between words in terms of word-formation, the categorisation of marked (e.g. non-standard) word forms, and the pronunciation of the headword and its declined forms. In addition, crowdsourcing could be used to expand the lexicon of word forms for further use in the development of language technologies for Slovene.

Grammar – In terms of grammar, solving microtasks could help determine the relationship between certain interchangeable suffixes (e.g. the plural of *študent* ‘student’, which can be either *študenti* or *študentje*) or word forms (e.g. the demonstrative pronouns *oni* and *tisti*).

Standard – Microtasks concerning standard Slovene might include checking lists of individual paradigms and their potential corrections, as well as adding information on pronunciation and syntax.

Stylistics – Microtasks in stylistics could contribute towards developing the taxonomy of stylistic qualifiers and determining (or confirming) the stylistic qualifiers for dictionary headwords (or at least those that are deemed problematic).

User feedback – Crowdsourcing could also contribute towards the development of a user-friendly interface for the dictionary. By solving microtasks, potential dictionary users could decide between various options in terms of design, transparency, usefulness, etc., and choose the one they find suits the best.

5. Conclusion and future work

Crowdsourcing has great potential in lexicography, as evidenced by a number of linguistic projects that have already successfully used crowdsourcing as an effective method for data processing. To ensure the successful implementation of crowdsourcing in the lexicographic workflow, many aspects need to be considered: from microtask design, data preparation, crowd profiling and motivation to legal and financial issues.

The SSSJ project aims to be one of the first dictionary projects to give crowdsourcing a prominent role in the development of a database for a new monolingual dictionary of Slovene. The experience from the project so far has already shown that the need for crowdsourcing input extends beyond the dictionary database to any related existing or future language resource, such as a lexicon or a user interface. In addition, the crowd could be used to establish a permanent user feedback channel through crowdsourcing.

It is noteworthy that the results obtained from lexicographic crowdsourcing tasks can also be used for other purposes, e.g. for the improvement of language tools used by lexicographers. For example, corpus examples identified as unclear could form a training corpus for the improvement of a tool for extracting good dictionary examples. Similarly, identifying incorrect examples of collocates under a particular grammatical relation can help fine-tune scripts for extracting grammatical relations and their collocates from the corpus.

Crowdsourcing may well become a common tool in the next generation of lexicographic projects, making it much less time- and resource-consuming to keep up with the constant changes in language use as well as the increased demand for linguistic data-processing. We can therefore envisage the emergence of in-house crowdsourcing teams focused solely on providing support to lexicographers, linguists and researchers with language-related crowdsourcing tasks.

6. Acknowledgement

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

7. References

- Abel, A. & Meyer, C. (2013). The dynamics outside the paper: user contributions to online dictionaries. In *Proceedings of eLex 2013*, pp. 179–194.
- Akkaya, C., Conrad, C., Wiebe, J. & Mihalcea, R. (2010). Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. In *Proceedings of NAACL-HLT 2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*.
- Biemann, C. & Nygaard, V. (2010). Crowdsourcing WordNet. In *Proceedings of the 5th Global WordNet Conference*. Mumbai, India.
- Chamberlain, J., Poesio, M. & Kruschwitz, U. (2008). Phrase Detectives: A Web-based collaborative annotation game. In *Proceedings of iSemantics*. Graz, Austria.
- El-Haj, M., Kruschwitz, U. & Fox, C. (2014). Creating Language Resources for Under-resourced Languages: Methodologies, and experiments with Arabic. In *Language Resources and Evaluation 2014*. Springer.
- Fišer, D., Tavčar, A. & Erjavec, T. (2014). sloWCrowd: A crowdsourcing tool for lexicographic tasks. In *Proceedings of LREC 2014*, pp. 4371–4375.
- Fossati, M., C. Giuliano & Tonelli, S. (2013). Outsourcing FrameNet to the Crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 742–747.
- Gut, U. & Bayerl, P. S. (2004). Measuring the Reliability of Manual Annotations of Speech Corpora. In *Proceedings of Speech Prosody 2004, Nara*, pp. 565–568.
- Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. New York: Crown Publishing Group.
- Joubert, A. & Lafourcade, M. (2012). A new dynamic approach for lexical networks evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 23-25 May 2012.
- Jurgens, D. & Navigli, R. (2014). It's All Fun and Games until Someone Annotates: Video Games with a Purpose for Linguistic Annotation. *Transactions of the Association for Computational Linguistics*, 2. Association for Computational Linguistics, pp. 449–463.
- Kallas, J., Kilgarriff, A. Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M. & Viks, Ü. (2015). Automatic generation of the Estonian Collocation Dictionary database. In Kosem, I., Jakubiček, M., Kallas, J., Krek, S. (eds.) *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*. Ljubljana/Brighton: Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd., pp. 1-20.
- Klubička, F. & Ljubešić, N. (2014). Using crowdsourcing in building a morphosyntactically annotated and lemmatized silver standard corpus of Croatian. *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*. Ljubljana.

- Kosem, I., Gantar, P. & Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In *Proceedings of eLex 2013*, pp. 33–48.
- Krek, S., Kosem, I., Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika* (A Proposal for a new Dictionary of Contemporary Slovene). Version 1.1. Accessed at: http://trojina.org/slovar-predlog/datoteke/Predlog_SSSJ_v1.1.pdf.
- Lanxon, N. (2011). *How the Oxford English Dictionary started out like Wikipedia*. <http://www.wired.co.uk/news/archive/2011-01/13/the-oxford-english-wiktionary> (Access: 25. 10. 2014)
- Lew, R. (2013). User-generated content (UGC) in online English dictionaries. *OPAL - Online publizierte Arbeiten zur Linguistik*.
- Oyama, S., Baba, Y., Sakurai, Y. & Kashima, H. (2013). Accurate Integration of Crowdsourced Labels Using Workers' Self-Reported Confidence Scores. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 2554–2560.
- Rumshisky, A. (2011). Crowdsourcing Word Sense Definition. In *Proceedings of the Fifth Law Workshop (LAW V)*. Portland: Association for Computational Linguistics, pp. 74–81.
- Rumshisky, A., Botchan, N., Kushkuley, S. & Pustejovsky, J. (2012). Word Sense Inventories by Non-Experts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12*. Istanbul, Turkey.
- Sabou, M., Bontcheva, K., Derczynski, L. & Scharl, A. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *Proceedings of LREC 2014*, pp. 859–866.
- Snow, R., O'Connor, B., Jurafsky, D. & Y Ng, A. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 254–263.
- Tavčar, A., Fišer, D. & Erjavec, T. (2012). sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. In *Proceedings of the Eighth Language Technologies Conference*. Ljubljana: Jožef Stefan Institute, pp. 197–202.
- Tiberius, C. & Schoonheim, T. (2014). The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. In V. Hildenbrandt (ed.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/2014). Preprint accessed at: http://www.elexicography.eu/wp-content/uploads/2014/05/TiberiusSchoonheim_The-ANW-and-its-Lexicographical-Process_Preprint.pdf
- von Ahn, L., Kedia, M. & Blum, M. (2006). Verbosity: a game for collecting common-sense facts. *Zbornik konference SIGCHI conference on Human Factors in computing systems*. ACM, pp. 75–78.

Żmigrodzki, P. (2014): Polish Academy of Sciences Great Dictionary of Polish [Wielki słownik języka polskiego PAN]. *Slovenščina 2.0*, 2 (2), pp. 37–52.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Mobile Lexicography: Let's Do it Right This Time!

Henrik Køhler Simonsen

Copenhagen Business School, Dalgas Have 15, 2000 Frederiksberg

E-mail: hks.abc@cbs.dk

Abstract

Mobile phones are ubiquitous and have completely transformed the way we live, work, learn and conduct our everyday activities. Mobile phones have also changed the way users access lexicographic data. In fact, it can be argued that mobile phones and lexicography are not yet compatible. Modern users are already mobile – but lexicography is not yet fully ready for the mobile challenge, mobile users and mobile user situations.

The article is based on empirical data from two surveys comprising 10 medical doctors, who were asked to look up five medical substances with the medical dictionary app *Medicin.dk* and five students, who were asked to look up five terms with the dictionary app *Gyldendal Engelsk-Dansk*. The empirical data comprise approximately 15 hours of recordings of user behavior, think-aloud data and interview data.

The data indicate that there is still much to be done in this area and that lexicographic innovation is needed. A new type of users, new user situations and new access methods call for new lexicographic solutions, and this article proposes a six-pointed hexagram model, which can be used during dictionary app design to lexicographically calibrate the six dimensions in mobile lexicography.

Keywords: mobile lexicography; mobile user situation; mobile data access

1. Introduction and Problem

Lexicography has gone mobile. Mobile phones are ubiquitous (cf. Google, 2013: 2) and are used by virtually everybody everywhere. Also publishing houses have caught the mobile wave and developed and marketed a host of dictionary apps. People are already mobile – but is lexicography as a discipline ready for the mobile challenge? Are lexicography and mobile devices compatible at all, and what characterises the mobile user situation? Questions like these can only be answered by means of user surveys with real users in real-life contexts. User research is serious business, but unfortunately is often unrightfully criticized by researchers, who prefer theory over practice (cf. for example Tarp, 2008: 44), who refers to user research of specific lexicographic situations as “...trying to fill the leaking jar of the Danaids...”. However, purely deductive procedures are not enough.

Like dictionaries, dictionary apps are utility tools designed and developed to be used (cf. Wiegand, 1988) and they should be designed and developed based on reliable user survey data. This argument is supported by Müller-Spitzer (2013), who argues that it is important to collect empirical data relating to dictionary users and Lew (2015), who offers an interesting discussion of the opportunities and limitations of user

surveys in lexicography. Collecting real-life empirical data is difficult and hard work, but like Müller-Spitzer (2013), it is argued that obtaining empirical data “with all the restrictions that go with it” is important.

Furthermore, as pointed out by Lew (2015: 8–9), the number of participants tends to be low in tests under the naturalistic paradigm, and this is in fact also the case in the two empirical surveys discussed in this paper. In fact, the answer to the question of how many users you should test in usability research was already given in 1989, when Nielsen argued that user testing with five participants was a cheap, fast and satisfactory evaluation (cf. Nielsen, 2000). Today, the answer is still the same as “this lets you find almost as many usability problems as you’d find using many more test participants” (cf. Nielsen, 2012).

First, the methodology and the empirical basis of this article will be outlined and next a number of important theoretical considerations on what characterizes mobile lexicography will be briefly discussed. Third, this article offers a discussion of six dimensions of paramount importance in mobile lexicography, and finally the article proposes a six-pointed hexagram model, which can be used during dictionary app design to lexicographically calibrate the six determining factors in mobile lexicography.

2. Methodology and Empirical Basis

As already briefly described, this article is based on data from two empirical analyses, and both surveys belong to the naturalistic paradigm (cf. Lew, 2015).

First, the article draws on the insights and conclusions from an intra-consultation survey of the consultation behaviour of 10 medical doctors. The data and the insights from this survey are discussed in (Simonsen, 2013: 416–429) and (Simonsen, 2014: 259–260). The 10 medical doctors were asked to look up medical terms by means of the app *Medicin.dk* on an iPhone 4S, which was wirelessly connected to a PC by means of Reflector, cf. <http://www.air squirrels.com/reflector/>. The 10 medical doctors were asked to participate in two tests. In Test A the test persons were asked to look up five medical terms while sitting down at a desk. In Test B the 10 test subjects were asked to look up the same five terms while slowly walking around a hospital bed. The survey of the mobile user situation focussed on a number of concrete task-dependent situations. Both tests were recorded while the tasks were performed both from the “inside” by means of Reflector, and at the same time the user activities were recorded from the “outside” by means of a digital camera. In addition to the recordings from the “inside” and the “outside”, the empirical basis also includes think-aloud data, as the test persons were asked to think aloud and verbalize what they did and saw, etc. To deduce additional qualitative comments, the empirical basis also includes interview data as the test persons were interviewed before and after the tests (cf. also Simonsen, 2014: 259–260 for a detailed discussion).

Tests A and B were designed to imitate two typical user situations for many doctors: knowledge acquisition and knowledge checking prior to patient consultation and knowledge checking during a patient consultation. During the two tests, the doctors were asked to solve five tasks. The five tasks included looking up the five product names Terbasmin (asthma), Tamoxifen (breast cancer), Antepsin (ulcer), Tredaptive (cholesterol) and Fludara (leukaemia) and can be summarized as follows:

Task 1: Look up “Terbasmin” – to find information

Task 2: Look up “Tamoxifen” – to extract information about side effects to inform patient

Task 3: Look up “Antepsin” – to extract information about dosage to check prescription

Task 4: Look up “Tredaptive” – to extract information about dosage to inform patient

Task 5: Look up “Fludara” – to find and check spelling of term to be able to write a text.

In other words, the first survey tests how the 10 doctors act in cognitive situations (Task 1), in operative situations (Tasks 2–4) and in communicative situations (Task 5), cf. also Tarp (2011). Furthermore, Fuertes-Olivera & Tarp (2014: 87) argue that the lexicographical process seen from the user’s perspective can be divided into three fundamental phases:

1. extra-lexicographical pre-consultation phase
2. intra-lexicographical consultation phase
3. extra-lexicographical post-consultation phase

The first survey thus primarily covers the intra-lexicographical consultation phase and the extra-lexicographical post-consultation phase.

Second, the article draws on the insights and conclusions from another intra-consultation survey of the consultation behaviour of five 13-year-olds. The five teenagers were asked to look up five terms from an official text used for testing the English proficiency levels of Danish students by means of an iPhone 4S with the dictionary app *Gyldendal Engelsk-Dansk*. In this survey, the iPhone was also wirelessly connected to a PC by means of Reflector, cf. <http://www.airsquirls.com/reflector/>. The five students were asked to participate in two tests. Test A investigated how the five 13-year-olds accessed bilingual dictionary data while sitting down at a desk. Test B looked at how the five 13-year-olds accessed the same bilingual dictionary data while walking around a table, thus alluding to a mobile user situation. Both tests were recorded while the tasks were performed both from the “inside” by means of Reflector, and at the same time the user activities were recorded from the “outside” by means of a digital camera.

The five teenagers were asked to look up the following five terms.

Task 1: Look up “wildlife programmes” – to translate into Danish

Task 2: Look up “cheetahs” – to translate into Danish

Task 3: Look up “fancy it” – to translate into Danish

Task 4: Look up “auntie” – to translate into Danish

Task 5: Look up “disappointed” – to translate into Danish

In other words, the second survey tests how the five teenagers act in communicative situations (Tasks 1–5) during primarily the intra-lexicographical consultation phase and the extra-lexicographical post-consultation phase.

The two surveys thus included a total of 10 medical doctors and five teenagers. The empirical data of the first survey comprises 20 internal recordings, 20 external recordings, 20 think-aloud data recordings and 10 interview data recordings. The empirical data of the second survey comprises 10 internal recordings, 10 external recordings and 10 think-aloud data recordings.

3. The DNA of mobile lexicography

Before discussing the mobile user situation and the challenges and opportunities of mobile lexicography on the basis of the insights and conclusions from the two surveys, we first need to outline six dimensions, which dictate and constitute the basic framework of mobile lexicography. The six dimensions are the mobile device as a lexicographic medium, the mobile lexicographic data, the mobile user, the mobile user situation, the mobile lexicographic task and the mobile access method (cf. also Simonsen, 2014: 249–262).

First, what characterizes a mobile device? According to Budiu (2015) and Simonsen (2014), the small screen and the size of the mobile device make it hard for users to access, understand, process and remember information on mobile devices. Furthermore, the size and the portability of the mobile phone make it hard for users to stay focused. According to Budiu (2015), the portability of mobile phones also means that attention is fragmented and sessions very often short and punctual. Furthermore, it is also twice as hard to understand mobile content compared to online content (cf. Budiu, 2015), so therefore mobile content should leave out any filler content and unnecessary information. Budiu (2015) also argues that there is an inherent problem with the size of the touchscreen keyboard, because it is hard to type proficiently on a mobile phone. This argument is supported by Simonsen (2014), who also found that medical doctors often experienced problems when typing during search operations on a medical dictionary app. In fact, one medical doctor specifically referred to the fact that the touchscreen was too small and his fingers were too large. All these characteristics of the mobile device contribute to the cognitive load of the user; and we have not yet even considered the DNA of the lexicographic data.

Second, what characterizes lexicographic data? The information density of lexicography is high and very often lexicographic articles are quite long and comprehensive. It is in the DNA of lexicography to give the user precise, but often also long definitions, examples, synonyms, idioms, etc. The complexity is even higher in bilingual dictionary apps. Furthermore, many dictionary apps are unfortunately merely abridged app versions of the paper version. This argument is also made by Tarp (2015: 17), who argues that “However, in spite of the existence of a number of relevant techniques to improve the lexicographical product, the overwhelming majority of e-dictionaries still present themselves as paper or paper-like dictionaries with traditional, static articles, which have been placed on digital platforms without taking the necessary steps towards a completely new generation of dictionaries much more adapted to the users’ real needs in each situation”. Many dictionary apps do feature Google-like search-as-you-type search functions, but the user still interacts with the mobile device by means of a very small touchscreen keyboard. The small screen also means that content is not easily accessed and processed. Lexicographic content thus needs to be revised and abridged for dictionary app purposes; otherwise the mobile user will suffer from information overload.

Third, the characteristics and backgrounds of the users play a paramount role. The test persons involved in the two surveys discussed below comprise both digital immigrants and digital natives (see Prensky, 2001 for an outline of the terms digital natives and digital immigrants). As outlined above, the test persons can also be divided into professionals (medical doctors) and non-professionals (teenagers) and – as will become apparent from the discussion below – the backgrounds, competence sets and experience levels of the users almost dictate the way they access data and process information. The 10 medical doctors could be described as digital immigrants and they still prefer accessing medical data on a computer screen. However, the five 13-year-olds are digital natives and have all grown up in a hyper-connected world, and they prefer accessing virtually everything on mobile devices. The surveys seem to indicate that digital natives in comparison to digital immigrants are impatient and surprisingly illiterate when it comes to basic reference and dictionary skills, i.e. they have never really learned how to use a dictionary. In conclusion, the characteristics and backgrounds of the users are important to keep in mind when designing dictionary apps.

Fourth, the actual user situation is crucial. Dictionary apps are utility tools designed and developed to be used (cf. also Wiegand, 1988), and they must be designed and developed to suit the different user situations in which the users operate. Clearly, the user situation has an important impact on the selection of lexicographic data to be shown and the type of access method by means of which the user should access lexicographic data.

Fifth, the type of task that the user is solving also plays an important role in mobile lexicography. Dictionary apps are utility tools, and utility tools are used to solve

specific tasks. The empirical data, which will be discussed below, also show that different tasks call for different data sets and different access methods are required when using a dictionary app, for example, to translate a word or to save a person's life in an ambulance or in an emergency. In other words, the task dictates a number of factors in mobile lexicography.

Finally, the way users access lexicographic data in dictionary apps is also important to keep in mind when discussing mobile lexicography and designing dictionary apps. The two dictionary apps tested in the two surveys differ considerably. The *Gyldendal Engelsk-Dansk* app is a standard bilingual dictionary app based on the well-proven Gyldendal dictionary concept used by almost all students in Danish schools. The *Medicin.dk* app is a medical dictionary app designed and developed for health care persons. The *Gyldendal Engelsk-Dansk* app does not have a search-as-you-type search function. The *Medicin.dk* app does, and it even allows the user to tailor-make which data categories to show. This feature is very useful for users, because they can tailor-make the amount and type of data that they need. Another feature offered to the users of the *Medicin.dk* app is the scan feature utilizing the camera of the mobile device. In fact, paramedics or emergency doctors use the scan feature of the *Medicin.dk* app to determine the type of medicine digested in situations where patients are suffering from poisoning and where doctors need to make quick decisions. In conclusion, different access methods are needed in different situations to solve different tasks.

4. Results and Discussion

First, a brief description of the two surveys and the tests performed is relevant. Figures 1 and 2 below show a 62-year old medical doctor (TP5) being tested during Test A (while sitting down at a desk) and during Test B (while walking around a hospital bed).

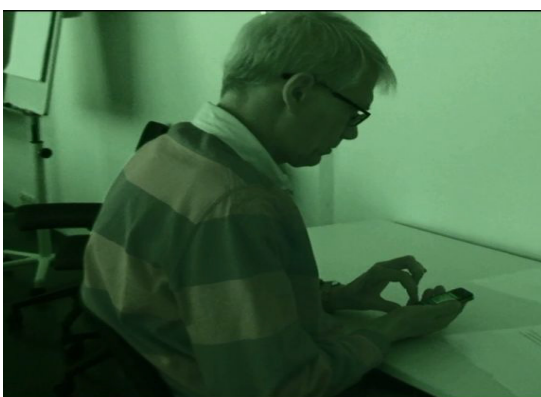


Figure 1: Survey 1 - Test A: Stationary Test



Figure 2: Survey 1 - Test B: Mobile Test

Figure 3 below shows a user situation with the same 62-year old medical doctor. Figure 3 shows the user situation seen from both the inside and the outside and is an edited figure of two video recordings. Figure 3 shows how TP5 sits at the table in the

left hand side of the picture interacting with the mobile device, and in the right hand side of the picture TP5's search behaviour on the iPhone is recorded and shown from the inside.

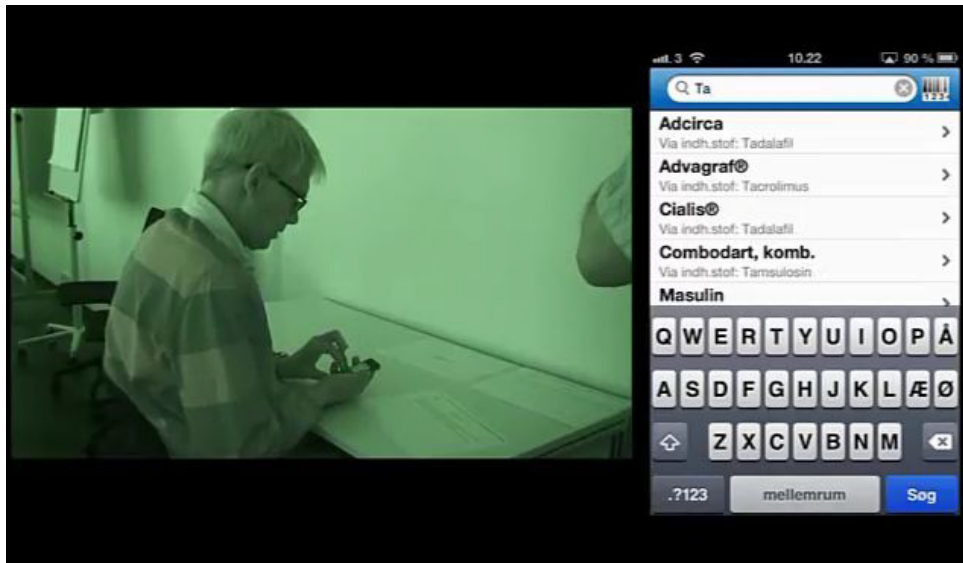


Figure 3: Survey 1 - Test A: Outside vs. Inside

Figures 4 and 5 below show a 13-year-old test person (TP15) being tested during Test A (while sitting down at a desk) and during Test B (while walking around).



Figure 4: Survey 1 - Test A: Stationary Test



Figure 5: Survey 2 - Test B: Mobile Test

Figure 6 shows TP15's user situation seen from both the inside and the outside. Figure 6 shows how TP15 sits at the table in the right hand side of the picture

interacting with the mobile device, and in the left hand side of the picture TP15's search behaviour on the iPhone is recorded and shown from the inside.

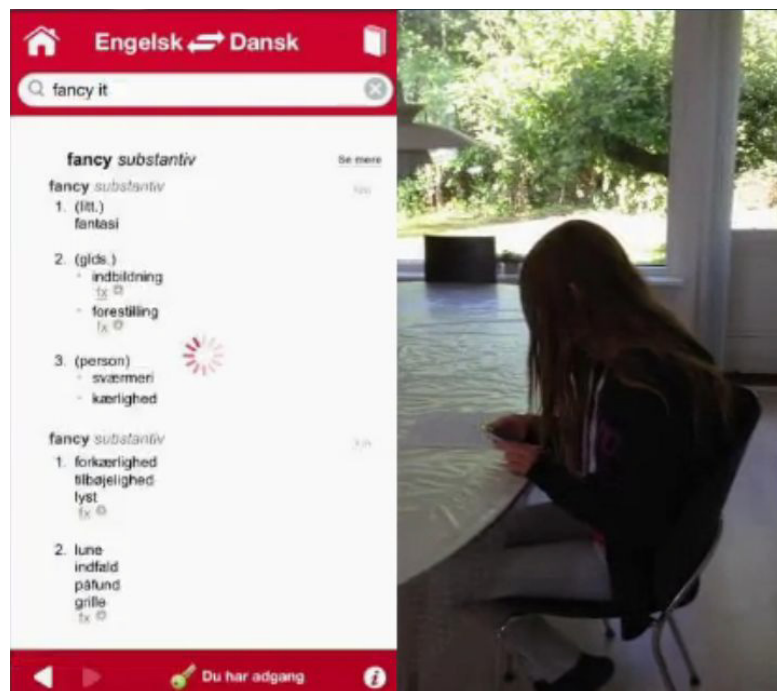


Figure 6: Survey 2 - Test A - Outside vs. Inside

A general observation on the basis of the data is that search speed, search quality, and ability to focus and interact with the mobile device was higher during the stationary user situation than during the mobile user situation. The digital natives were marginally quicker interacting with the device than were the digital immigrants, but they also seemed to have poorer reference skills.

The discussion of the data and the results will be based on data relevant to the six characteristics of mobile lexicography: the mobile device, the lexicographic data, the mobile user, the mobile user situation, the mobile task and the mobile access method.

4.1 The Mobile Device as a Lexicographic Medium

For decades the limitations and opportunities of both paper and online dictionaries have been discussed (e.g. Almind, 2005). Now, a new lexicographic medium is used and theoretical considerations on the characteristics of the mobile phone as a lexicographic medium are needed. No doubt the limitations and opportunities of the mobile device are relevant when discussing mobile lexicography. The trend in mobile telephones is that touchscreens are getting bigger, but the trade-off between portability and size still means that size is limited. A number of relevant considerations on mobile user surveys, mobile devices and interaction with a mobile device during movement can be found in Budiu & Nielsen (2013); Budiu (2015); Cerejo (2012); Church (2009); and Google (2013).

However, in the field of lexicography, only a few contributions have been published (including, in particular Curcio, 2014; Marellò, 2014; Simonsen, 2013; Simonsen, 2014), which each offer a number of theoretical considerations on how mobile users consult and use different dictionary apps.

The two surveys upon which this discussion is based do however seem to indicate that interacting with a mobile phone such as the iPhone 4S is difficult. Both surveys show that interacting with a mobile phone during movement is possible, but difficult, because the user both has to navigate in the search functions on the touchscreen and in the physical world at the same time.

Survey 1 tested 10 medical doctors in two user situations, and when I asked TP5 “Do you use your mobile device while moving?” he said “No – not really. I mostly use my mobile phone when I am sitting down because I think the screen is too small and my fingers are too big for the touchscreen”. TP5 can be seen in Figures 1–3 above, and at the time of the test he was a 62-year old medical doctor. He was the oldest test person among the 15 people tested, which seems to indicate that age plays a role in mobile information access behaviour. This in fact corresponds with the discussion of digital natives vs. digital immigrants (cf. Prensky, 2001). The 5-inch screen on a standard smartphone such as the iPhone 4S is simply not enough. Size does matter when it comes to successful data access and information processing. The design of dictionaries has always been relevant for lexicography (e.g. Almind, 2005), but when it comes to mobile lexicography there is still much to be done.

The input device (the finger) and the small letters displayed on a 5-inch screen are not a perfect match as one of the test persons surveyed actually pointed out. The data from TP7 and TP8, who chose to hold the mobile device horizontally, show that they in fact were quicker and better at locating information. A similar conclusion can be made on the basis of Survey 2, which included five teenagers. The digital natives (the teenagers) were no doubt quicker than the digital immigrants (the doctors); however, they also used the backspace button all the time, indicating that they might be quick at interacting with the device, but that they made a large number of typos. All five teenagers held the mobile device with both hands during movement while they typed with their thumbs. Observations from the outside during both surveys indicate that the majority of users hold the mobile device in a vertical position allowing them to use both thumbs while either sitting or walking. Observations from the inside during both surveys indicate that the majority of users make a large number of typos and that they use the backspace button to delete and retype. Other observations indicate that the autofill function of the iPhone 4S is not a help but more a source of frustration. Only TP14 and TP15 use the pinch and pan gesture and the magnifying glass to make it easier to select the type of information they want and both TP14 and TP15 are digital natives.

In conclusion, the physical characteristics of a mobile phone must be taken into consideration when designing dictionary apps. The size and the user situation make it

impossible to access information the same way we do in electronic dictionaries, for example. Consequently, we need to carefully select the type and amount of dictionary data to show and even leave out data. This will be discussed in detail below.

4.2 Lexicographic Data on Mobile Devices

The type and amount of lexicographic data to be included in dictionary apps is a new discussion. In fact, it is argued that this discussion is of paramount importance, because users may otherwise suffer from information overload; see also Tarp (2015: 17) who eloquently argues that “One of the major problems in past and present dictionaries is information overload...”. The fact that data overload may obstruct and even hinder both access to the relevant data and retrieval of the required information from these data, (cf. also Bergenholtz & Gouws, 2010) has been empirically demonstrated in these surveys. In fact, the discussion was started by Simonsen (2014), who proposes four principles of mobile lexicography. One of the principles is called “Mobile Data Principle”. Simonsen (2014: 260) argues that “The mobile user situation also dictates the type and complexity of the mobile data. The size of the user interface and the punctuality of the user situation mean that complex data and long text segments are not an optimum way of displaying mobile data”.

The data from the surveys support the argument that data overload may obstruct and even hinder both access to the relevant data and retrieval of the information required from these data (cf. Bergenholtz & Gouws, 2010). Nielsen (2011) argues that “if in doubt – leave it out” and empirically proves that “writing for mobile readers requires even harsher editing than writing for the web”. The two dictionary apps tested in this article clearly contain way too much information in a number of situations, and it can be argued on the basis of my own empirical data that some information overload does in fact take place, especially in *Gyldendal Engelsk-Dansk*. Sometimes you get the impression that publishing houses publish dictionary apps simply because everybody else does and that include as much lexicographic data as possible. The question of information overload is discussed by Tarp (2015: 17) who uses the following terms to describe information overload:

“absolute overload”, which takes place if there are more data than required to meet the users’ needs

“relative overload”, which takes place if there are more data than can be visualised without scrolling down or than the predicted user can be expected to overview

“functional overload”, which is a case of absolute data overload when it relates to the needs of a specific user in a specific type of situation

“concrete overload”, which is a case of absolute data overload when it relates to the needs that a concrete, individual user may have in a concrete situation.

In fact, I argue that all four types of information overload can be demonstrated using empirical data. Less is in fact more sometimes, and it is argued that the characteristics of the mobile device, the characteristics of the mobile user, the size of the user interface and the complexity of the mobile user situation may sometimes have been sacrificed on the altar of lexicographic and technical perfectionism.

The dictionary app tested in Survey 1 was a medical dictionary app developed for health care professionals (HCPs). Figure 7 below shows three screen dumps from the app.

As will appear from the circled spot in the screen dump to the left, the dictionary app features a very useful “search-as-you-type” feature similar to that used by Google. The centre screen dump shows a standard display of the search result, but as will appear from the circled spots the user can tailor-make what and how much lexicographic data he wants when he clicks “Min visning” (My profile). The circled spots in the screen dump at the right show how the user may select the type of lexicographic data he needs the next time he uses the dictionary app. This sort of situational adaptation is a step forward in mobile lexicography and resembles principles 1, 2 and 6 described by Fuertes-Olivera & Tarp (2014: 64), because the customization allows the user to avoid information overload, to access the data required in each consultation and finally ensures that the article contains no more than needed.

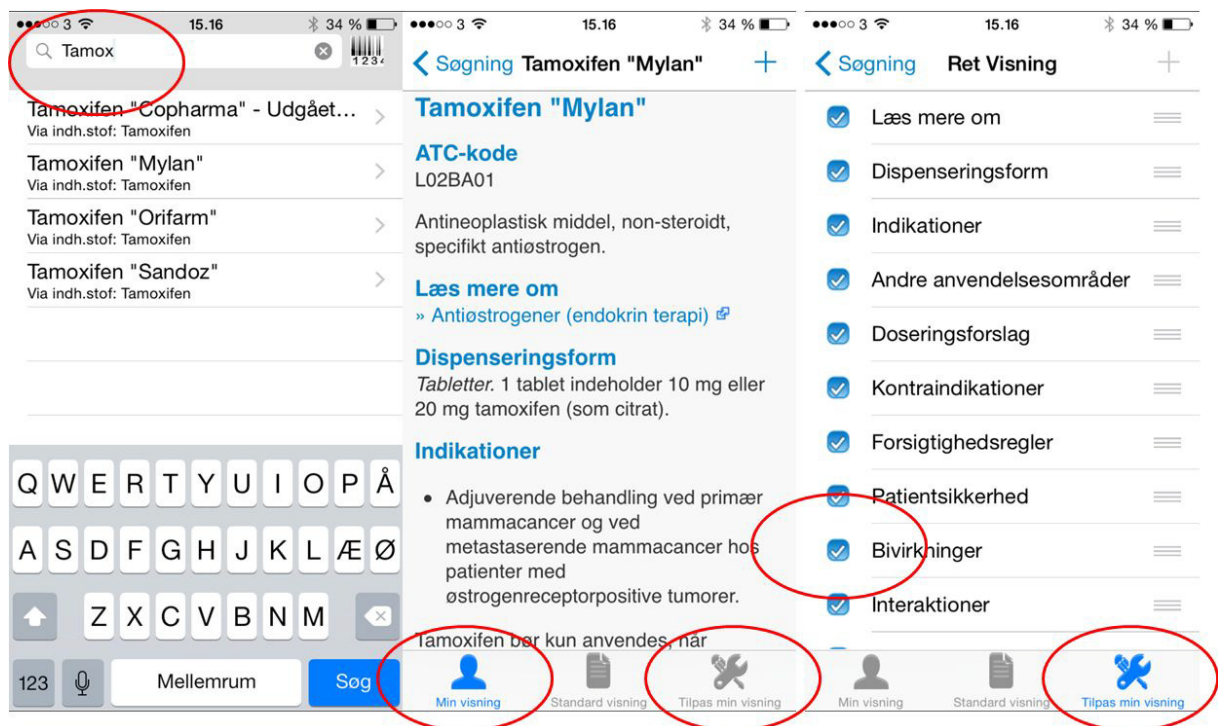


Figure 7: Medicin.dk

Observations from the inside reveal that the 10 doctors quickly find and access the article they need, primarily because of the powerful search-as-you-type feature. When

they look for a specific type of information, for example information on side effects (Bivirkninger), they quickly scroll down to the lexicographic data type needed by navigating on the basis of the bold, blue headlines. The user situation and the actual task also affect the type of data needed. As will be discussed below, the mobile user situation is characterized by being volatile and punctual. The mobile user typically checks knowledge and performs simple searches. The mobile user situation primarily supports simple, punctual, communicative lexicographic functions, but is not suited to support complex, cognitive lexicographic and bilingual communicative functions.

Recordings from the inside of the consultation behaviour of the five teenagers indicate that information overload does take place and that this information overload in fact hinders both access to the right type of data and the extraction of the required information. Figure 8 below shows a number of screen dumps from the dictionary app *Gyldendal Engelsk-Dansk*.

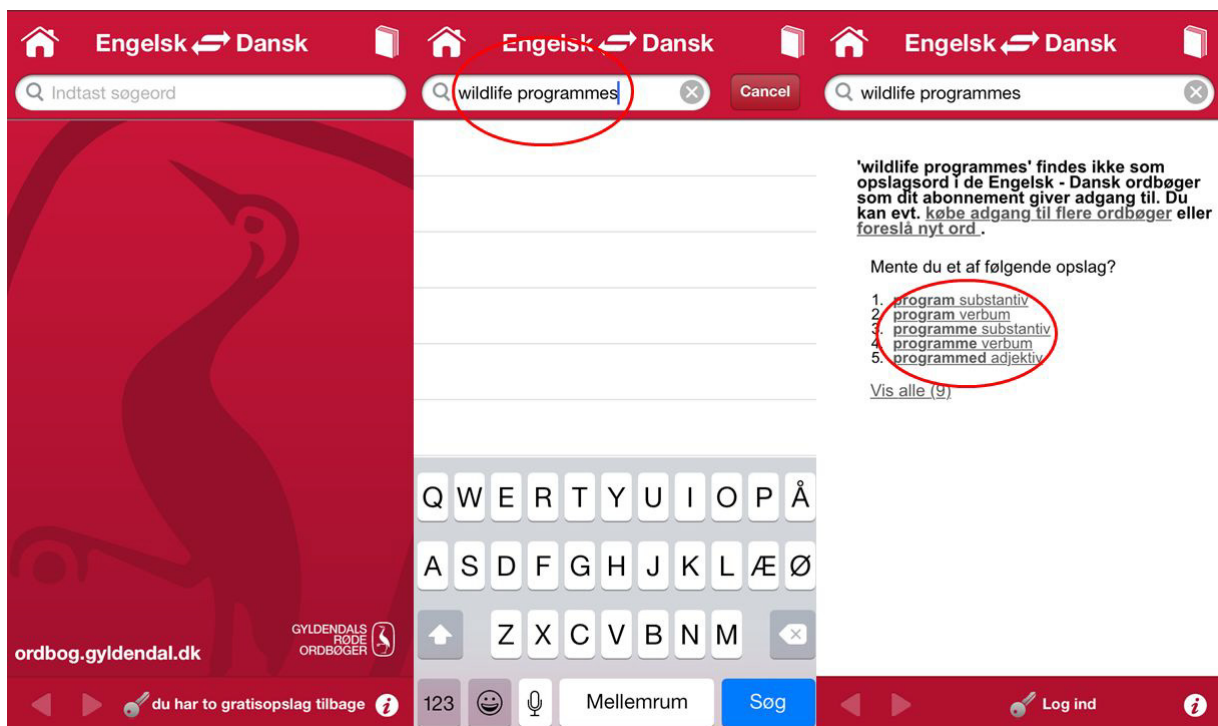


Figure 8: *Gyldendal Engelsk-Dansk*

This dictionary app does not offer a search-as-you-type feature, which is unsatisfactory if the primary user group (students) is borne in mind. A search-as-you-type feature seems to be a standard solution in mobile lexicography, cf. for example *Merriam-Webster Dictionary App (MW)*, *Den Danske Ordbog (DDO)*, *Advanced English Dictionary and Thesaurus (AEDT)* and *Ordbogen.com (OC)*, etc. The recordings from the inside clearly show that users make a lot of typos, and that the consultation process is negatively affected because users have to use the backspace button all the time. The recordings also show that the *Gyldendal Engelsk-Dansk* app, in some situations, seems to display way too much data and that some data should be

offered earlier in the consultation process.

Obviously, this may have to do with the argument that the five teenagers tested seem to lack basic reference skills, but the empirical data also show that the five digital natives search as they would on Google and it seems as if they expect a search-as-you-type feature. TP11, TP12, TP13 and TP14 all typed “wildlife programmes”, that is, they entered a multiword item in the search field and clicked search to find the translation. Only TP15 performed a search for “wildlife” and then “programmes”. So it seems that the digital natives expect a search-as-you-type feature.

Furthermore, the recordings from the inside show that the teenagers do not explore the possibilities of the *Gyldendal Engelsk-Dansk* app. Even though the app suggests a number of possible meanings, none of the five digital teenagers used this feature. Not even when the app actively asked “Do you mean one of the following terms”, did they explore further possibilities. TP11, for example, entered “wildlife programmes” in the search field and even though the app suggested a number of options, she did not click any of them. Instead she deleted what she wrote in the search field and entered the word “wild” and subsequently the word “wildlife”. In conclusion, the empirical data support the argument made above that too much information may both hinder access to the right data and extraction of the information required, because none of the teenagers except TP15 came up with the right Danish translation of “wildlife programmes”. The next step in this discussion is to look at the characteristics of the mobile user.

4.3 The Mobile User

Wiegand once called the user the “Bekanntes Unbekanntes” (Wiegand, 1988), but it is argued that we now have much more knowledge of who the user actually is. A number of relevant theoretical contributions have discussed how mobile dictionary users use different dictionary apps (for example Curcio, 2014; Marellò, 2014; Simonsen, 2013; Simonsen, 2014). Simonsen (2014) describes the mobile user as follows: “The mobile user is on the move and needs and accesses information while on the go. This makes the mobile user punctual, impatient, imprecise and preoccupied with other things”. Background, education, age and experience level of the user play a paramount role in all types of information access discussions. The test persons involved in the two surveys can be divided into professionals (medical doctors) and non-professionals (teenagers); into digital immigrants (medical doctors) and digital natives (teenagers); into educated and experienced (medical doctors) and uneducated and inexperienced (teenagers); and into old (medical doctors) and young (teenagers). Obviously, the user’s background, competence set and experience level almost dictate the way they access data and process information. This is also evident from the empirical data. As already discussed above, the digital natives seem to be really impatient and lacking reference skills. Only TP15 chose to explore the additional suggestions offered by the app while the other four test persons ignored the full

potential of the app. Another general observation is that mobile users *per se* are mobile and able to move around. This very fact makes them sporadic and impatient multi-taskers, which means that accessing data on a mobile device is not the same as accessing data on a 17-inch computer screen. The empirical data produced in the two surveys also indicate that consultation behaviour is naturally individual and dependent upon the task. The emergency doctor prefers the mobile device and loves accessing medical data on the mobile device because she uses the app at emergency sites or in the ambulance. The characteristics of the mobile user situation will be the topic of the next section of this article.

4.4 The Mobile User Situation

As already argued the mobile user situation affects a number of dimensions. The data show that there is a significant difference between the two user situations, sitting (Test A) and moving (Test B), when it comes to access speed; that is, from the moment the test person started the data access operation to the moment he ended the search operation. A dictionary app is no doubt a utility tool designed and developed to be used in specific situations and, according to Tarp (2011), online dictionaries should be developed to help users perform activities in four situations:

1. In communicative situations, to listen to – and to read, write or translate oral and written texts in specific professional situations
2. In cognitive situations, to store information and learn about the profession (theories, methods, etc.) and about carrying out professional activities
3. In operative situations, to perform specific activities and solve problems in specific situations
4. In interpretive situations, to interpret and extract information from opaque, non-verbal signs such as figures, graphs, visual illustrations etc. that are used as information units in texts in specific professional situations, or as independent items.

The two surveys in this paper cover the first three situations and show that it does make a difference whether a dictionary app is used professionally or in school, or when sitting down or walking and that the user situation does affect which data are accessed, how data are accessed and how information is extracted from the data and used. Simonsen (2014) argues that “the mobile user situation is characterized by being volatile, punctual and by often taking place while the user does other things. The mobile user typically checks knowledge and performs simple searches. The mobile user situation primarily supports simple, punctual, communicative lexicographic functions, and is not suited to support complex, cognitive lexicographic functions”.

The data clearly substantiate this argument. The data seem to indicate that the mobile user situation primarily supports simple, punctual, communicative lexicographic functions, but that mobile devices and dictionary apps are also suitable

in operative situations, for example when an emergency doctor needs to find a medical product and decide what does to dispense to the patient.

The data also show that mobile lexicography is not a perfect match when it comes to heavy cognitive situations, where users are researching a specific complex question. In Survey 1, it was found that the information access success of the 10 medical doctors was reduced in cognitive user situations, especially Tasks 2, 3 and 4, which were all about locating complex information with a view to making decisions as to side effects, dosage and how to take the medicine, etc. In fact, TP7 stated during the follow-up interview that “If I have to look a little bit deeper into a question then I clearly prefer the computer. I would definitely use the computer if I were to prescribe medicine that I have never used before”. In other words, the mobile user situation and cognitive lexicographic functions does not make a perfect match.

In conclusion, the user situation has an important impact on the selection of lexicographic data to be shown and the type of access method by which the user should access lexicographic data. This question will be addressed in the next section of this article.

4.5 The Mobile Lexicographic Task

The mobile lexicographic task that the user is solving constitutes perhaps the most important dimension. Apps are utility tools and are designed so that the user can solve specific tasks. And different tasks call for different tools, etc. Unfortunately, the importance of the task has so far received little attention in lexicography, but it is argued that the task which the user is solving is of paramount importance for a number of aspects.

The data harvested during the two surveys also suggest that there is a clear connection between the user’s competence set, the task that the user is solving, the way the user prefers to access the data and last, but not least, the type of data the user needs. One example from Survey 1 reveals that a paramedic doctor uses the *Medicin.dk* app differently than do, for example, the hospital doctors. When asked “Which platform and user situation do you prefer?”, one of the hospital doctors said “I prefer the website version of *Medicin.dk*, if my problem is complex. The app and the iPhone are handy, if I suddenly have a problem that I know can be solved by using the app. However, if I need more in-depth knowledge I would rather use the website”. On the other hand, the test person working as an emergency doctor stated that “I prefer the app and I noticed that using it comes naturally for me, because I use it all the time. As an emergency doctor the app is much better. It is quicker and I do not have the time to use the website version”.

Such choices are in fact only natural. When you want to hammer a nail into wood

you use a hammer. The task dictates that you use a hammer. The task comes first – not the tool, which in fact is also the essence of the popular expression “If all you have is a hammer, everything looks like a nail”. In other words, if the tool you have is limited, simple-minded people (users?) apply the tool inappropriately. It is argued that this is what sometimes happens in mobile lexicography.

As will be evident from Figure 7 above, the user searched for a medicinal product called Tamoxifen. The autofill search function also works in the app as shown in the left-hand screen dump. If the user wants to tailor-make the data structuring of the app he can open the actual article as shown in the middle screen dump and click the option “Min visning” (My profile). Then a customization window appears as shown in the right-hand screen dump, and the user can select the data he wants. In other words, an oncologist for example may first of all select the groups of medicinal products that he often prescribes, and which is recommended in the treatment guides. Second, he can select the exact types of data that he needs when solving different tasks. If, for example, the doctor is going to inform a breast cancer patient about possible side effects, he may choose to enable “Bivirkninger” (side effects) and disable all other data types. In other words, you use the tool required to solve a specific task. Tarp (2014: 17) argues for the use of mono-functional dictionaries to avoid functional overload and for the development of personalised dictionary tools to avoid concrete overload and, as shown in Figure 7, this is in fact possible in the medical dictionary app *Medicin.dk*.

4.6 The Mobile Access Method

The way users access data is yet another important dimension when discussing mobile lexicography. According to Simonsen (2014: 260) “the mobile user navigates in both the physical world and in the user interface of the mobile device at the same time. This calls for a very simple and easy-to-use data access method, for example a very intelligent search engine or even better a voice-activated search engine like Siri in an iPhone”.

The data seem to suggest that simple search-as-you-type search engines with a large search field are preferred by most users: Budiu (2015) argues that content and prioritization are extremely important issues to take into account on mobile devices. Scrolling through large text blocks reduces the information access success of users and as data from Survey 2 indicates, users do not explore the many possibilities of standard dictionary apps.

During the two surveys the 10 medical doctors and five teenagers exclusively used a semasiological data access method of typing letters in the search field. All test persons used this access method, probably because it is the most natural access method for most users, even though other ones are possible. Figure 9 below shows a section of the search fields in the two apps tested *Medicin.dk* and *Gyldendal Engelsk-Dansk*.

Both apps feature a standard search field of 4 cm x 0.5 cm, and as data from the two surveys show it is in fact quite difficult for both digital immigrants and digital natives to type the right letters by means of the touchscreen and at the same time monitor the correct spelling. That is why a search-as-you-type search feature is so important in mobile lexicography.

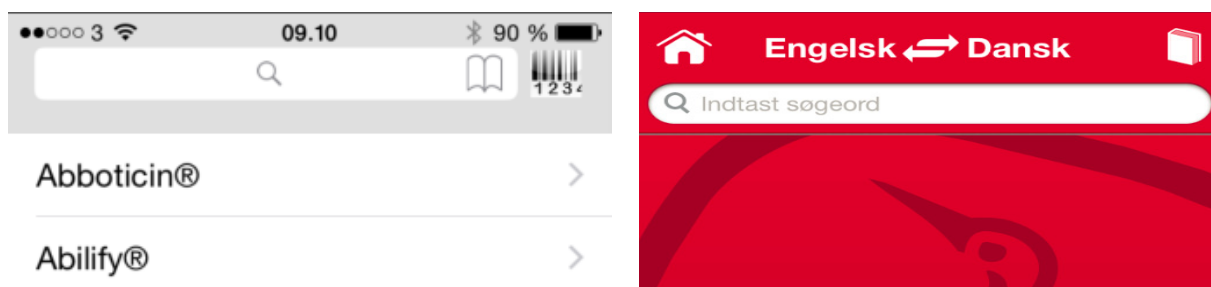


Figure 9: Search fields in *Medicin.dk* and *Gyldendal Engelsk-Dansk*

None of the 15 test persons used an onomasiological access method for looking up on the basis of concepts, etc. The medical dictionary app *Medicin.dk* does in fact offer a bookmark feature, where users can store frequently-used look-ups, just as the app allows users to access information on reimbursement, dispensation of medicine, etc. Finally, the app *Medicin.dk* also features an optical character recognition feature whereby health care persons can use the inbuilt camera of the mobile device to scan the bar code of medicinal products and this way check the type of medicine being administered to a patient.

The method by which users access lexicographic data on mobile devices is no doubt an area where more research is needed. As demonstrated above, users find it relatively hard to type correctly simply because the touchscreen is too small compared to the size of the index finger and thumb. At the same time users are often mobile when using mobile devices, thus rendering it even harder to type on the touchscreen and simultaneously navigate in the physical world. Consequently, new access methods and technologies are needed and one of the most promising solutions might be a voice-activated access method like Siri in most iPhones.

Too much focus on a single aspect in a complex situation very often results in failure. Other researchers have discussed this dilemma (e.g. Verlinde et al., 2010; Simonsen, 2011; Simonsen, 2013; Simonsen, 2014; Tarp, 2015 to mention just a few). Verlinde et al. (2010: 5) make a case for a “Lexicographic Triangle”, Simonsen (2011) proposes the “Information Scientific Star Model”, and Tarp (2015) argues for a back to basics approach where a mono-functional solution is recommended.

The above discussion can be illustrated in the hexagon model for mobile lexicography given in Figure 10.

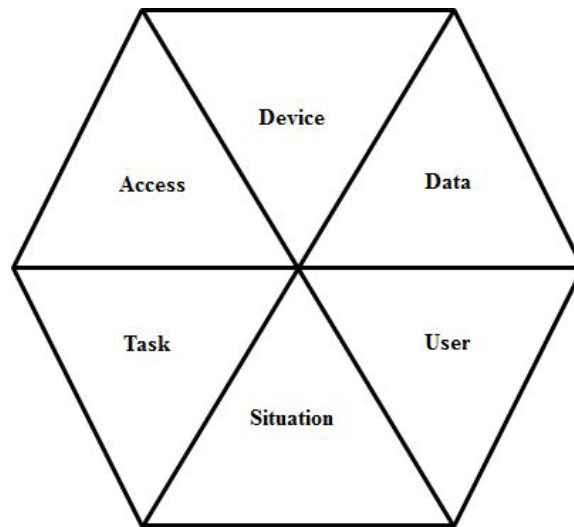


Figure 10: Mobile Lexicography Model

5. Conclusion

In this article the DNA of mobile lexicography has been discussed and a model for mobile lexicography proposed. Users have already gone mobile and to avoid the different types of information overload discussed by Tarp (2015), new more balanced solutions are required. All six dimensions discussed above should be taken into account. So no more lexicographic data dictatorship! No more user dictatorship!

What mobile lexicography needs is a balanced distribution of power whereby all six dimensions are calibrated vis-à-vis each other. The hexagon model proposed above illustrates that all six dimensions are interconnected, and it is argued that the hexagon model may enable lexicographers to design better dictionary apps.

This article has demonstrated how doctors and students use two different dictionary apps and has proposed a number of theoretical considerations regarding mobile lexicography.

Lexicographic innovation is required. Now is the time to do it right, otherwise lexicography as a discipline may die from a fatal “*identity crisis*”, as Tarp (2015: 16) argues. Therefore, much more research in mobile lexicography is needed and timely; because users have already gone mobile.

6. References

- Almind, R. (2005). Designing Internet Dictionaries. *Hermes, Journal of Linguistics*, 34, pp. 37-54.
- Bergenholtz, H. & Gouws, R. H. (2010). A new perspective on the access process. *Hermes. Journal of Language and Communication in Business*, 44, pp. 103-127.
- Budiu, R. & Nielsen, J. (2012). *Mobile Usability*. New Riders Press. Berkeley.
- Budiu, R. (2015). *Mobile User Experience: Limitations and Strengths*. In: NN/g Nielsen Norman Group: Accessed at: http://www.nngroup.com/articles/mobile-ux/?utm_source=Alertbox&utm_campaign=205de653eb-Mobile_UX_long_04_20_2015&utm_medium=email&utm_term=0_7f29a2b335-205de653eb-40153273 [21/04/2015]
- Cerejo, L. (2012). The elements of the mobile user experience. *Mobile design patterns* (1st ed., pp. 5-20). Freiburg, Germany: Smashing Media GmbH.
- Church, K. & Smyth, B. (2009): Understanding the intent behind mobile information needs. In: *IUI 2009 International Conference on Intelligent User Interfaces*, pp. 247-256.
- Curcio, M. N. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In: *Proceedings of the XVI EURALEX International Congress: The User in Focus 15-19 July 2014, Bolzano/Bozen*. Accessed at: <http://www.eurac.edu/en/research/autonomies/commul/Publications/Pages/default.aspx> [21/04/2015].
- Google (2013). *Our Mobile Planet: Denmark – Understanding the Mobile Consumer*. Accessed at: <http://services.google.com/fh/files/misc/omp-2013-dk-en.pdf> [22/04/2015].
- Marello, C. (2014). Using Mobile Bilingual Dictionaries in an EFL Class. In: *Proceedings of the XVI EURALEX International Congress: The User in Focus 15-19 July 2014, Bolzano/Bozen*. Accessed at: <http://www.eurac.edu/en/research/autonomies/commul/Publications/Pages/default.aspx> [21/04/2015].
- Müller-Spitzer, C. (2013). Contexts of dictionary use. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 1-15.
- Nielsen, J. (2000). Why You Only Need to Test With 5 Users. In: NN/g Nielsen Norman Group: Accessed at <http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/> [21/04/2015].
- Nielsen, J. (2011). *When in doubt, leave it out*. In: NN/g Nielsen Norman Group: Accessed at <http://www.nngroup.com/articles/condense-mobile-content/> [21/04/2015].
- Nielsen, J. (2012). How Many Test Users in a Usability Study? In: NN/g Nielsen Norman Group: Accessed at <http://www.nngroup.com/articles/how-many-test-users/> [21/04/2015].

- Prensky, M. (2001). Digital natives, digital immigrants part 1. *On the Horizon*, 9(5), 1–6: Accessed at <http://www.emeraldinsight.com/journals.htm?issn=1074-8121> [01/04/2014].
- Simonsen, H. K. (2013). Brugerne er allerede mobile! In. *Nordiska studier i lexicografi 12 – 2013*, pp. 416-429.
- Simonsen, H. K. (2014). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014, Bolzano/Bozen*, pp. 249-261.
- Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In P. A. Fuertes-Olivera & H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London, New York: Continuum, pp. 54-70.
- Tarp, S. (2012). Theoretical challenges in the transition from lexicographical p-works to e-tools. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 107–118.
- Tarp, S. (2015). Detecting user needs for new online dictionary projects: Business as usual, user research or...? In: C. Tiberius & C. Müller-Spitzer (eds.) *Research into dictionary use/Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. - Mannheim: Institut für Deutsche Sprache. (erscheint in: OPAL - Online publizierte Arbeiten zur Linguistik 2015) <http://multimedia.ids-mannheim.de/mediawiki/web/images/7/7f/Preprint-V1.pdf> [21/04/2015]
- Verlinde, S., Leroyer, P. & Binon, J. (2010). Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats. *International Journal of Lexicography*, 23(1), pp. 1–17.
- Wiegand, H. E. (1988). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlin/New York: de Gruyter.

Websites:

- Reflectorapp.com (2015): Accessed at: <http://www.airsquirrels.com/reflector/> [21/04/2015]

Dictionary apps:

- Advanced English Dictionary and Thesaurus at App Store [21/04/2015]
- Den Danske Ordbog at App Store [21/04/2015]
- Gyldendal Dansk-Engelsk/Engelsk-Dansk at App Store [21/04/2015]
- Merriam-Webster Dictionary App at App Store [21/04/2015]

Ordbogen.com at App Store [21/04/2015]

Pro.medicin.dk app at App Store [21/04/2015]

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



What can a social network profile be used for in monolingual lexicography? Examples, strategies, desiderata

Monika Biesaga

The Institute of the Polish Language at the Polish Academy of Sciences, Cracow

E-mail: monika.biesaga@interia.pl

Abstract

The aim of this paper is to introduce the phenomenon of social network tools used in contemporary European e-lexicography. Because of their central role in this field of lexicography the monolingual dictionaries of national and regional languages have been chosen as the corpus for this study. The analysis of the Lexilogos portal resources (namely an alphabetical list of the European dictionaries) has shown that social media tools are used in 21 dictionaries. Concerning the list of arguments to be presented, firstly, the linking of the dictionary website to social network profiles was analyzed (ways of linking: sharing and following, as well as some issues related to graphic matters). Secondly, the most important characteristics of social network profiles were introduced (number of users, frequency of entries, types of content and their marketing role). Thirdly, some of the advantages of lexicographical social networks were shown. In conclusion I have expressed the most important desiderata concerning lexicographical social media profiles.

Keywords: e-lexicography; social networks; linking dictionary resources; user-friendly lexicography

1. Introduction

Today for many of us it is hard to believe that just over 10 years ago there were no social network websites¹. During the last decade they have become revolutionary facilities which help to maintain private social contacts as well as to send and receive various types of other personalized information. Because of their functionality they are undoubtedly indispensable marketing tools, and are very often used to communicate with public users, not only by commercial companies, but also by various public institutions.

Therefore, it is not surprising that social media are also being used by lexicographers and others involved in dictionary projects (e.g. marketing specialists for the production of big commercial dictionaries). Because of the effectiveness of social media and the lack of common information regarding this type of lexicographical initiative, I have decided to create an inventory of social media tools² and particular

¹ For example Facebook was introduced in 2004, Twitter in 2006.

² To avoid repetition, social media are also being called social networks in this paper; the same thing relates to social media tools which are also described as social media functions or facilities.

profiles connected with these dictionaries.

This paper constitutes an introduction to the subject; therefore, it will contain basic information. This will help us to comprehend the matter and locate our own projects in the existing lexicographical networking universe. Firstly, I would like to focus on graphic matters concerning linking resources, namely how dictionary pages (main pages and particular entry pages, and if there are differences between them) are connected within social media profiles. In this paragraph I will also show the variety of social media facilities used for lexicographical purposes. Secondly, I will focus on existing dictionary Facebook profiles by indicating their main characteristics, such as the number of followers, the frequency of entries and most importantly, the thematic content, including ways of linking to various dictionary resources.

2. Inventory

The first and most challenging part was to create a homogenous inventory of dictionary projects connected with social media profiles and facilities. For that purpose I have chosen one of the biggest existing resources of worldwide dictionaries: the Lexilogos Internet portal. It contains links to hundreds of dictionaries gathered accordingly to various criteria. For the purpose of this analysis I used an alphabetical order of languages³. After a brief overview it became obvious that this portal, although very helpful and rich in terms of the content, is centred on European languages. Therefore, it cannot be used as a reliable source of information regarding worldwide languages⁴. Because of this factor as well as a linguistic barrier, I have decided to analyze only European dictionaries in this paper.



Figure 1: Europe Political Map by Aotearoa – Own work, CC BY-SA 3.0, Wikimedia Commons

This decision forced me to find a scientifically approved geographical division of the Earth`s continents. Therefore, in this paper I am using the Europe map specified by

³ http://www.lexilogos.com/dictionnaire_langues.htm

⁴ It is worth mentioning that the Working Group 1 from the European Network of e-Lexicography is preparing an exhaustive inventory of European academic dictionaries. Further information can be found at elexicography.eu.

the International Geographic Union (see Figure 1). If a country belongs partially to the European continent I also analyzed the corresponding linguistic resources in the Lexilogos (e.g. Turkey, Russia).

The Lexilogos profile contains links to the various types of dictionaries (monolingual⁵, bilingual, etymological etc.). Due to the homogeneity of the inventory, I have taken into account only monolingual and general dictionaries of contemporary European languages. In my opinion these types of lexicographical products create the central part of this lexicography. Their task is to transmit not only the language itself but also a kind of cognitive entity connected with the language. For the purpose of this inventory I have analyzed both dictionaries of the official languages as well as dictionaries of territorial languages (e.g. Asturian, Basque, Catalan).

Social multidictionaries such as Wikitionary, FreeDictionary and Wordreference were not included in the inventory because of their secondary nature and the lack of methodological basis.

As a result I have received an inventory consisting of 21 electronic dictionaries which use social media facilities⁶:

1. Cambridge Advanced Learner`s Dictionary
2. Chambers Free English Dictionary
3. Collins English Dictionary
4. Den danske Ordbog. Moderne Dansk Sprog (The Danish Dictionary. Modern Danish Language)
5. Dex Online. Dicționare ale limbii române (Romanian Dictionary)
6. Diccionario de la lengua Española (Spanish Dictionary), Diccionario esencial de la lengua Española (The Essential Dictionary of Spanish)⁷
7. Dicionário Priberam da Língua Portuguesa (Priberam Dictionary of Portuguese)
8. Dictionariu de la Llingua Asturiana (Dictionary of the Asturian Language)
9. ДИГИТАЛЕН РЕЧНИК НА МАКЕДОНСКИОТ ЈАЗИК (Digital Dictionary of the Macedonian Language)
10. Dizionario Treccani (Treccani Italian Dictionary)
11. Duden (German Dictionary)
12. Грамота.ру (Gramota.ru, Russian Dictionary)
13. Gran diccionari de la llengua catalana (Great Dictionary of Catalan)
14. Larousse Dictionnaire de Française (Larousse French Dictionary)

⁵ In the monolingual dictionary the lemmas from language x are defined with the words from language x; in the bilingual dictionary the lemmas from language x are defined with the words from language y.

⁶ This paper reflects the state of the art in May 2015, as for the analysis of Facebook entries gathered, data includes information from the last six months (XII 2014 – V 2015).

⁷ Both dictionaries are on the same website, they share the same social media tools.

15. Macmillan Dictionary
16. Oxford English Dictionary
17. Речник на думите в българския език (The Dictionary of Words in the Bulgarian Language)
18. SLex. Elektronický lexikón slovenského jazyka (SLex. Electronic Dictionary of Slovak)
19. Sproget (The Danish Dictionaries Portal), consists of among other resources: Den Danske Ordbog (The Danish Dictionary)
20. Van Dale (Dutch dictionary)
21. Wielki słownik języka polskiego (Great Dictionary of Polish).

3. General remarks

As we can see, social networks are used for linking resources not only in commercial dictionaries (Larousse, Oxford English Dictionaries, Van Dale etc.) but also in academic projects (Diccionario de la lengua Española, Dictionariu de la Llingua Asturiana, Wielki słownik języka polskiego). However, one must admit that the usage of social media is not common. Because of the enormous differences between European dictionary projects (financial background, number of employees, lexicographical tradition) I would not want to indicate the exact percentages. Conducting a profile is a relatively time consuming occupation. One must find the topic for a future Facebook or Twitter entry. It needs to be interesting for users and at the same time be connected with the particular dictionary resources (specific entry or a group of entries). It is not rare that after creating a social network post, users pose further questions, formulate remarks or express doubts, sometimes even involving themselves in some kind of dispute; therefore, constant attention by the administrator person is essential. Considering this, it is understandable that because of limited time or human resources, many dictionaries withdraw from using social media tools. In other kinds of projects, especially academic ones, user orientation does not exist while the main goal is to finish the project and satisfy scientific reviewers. To summarize, in my opinion, less than 10 percent of the European dictionaries linked to the Lexilogos use social media tools.

The most popular social networks are Facebook and Twitter (global tendency). The general rule is that if a project is commercial and relatively popular (in terms of the number of users), linking to social media is beneficial (e.g. sharing content buttons) and the lexicographic project itself consists of, aside from the website, a few social profiles. The less frequently used networks are also popular, however: Google+, Flickr, Instagram and YouTube. It is worth noting that in a few cases, the content of some dictionaries can be shared via national social networks, e.g. VKontakte for Russian (Gramota Dictionary), bgHot for Bulgarian (Rechnik Dictionary).

4. Linking from dictionary websites

If we look at the structure of dictionary websites, we can see that there are two ways

of linking lexicographical content to social media. One is sharing. In this case, somewhere on the page, usually at the top, we find buttons which enable us to share the content of the page on our private social network account. This facility is widely used especially in the case of particular dictionary entries (see Figure 1⁸).



Figure 1: Sharing buttons connected with the entry (*bois*), Larousse Dictionary.

We can also encounter pages where sharing buttons are located in the central part of the page (see Figure 2).



Figure 2: The main page of the dictionary with sharing buttons is in its central part, Rechnik Dictionary

⁸ To make figures more readable, commercial frames were framed with grey and filled with white.

Another alternative technique of sharing is observed in the Cambridge dictionary. On the main page we find a separate frame with a Word of the Day and additional sharing buttons (see Figure 3). As can be expected, it is not a random word but one chosen in advance by the lexicographers (for example the word must have only one meaning to fit to the frame). This strategy can be considered very useful because it gives the user the opportunity to read an entry he probably would not have searched for, but which might be interesting for him (in Figure 3, a rare verb *warble* is introduced that could enrich the user`s vocabulary). In this way, the dictionary team strives to maximize the attention of the user and promote additional entries apart from the one methodically searched for. As we will see below in the text this strategy, called “Word of the Day”, is also very common in social network profiles themselves.

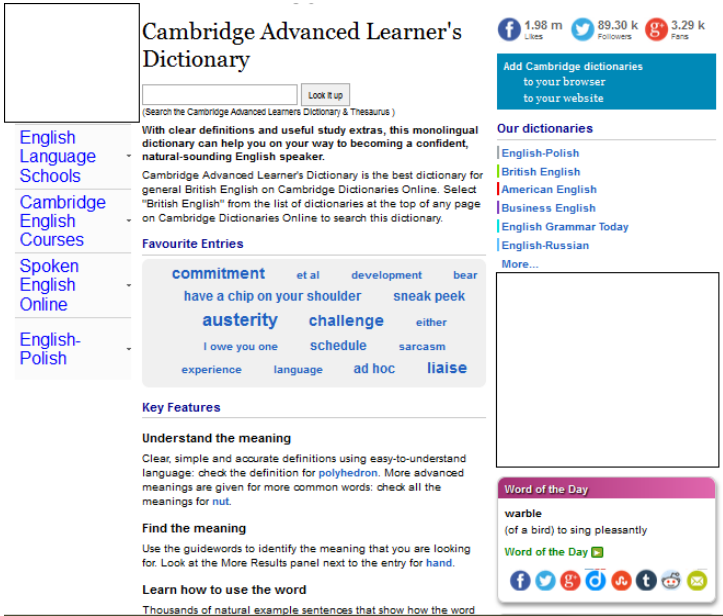


Figure 3: Alternative sharing content technique – Word of the Day, Cambridge Dictionary

Besides sharing content we can also follow (subscribe) to this dictionary social network profile which means that on our private social media account we will see the entries published regularly by someone from the dictionary team (lexicographer or marketing specialist). The frequency of the entries varies between different dictionaries. This problem will be discussed below as related to the example of Facebook profiles.

In the inventory I have discerned two main techniques of following dictionary profiles. One of them could be called a voluntary following. In this method we will have social media buttons somewhere on the page. If we click on them we will be led automatically to the dictionary profile. In this case we can see the content and subsequently, if we like the dictionary profile, we can subscribe to it by clicking a button dedicated to this purpose. Opposite to where the sharing buttons are located, following buttons can be located in many other places on the page. We see them at

the top of the page (see Figure 4).

The screenshot shows the top of the website with the logo and name 'Academia de la Llingua Asturiana'. Below the navigation bar, there are two search buttons: 'Busca cenciella' and 'Busca avanzada'. A search form contains the text 'Pallabra: llingua' and a 'Facer consulta' button. Below the form is a detailed dictionary entry for 'llingua, la' with various definitions and examples. At the bottom right, there is a 'Política de cookies +' button.

Figure 4: Following buttons at the top of the page (Asturian Dictionary)
They can also be located at the bottom of the page (see Figure 5):

The screenshot shows the Sproget website with a search bar at the top. Below the search bar, there are several content sections: 'Nyheder' (News), 'Grammatik for dummies', 'HÅNDENS TEMA', 'Han var en mørkglødet mand', 'VISTE DU?', 'SENI ET ORD', 'UGENS ORD', 'RÅ OG REGLER', 'TEMAER', 'LEG OG LER', and 'LINKS'. Each section contains text and images related to the Danish language. At the bottom, there are social media icons for Facebook, LinkedIn, and Twitter.

Figure 5: The following buttons at the bottom of the page, Sproget
Apart from simple buttons (an icon with a social network symbol) we can also encounter a special frame consisting of a following button with the total number of profile followers (individual subscriptions). It is one of many marketing tricks to show

that there are a certain number of people who subscribe to the profile so the profile itself is valuable and should be subscribed to.

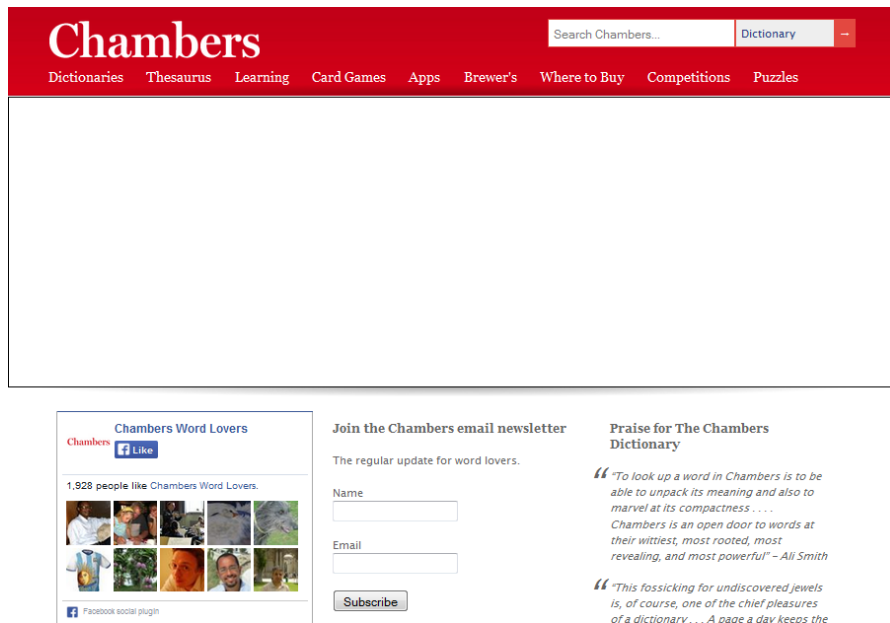


Figure 7: Facebook following button with the number of subscriptions and the selected photos of the followers (left side, bottom of the page), Chambers dictionary

In the inventory I have also found another, less used technique connected with following buttons. In this case we have a separate frame which enables us to see the latest entries from the social network profile and the total number of followers (see Figure 8). This way, users do not need to enter the profile to see its content. They can read and make a decision concerning the subscription without leaving the main or entry page.



Figure 8: Separate social media frames with latest updates (bottom of the page), Larousse Dictionary

Instead of a voluntary subscription (the user enters or recognizes a social media profile via the dictionary page and decides whether he wants to subscribe to it) a few dictionaries use involuntary following. In this case if the user clicks on the network profile button he would automatically subscribe to the content and become a follower. This technique is used for example by the Collins Dictionary (see Figure 9).

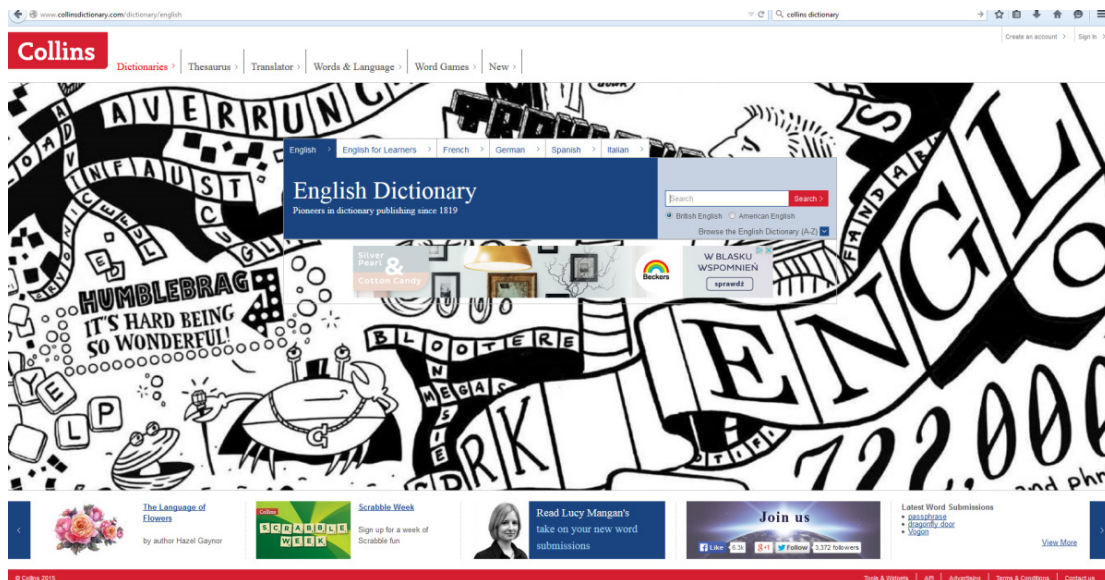


Figure 9: Involuntary following buttons example (“Join us” at the bottom of the page), Collins Dictionary

5. The content of dictionary social media profiles (Facebook example)

As aforementioned, the total number of dictionaries which use social network facilities (there is a link from the dictionary website) is 21. Among them, two do not have any visible button or frame which would lead us from the dictionary page itself to the separate social media profile (SLEX. Elektronický lexikón slovenského jazyka and ДИГИТАЛЕН РЕЧНИК НА МАКЕДОНСКИОТ ЈАЗИК). This means that these two dictionaries enable their users to share information about the dictionary on the user Facebook profile; however at the same time, the lexicographical team does not provide a separate social network profile. In such cases we can observe a simplified link between the dictionary website and the social network.

On the other hand, if the dictionary team decides to launch a social medium there is usually more than one social network involved. In most cases we can encounter Facebook and Twitter profiles, however Google+, YouTube, Flickr and Instagram are also quite popular. By multiplying profiles, the lexicographical team can achieve many goals. First of all every social network has its own characteristic (e.g. Twitter is used mainly to communicate via short messages, Flickr and Instagram are used for sharing elaborated photos and graphics). Therefore, each network might appeal to a slightly different group of users. Multiplying profiles also helps in the website

positioning process. Furthermore, the content of any particular social network (very specific) can be linked to in a simplified form on another social media profile. For example, the dictionary team creates an entry on the blog, and later informs users about this content on their Facebook profiles. This strategy leads the user to the impression that this particular dictionary is a very dynamic entity with a rich content and strong focus on the user`s needs.

Because of its immense impact and the various possibilities of sharing content on the profile, I have chosen the Facebook network as the subject for further analysis. What is very interesting is that not all Facebook profiles which are linked to the dictionary website lead users to the dictionary networks. Some of the dictionaries are also linked to the institutions that provide this lexicographical work for the Facebook profiles (Academia de la Llingua Asturiana profile in the case of Diccionariu de la Llingua Asturiana, Real Academia Española in the case of Diccionario de la lengua Española). Other sources link to the Facebook profile of a general product also consisting of particular dictionaries (Gran diccionari de la llengua catalana and Enciclopèdia Catalana profile) or the publishing house profile (Treccani publishing house related to Dizionário Treccani, Priberam company related to Dicionário Priberam da Língua Portuguesa, Van Dale related to Van Dale Dutch dictionary).

In these types of Facebook profiles, the kind of dictionary content varies. There are institutional profiles which do not reflect any kind of dictionary content (therefore in this case we have only one side called “blind” linking). To this group belongs, for example, the Treccani publishing house profile (there is no information about the dictionary itself, although we can read about various cultural facts, meetings with the authors and discover interesting quotations), Academia de la Llingua Asturiana profile (concerned mostly with events connected to the popularization of the Asturian language) and Real Academia Española (a profile focus on institutional events as well as the latest books published by RAE).

In the other non-dictionary profiles, lexicographic interests play a crucial or at least significant role. This method is used by the Priberam publishing house (we have “Word of the Day” content with a link to the dictionary entry, also guessing the subsequent day’s word) or Enciclopèdia Catalana (besides information about Catalan history and culture we can also acquaint ourselves with the various facts presented in the dictionary, e.g. grammar information, interesting phrasal verbs, correct word forms from Catalan; naturally each profile entry is linked to the dictionary page).



Figure 10: The example of linked dictionary content on the institution`s profile, Enciclopèdia Catalana Facebook profile

In the second analyzed group, dictionaries have their own Facebook profiles and this appears to be the dominant tendency. The first thing that should be discussed is the number of followers. In my inventory this measure varies from over 2 million (Cambridge Dictionary) to 2,000 followers of the Chambers English dictionary profile. The most important factor is probably the role of the language in international communication connected with the popularity of the lexicographic project (it is very visible in English dictionaries, e.g. Cambridge and Oxford Dictionaries vs. Chambers and Collins dictionaries). However, even if we are discussing the less used languages on the European scale (e.g. Danish, Romanian, Polish or Bulgarian), the number of followers always exceeds 2,000. This provides visible information regarding the popularity of interest in vocabulary among Facebook users.

The second measure concerning Facebook profiles is the frequency of entries. Relating to lexicographical projects, the keyword would probably be “irregularity”. Most profiles follow a particular pattern; for example, some are updated a few times each day (Oxford Dictionaries, Macmillan Dictionary; in such cases someone is definitely responsible for project promotion), others once a day (Romanian DexOnline), three or four times a week (Duden dictionary) or even more rarely (Wielki słowniki języka polskiego). However, every profile has moments when the gap between the entries becomes bigger. That gives valuable human resources information. Usually there is one person in the lexicographical project responsible for social networks. If this person is not present or is simply overwhelmed by other duties the social network is left without an update. Therefore, it is strongly recommended to delegate at least two people to work together or interchangeably to give a more professional impression.

When it comes to dictionary profiles, it must be mentioned that each and every lexicographical profile is a unique entity with a separate universe of its own (user

orientation, aesthetics, content, techniques for linking resources). However, there are also strategies which are quite common despite the diversity of the analyzed projects.

Probably the most popular and “lexicographical-like” technique in the European dictionary profiles is to publish the “Word of the Day”. This technique is used by the Oxford Dictionaries, Cambridge Dictionary, Macmillan Dictionary, Collins Dictionary, Priberam Dictionary and DexOnline.



Figure 11: Examples of the “Word of the Day” strategy, Priberam, Oxford Dictionaries, DexOnline

It is worth mentioning that in every entry representing this technique, we have a visible link to the dictionary website concerning this word. Also, the technique of creating attractive graphics or uploading visual illustrations seem to be very valuable, hence it encourages users to share the Facebook entry on their own private social accounts. The only aspect that causes concern, and not only in the case of the “Word of the Day” illustrations, is the lack of attribution. Sometimes photos and illustrations of a very high artistic quality are uploaded on dictionary profiles with no any caption. The person responsible for social networking should always consider the issue of royalties.

The second technique, typical not only for dictionary profiles but also for social network profiles in general, is to devote an entry to the subject that will be fully presented on a separate website (there is a link attached). This method has two variants. The first and more valuable comes from a marketing point of view (the number of users), and comprises mentioning our self-created content and resources. Depending on the project it could be a paper from the blog or from another part of our website (apart from the entries). This technique is widely used in English dictionaries (see Figure 12). Once launched, it could probably encourage lexicographers to write short essays commenting on dictionary resources.



Figure 12: Examples of linking different dictionary resources in a social network profile, Oxford Dictionary, Macmillan Dictionary, Collins Dictionary

As well as linking to our self-created dictionary resources, social network profiles also consist of many outside links connected to broadly defined linguistics. The type and intellectual level of the linked content depends on the person editing the profile. Probably the most entertaining technique is to gather different photos illustrating actual language errors (see Figure 13).

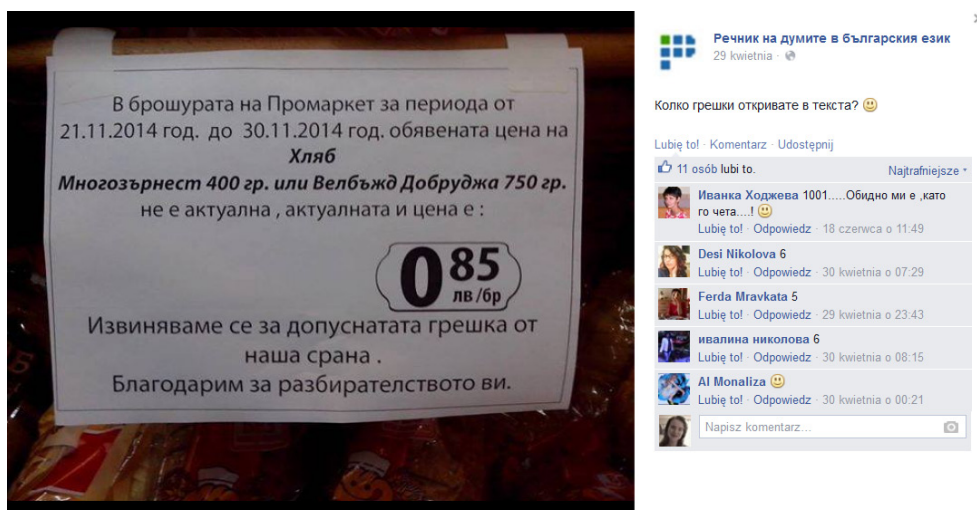


Figure 13: Example of an entertaining entry with the dictionary photo of the supermarket shelf sourced from elsewhere (*How many errors can you make in one text?*), Rechnik dictionary

One of the lexicographical goals, even in the case of monolingual and descriptive dictionaries, is to present correct and appropriate language usage: in the inventory I have found many interesting examples of entries devoted to this subject. There, such phenomena as idioms, paronyms, rare words and their meanings, common grammar and spelling mistakes, etc. were discussed. In the case of regional languages (Catalan) correct regional word forms were mentioned.

Among the techniques used, there were two which strongly encourage interaction. One is to give a short linguistic test, usually of only one question (see Figure 14). This technique is used for example in the Chambers, Van Dale Dictionary, Priberam and Duden Dictionaries. In the case of commercial and published dictionaries there could also be the possibility of winning a book.



Figure 14: Examples of linguistic quizzes on these dictionary profiles, Sproget, Duden Dictionary

The second highly interactive technique is to ask the users for help; for example in the case of rare meanings which are not well illustrated in the dictionary corpus (professional usages, meanings connected with strongly spoken jargons). This method is used in the Oxford Dictionaries and can be found to be very fruitful in the case of problematic lemmas which are corpus resistant (see Figure 15).



Figure 15: An example of asking for lexicographical help technique, Oxford Dictionaries

6. Advantages and desiderata

As was shown in the above examples, various techniques connected with social networks are being used in European monolingual lexicography. All have one goal: to increase the number of active users of the dictionaries. Aside from this, profiles can fulfill other important functions. It is a topic for further discussion whether social media profiles should educate or rather entertain. Is marketing our only goal or do we also feel obliged to share our knowledge with users? This question is also raised when considering the intellectual level of our entries. Is it ethical to laugh at somebody's lack of education by posting photos of wrongly written words or phrases? If we focus only on education will we, by doing so, deprive ourselves of the users who are focused purely on internet entertainment?

Concerning the lexicographer, profiles could enable us to think differently about dictionary resources and the needs of our users. The role of social network profiles in contemporary electronic lexicography seems to be irreplaceable; hence they offer a unique opportunity to connect and link various lexicographical data.

While administrating social network profiles is only a small part of our lexicographical work, it could also be useful to create an inventory of dictionaries using networking techniques. This way we could share our experiences, influence and inspire each other.

As for the future, it could also be interesting to repeat the analysis of social network profiles for bilingual dictionaries and other types of monolingual dictionaries (e.g. historical, etymological dictionaries or dictionaries of discontinuous units like textual units or idioms). While this paper focused mainly on matters important to the lexicographer acting as the social network administrator, it would be useful also to analyze feedback from Facebook or Twitter users. This would bring us closer to the relatively complete picture of the user-oriented contemporary lexicography.

6. Acknowledgements

Praca naukowa finansowana w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Narodowy Program Rozwoju Humanistyki” w latach 2013-2018, nr projektu: 0016/NPRH2/H11/81/2013.

Scientific work financed under the program of the Minister of Science and Higher Education under the name "National Program for the Development of Humanities" in the years 2013-2018, Project No.: 0016/NPRH2/H11/81/2013.

7. References

Websites

- Cambridge Advanced Learner`s Dictionary. Accessed at: <http://dictionary.cambridge.org>. (21.05.2015)
- Chambers Free English Dictionary. Accessed at: <http://www.chambers.co.uk>. (21.05.2015)
- Collins English Dictionary. Accessed at: <http://www.collinsdictionary.com> (21.05.2015)
- Den danske Ordbog. Moderne Dansk Sprog. Accessed at: <http://www.ordnet.dk/ddo> (21.05.2015)
- Dex Online. Dicționare ale limbii române. Accessed at: <http://www.dexonline.ro>. (21.05.2015)
- Diccionario de la lengua Española, Diccionario esencial de la lengua Española. Accessed at: <http://www.rae.es/obras-academicas/diccionarios>. (21.05.2015)
- Diccionariu de la Llingua Asturiana. Accessed at: <http://www.academiadelalingua.com/diccionariu/> (21.05.2015)
- Dicionário Priberam da Língua Portuguesa. Accessed at: <http://www.priberam.pt/dlpo>. (21.05.2015)
- Dizionario Treccani. Accessed at: <http://www.treccani.it>. (21.05.2015)
- ДИГИТАЛЕН РЕЧНИК НА МАКЕДОНСКИОТ ЈАЗИК. Accessed at: <http://www.makedonski.info>. (21.05.2015)

Duden Wörterbuch. Accessed at: <http://www.duden.de>. (21.05.2015)
Грамота.ру. Accessed at: gramota.ru. (21.05.2015)
Gran diccionari de la llengua catalana. Accessed at: <http://www.diccionari.cat>.
(21.05.2015)
Larousse Dictionnaire de Française. Accessed at:
<http://www.larousse.fr/dictionnaires>. (21.05.2015)
Lexilogos. Accessed at: <http://www.lexilogos.com>. (21.05.2015)
Macmillan Dictionary. Accessed at: <http://www.macmillandictionary.com>.
(21.05.2015)
Oxford English Dictionary. Accessed at: <http://www.oxforddictionaries.com>.
(21.05.2015)
Речник на думите в българския език. Accessed at: <http://www.rechnik.info>.
(21.05.2015)
SLex. Elektronický lexikón slovenského jazyka. Accessed at: <http://www.slex.sk>.
(21.05.2015)
Sproget (Den Danske Ordbog). Accessed at: <http://www.sproget.dk>. (21.05.2015)
Van Dale. Accessed at: vandale.nl. (21.05.2015)
Wielki słownik języka polskiego. Accessed at: <http://www.wsjp.pl>. (21.05.2015)

Facebook Profiles

Cambridge Dictionaries Online. Accessed at:
<http://www.facebook.com/CambridgeDictionariesOnline>. (21.05.2015)
Chambers Word Lovers. Accessed at: <http://www.facebook.com/wordlovers>.
(21.05.2015)
CollinsDictionary.com. Accessed at: <http://www.facebook.com/collinsdictionary>.
(21.05.2015)
sproget.dk. Accessed at: <http://www.facebook.com/sprogetdk>. (21.05.2015)
dexonline. Accessed at: <http://www.facebook.com/dexonline>. (21.05.2015)
Real Academia Española. Accessed at: <http://www.facebook.com/RAE>. (21.05.2015)
Academia de la Llingua Asturiana. Accessed at: <http://www.facebook.com/AcademiadelaLlinguaAsturiana>. (21.05.2015)
Priberam. Accessed at: <http://www.facebook.com/priberam>. (21.05.2015)
Treccani.it. Accessed at: <http://www.facebook.com/treccani>. (21.05.2015)
Duden. Accessed at: <http://www.facebook.com/Duden>. (21.05.2015)
GRAMOTA.RU. Accessed at: <http://www.facebook.com/gramota.ru>. (21.05.2015)
Enciclopèdia.cat. Accessed at: <http://www.facebook.com/Enciclopedia.cat>.
(21.05.2015)
LAROUSSE. Accessed at: <http://www.facebook.com/larousse.fr>. (21.05.2015)
MacDictionary. Accessed at: <http://www.facebook.com/pages/MacDictionary>.
(21.05.2015)
Oxford Dictionaries. Accessed at: <http://www.facebook.com/OxfordDictionaries>
(21.05.2015)
Речник на думите в българския език. Accessed at:
<http://www.facebook.com/rechnikinfo>. (21.05.2015)

Van Dale Uitgevers. Accessed at: <http://www.facebook.com/VanDaleUitgevers>.
(21.05.2015)

Wielki słownik języka polskiego. Accessed at: <http://www.facebook.com/wsjpgpan>.
(21.05.2015)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



The Construction of Online Health TermFinder and its English–Chinese Bilingualization

Jun Ding¹, Pam Peters², Adam Smith²

¹ Fudan University, No 220 Handan Rd. Shanghai, China.

² Macquarie University, NSW 2109, Australia.

E-mail: jdming@fudan.edu.cn, pam.peters@mq.edu.au, adam.smith@mq.edu.au

Abstract

Health TermFinder (HTF) is an online platform and information tool designed to support medical and health terminologies. Pilot termbanks in selected fields such as breast cancer are currently under construction at Macquarie University in Sydney. Cooperation with Fudan University in Shanghai is underway to develop a bilingualized English–Chinese version of HTF. This paper provides a theoretical overview of HTF as a customized electronic information tool, with reflections on its structure, data organization, user interface and overall principles of construction. Following a discussion of the macrostructure of HTF, i.e., whether it is essentially a lexicographic or terminological work, two sections of the paper are devoted to discussions of its corpus-based selection of headwords and design of the microstructure, with emphasis on the user-oriented philosophy underlying both and based on best principles/practice in lexicography and multimodal language learning. The status quo of the cooperative bilingualization project is given close examination in Section 5, and in Section 6 the possible use of adaptive hypermedia in its future development is proposed.

Keywords: Online dictionary; Health TermFinder; user-oriented; bilingualized; adaptive hypermedia

1. Introduction

The difficulties and problems arising from the use of medical terminology cannot be overestimated in either medical research or in practice. The high linguistic demands of the language found in online health information, which could cause problems for those with low levels of literacy in English, motivated researchers at Macquarie University, Sydney¹, to construct a public online information tool for medical terminologies, codenamed Health TermFinder (HTF). Its target users include second-language health professionals in Australia and native English speakers without tertiary education. The Macquarie team is currently working on the first of the HTF termbanks consisting of breast cancer terminology, which currently comprises 51 pages.

¹ This team includes the two coauthors for this paper, Pam Peters, director of the TermFinder project and Adam Smith, researcher. Others are lexicographer Yusmin Funk, and Professor John Boyages of the Macquarie University Cancer Institute, who reviews the termbank's medical content for accuracy.

Meanwhile, the cooperative project of bilingualizing HTF into Chinese at Fudan University, Shanghai, is under negotiation with a team of English–Chinese bilingual lexicographers. The bilingualized Online Health TermFinder (BHTF) is expected to meet the needs of medical students at the Medical School of Fudan University (both undergraduates and graduate students) at its initial stage of development. Once in its later and more full-fledged form, BHTF will be made accessible to the whole Mandarin-speaking community in China.

So what is the nature of this Online Health TermFinder? Is it essentially a lexicographical or a terminological work? If, as described above, the project seems to have begun with observations on specific needs of specific sets of users, upon which principles is its design based; what are its macro- and micro-structures? And what makes the English–Chinese bilingualized version special in comparison to the plain translations into other Australian community languages (including Chinese) offered on the HTF platform? These are the questions to be addressed in this paper which attempts to examine not only the design and input data, but also the construction philosophy of HTF.

2. Lexicographic or Terminological?

In a broad sense, HTF is designed to be an online dictionary-type tool, providing help with health-related and medical terms in English. Yet initially it follows the so-called onomasiological model: a certain health issue is selected as the subject field for the new termbank. For instance, HTF currently includes only one such specialized area, the breast cancer termbank. However, the contents of the termbank do not represent a structured vocabulary of terms used in the field, nor are they restricted to concepts related only to breast cancer. HTF termbanks deal with not only medical terminologies, but also semi-technical terms. This is because their target users are people with low literacy levels in English, including both second-language health professionals and native-English-speaking patients and carers without tertiary education. Since semi-technical terms are usually inherently polysemous, they are likely to pose difficulties to the target users. Terms such as *treatment* will be searchable from one termbank to another, as many are generic medical terms useful to people with different medical problems. Therefore, despite its essentially onomasiological structure (consisting of distinct medical fields), HTF could hardly be considered a strictly terminological project (Riggs, 1989: 89) in view of the mixed lexical content of individual termbanks. Moreover, HTF is designed to serve decoding, or interpretive, purposes at the functional level; another reason to categorize it as essentially a lexicographic rather than terminological work, since the latter is usually also defined by its aim “to help writers produce texts” (Riggs, 1989: 90).

Though lexicographic by nature, HTF also differs considerably from a medical dictionary. For one thing, it lacks the scale or all-inclusiveness of a standard print dictionary. Unlike many online specialized dictionaries, it does not have a printed

counterpart. In other words, it is not adapted from a medical dictionary already in existence. The entry terms included in the breast cancer termbank are instead extracted from a database of online documents on breast cancer care, built by the team at Macquarie University. This practice of building reference databases from scratch will be replicated for other fields of healthcare. Based on such databases, HTF will eventually develop into a huge online multidisciplinary clearing-house in healthcare, rather than a conventional medical dictionary.

This also means that each individual termbank will have a claim to independence, and thus can be made available to users as a stand-alone termbank. In other words, it is not necessary to wait until the whole project is completed before launching it for public use, unlike the case of most dictionaries which have to be finished from A to Z before going into print or online. The HTF project ought thus to be looked upon as a process rather than a product. Its construction would simply go on until all the important health and medical areas are dealt with, and after that it could still be maintained in a continuously updatable form. Since users' needs are not static, but change and develop throughout time, the updatable form of HTF makes it a lexicographic work which can be constantly adapted and modified to meet the new or evolving needs of its users.

3. User-oriented data

In his discussion of lexicography for the language learner, Tarp (2008) elaborated on the importance of knowing the user profile, user situation, and user needs when creating an online dictionary tool. HTF is exactly such a lexicographic work, designed with a clear extra-lexicographic identification of its specific set of users and their specific needs.

The problems caused by medical terminology are a constant challenge for those health professionals in Australia who speak English as a second language. Native speakers with low levels of literacy encounter similar difficulties in understanding the “jargon” of medicine when either communicating with their doctors or reading printed factsheets or medical websites to access more in-depth information. Researchers at Macquarie University were thus motivated to construct an online information tool for medical terms so as to provide post-consultation help to patients and carers, as well as linguistic support to second-language health professionals.

A large body of online documents on breast cancer were collected from one of Macquarie University Library's specialized online LibGuides². They were categorized into two types in view of different readerships: those designed for the general public and those for health professionals. The documents were accordingly extracted into two separate databases: public (521,232 words) and professional (514,830 words, as of December 2014). Contents of the documents and their respective target audience are

² <http://libguides.mq.edu.au/content.php?pid=379776&sid=3605261>

listed in Appendix 1. Data analyses were carried out by the Macquarie research team to extract word frequencies and other lexical statistics (Peters et al., 2015)³. A preliminary table listing the top 24 words and terms in the professional and public databases are presented in Appendix 2. The very high levels of medical and semi-technical health management terms (*clinical, biopsy, carcinoma, screening*) in the professional listing show the demands on second-language professionals, let alone lay readers (patients and their carers) with low literacy levels in English.

All this preliminary terminological research demonstrates the user-oriented philosophy for HTF. The two databases are also being used as a corpus for the compilation of the breast cancer termbank; namely, for identifying terms, prioritizing them for attention, and providing examples of their usage in technical texts. Since the medical documents in the corpus are up-to-date and have specific readerships (breast cancer professionals and patients), the data extracted from them are highly user-oriented and consequently ensure the uptodateness and usefulness of the definitions and examples for the headwords entered in the termbank.

4. User-oriented Microstructure

The microstructure of each head entry is based on best practice for learners' dictionaries as well as multimodal language learning (Lemke, 1998). The users' actual needs are not easily ascertained through questionnaires or interviews, since "users may frequently only have a vague or approximate idea of the objective needs" (Tarp, 2009: 281). On the other hand, profiling the vocabularies of professional vs. public online documents on breast cancer is a practical and productive way of discovering the "genuine or objective needs occurring before the consultation process, i.e., extra-lexicographically" (Tarp, 2009: 282). That the needs thus identified are hypothetical does not make them invalid; though of course the validity still needs to be assessed by users once they start using the HTF termbanks.

Because HTF is a nonprofit research project, freely available to the public, no consideration would be given to the artificial needs of potential users, which are defined by Tarp (2009: 282) as publicity-created subjective needs mainly of interest to commercial publishers. Instead, the content and arrangement of information on each HTF page is designed to meet the genuine, objective needs of the target users. Below is a screenshot of the breast cancer termbank page for the word "lymphoedema", showing the essential English content.

³ This article is titled "Language, Terminology and the Readability of Online Cancer Information".

The screenshot shows the HealthTermFinder interface. On the left is a search sidebar with a 'Search' box containing 'lymphoedema', a 'Go' button, and a 'Translation Options' dropdown menu. The main content area features the term 'lymphoedema' in a large font, followed by '(Breast Cancer)', 'English pronunciation: [audio icon]', 'Grammar: noun', 'Definition: swelling of a limb due to the build-up of lymph', and 'Example: 1. Lymphoedema of the arm can occur after axillary treatment of any sort: dissection, radiation, or even after a sentinel node biopsy. 2. Early symptoms of lymphedema include heaviness, aching, fluctuating swelling in the hands or fingers; and later, swelling of the forearm, upper arm or the whole arm.' An image of a person's arm with lymphoedema is shown on the right, captioned 'Lymphoedema of the arm'. Below the examples is the 'Alternative Form: lymphedema'.

Figure 1: Screenshot of the page for “lymphoedema”

As we can see from the Figure, each term page in HTF includes five elements of lexicographic information:

- 1) lemma: *lymphoedema*
- 2) grammatical label: *noun*
- 3) definition: *swelling of a limb due to the build-up of lymph*
- 4) examples: *1. Lymphoedema of the arm can occur after axillary treatment of any sort: dissection, radiation, or even after a sentinel node biopsy. 2. Early symptoms of lymphedema include heaviness, aching, fluctuating swelling in the hands or fingers; and later, swelling of the forearm, upper arm or the whole arm*
- 5) alternative form: *lymphedema*

One of the foremost features of HTF is that the definitions are drafted in plain, highly-accessible English, accordant with the needs of second-language users and those with low reading skills. The definitions are induced from actual instances of the term’s use in the corpus, to cover both intensive and functional aspects of its meaning as far as possible.

Also noteworthy is the fact that neither captions nor labels on HTF are given in abbreviations or initials. It is common practice in dictionary making to avoid using captions (such as “Definition”, “Examples”) and to present grammatical labels as briefly as possible (“n” for “noun”, for instance), so as to save precious space for more indispensable information. Space is no longer a problem with online information tools and given that the main concern of HTF is to make the look-up process as easy and

friendly as possible for its target users, we have retained captions and labels spelled out in full to serve as important signposts.

For each entry term two examples of usage are selected from the corpus to complement the definition and provide users with both linguistic and factual information for the term in question. Illustrative materials are also sought in the corpus to show the term's place among other related terms, usually arranged in labeled diagrams or tables of parallel terms. Diagrams, tables, and pictures of relevant images, such as the above one showing “lymphoedema of the arm”, are introduced based on Lemke's (1998) theories about meaning-making via various semiotic “channels”. Lemke purports that information passed through different channels, such as linguistic, visual, pictorial and acoustic, can be equivalent or complementary, and may or may not reach the person simultaneously. Multimedia facilities make it possible to incorporate multiple semiotic systems on HTF. Besides the visual presentations of graphs, tables, pictures, etc., audio files providing the pronunciation and definition of the term are also available on each term page.

On the left-hand part of the page one can select the relevant termbank (breast cancer, for instance), and can then search terms. Below the look-up box, translations are offered in four of the major community languages in Australia, namely Arabic, Spanish, Vietnamese and Chinese (both traditional and simplified), which, when selected, raise translation boxes for the head term and its definition, as well as for the captions on graphics and labels on diagrams. The translated elements are expected to provide the second-language users with a more efficient “channel” for accessing the relevant information and anchoring their understanding. All the primary contents of HTF (definitions, examples, images, tables) are reviewed by medical experts, and “checkers” are appointed to review the primary translations for each language.

5. English–Chinese Bilingualization

Though translations of the primary contents are available in four languages for selected elements on HTF pages, the system will be fully bilingualized into Chinese (simplified) only in the second stage of the project, which will be carried out at the English department of Fudan University, Shanghai. Again, the English–Chinese bilingualization project is based on a clear identification of target users and their needs. The plan is to make the bilingualized Health TermFinder (BHTF) first accessible to Chinese students of the Medical School at Fudan for training purposes. With its medical terms related to different specific diseases, BHTF can be used as a specialized reference tool alongside more general English–Chinese medical dictionaries.

Ever since Benjamin Hobson (1816–1873) published *A Medical Vocabulary in English and Chinese* (1858), the earliest English–Chinese medical glossary of its kind known in China, the translation of medical terminologies from English into Chinese has played a role of pivotal importance in the development of medical science in the country (Wu &

Wong, 1932; Chen, 1984; Fu, 1990; Ma et al., 1993; Sun, 2010). One hundred and fifty years later, most important medical terms now have Chinese equivalents well established in the language (for instance, 乳腺癌[ruxian' ai] for *breast cancer*, 淋巴[linba] for *lymph*, 血管[xueguan] for *blood vessel*, etc.). For obvious reasons, medical and health professionals in China have to learn English and conduct their research and practice medicine using English as a second language. As a result, medical dictionaries are always in great demand and the best ones are often based on authoritative monolingual English medical dictionaries. For instance, *An English–Chinese Medical Dictionary* (ECMD, editor-in-chief Weiyi Chen, 1984, 1997, 2009, 2014) was largely a translation work of *Dorland's Illustrated Medical Dictionary* (Li & Chen, 2006). These bilingual medical dictionaries are also being increasingly converted into digital forms. The 3rd edition of ECMD was developed into a mobile phone application and is available for free downloading. Yet, the majority of these dictionaries offer only Chinese equivalents. No definitions for the head entries either in English or Chinese are included.

BHTF aims to provide Chinese medical students with more complete information about medical terms, including all the English texts for the head entry, its Chinese equivalent(s), Chinese translations of the definitions and examples and also of all English terms and texts in the diagram/table/illustration and usage notes. Although the Chinese equivalents of the English definitions would be sufficient for Chinese medical students to understand the looked-up term, their constant need to improve their level of English would drive them to read the English texts. In fact, BHTF can also serve as an alternative language learning tool for such users. Meanwhile, it is also necessary to provide Chinese translations of the definitions because of the students' limited English proficiency. Moreover, the independently-drafted definitions could also deviate from the orthodox ones (i.e. those found in traditional medical dictionaries) because the rapid development in medical science may impose some newly acquired, context-specific meanings on established terms. Definitions in Chinese could thus more efficiently alert the users to these differences. It is common practice for bilingual dictionaries in China to present translations of all illustrative examples, and would therefore be expected by Chinese-speaking users and would aid their comprehension of the head term and their English learning in general.

At a later stage, BHTF is to be made available to the general Chinese-speaking public, who often need help with medical terms after consulting a health professional. This stage will occur after the Australian HTF is made bidirectional, i.e., equipped with a redeveloped version of the present platform as a Chinese–English structure. Medical terms in both English and Chinese will then be searchable on the platform, each navigating users to the same entry page for information. Users from the Chinese public, with a limited command of English, are likely to benefit most from the Chinese translations for the looked-up term. However, if able to look up Chinese terms on BHTF, this online information tool will be doubly useful, providing a tool for Chinese citizens as well as for medical students. Also under consideration is the Romanization

of the Chinese equivalents, i.e., the inclusion of their Pinyin forms. This is because an increasing number of foreign students are coming to study at Fudan University each year. These non-Chinese-speaking students may need to communicate about medical issues in Chinese, and the Romanization of Chinese characters would considerably facilitate their pronunciation (the pronunciation of the Chinese characters cannot be inferred from their form). Audio files of Chinese equivalents and definitions can also be provided for their benefit as well as for that of Chinese citizens who speak a regional dialect.

6. Adaptive Hypermedia – Future Development for BHTF

The Fudan team working on the BHTF are considering the application of hypermedia to the termbank as part of its future development. *Hypermedia* in this context refers to user-adaptive software systems which can select and prioritize items of information for users depending on their individual needs (Brusilovsky & Millan, 2007). Adaptive hypermedia has been applied to an English dictionary of finance for Indonesian students (Kwary, 2011), in which the adaptive search system directs the user's search action to the results decided by the system to be preferable or most suited to the user's needs. It requires lexicographers to decide upon the most suitable result when searching for a particular term, and to set it up accordingly in the dictionary system. However, these decisions must be based on the user profile.

Since we have a very specific group of target users for the first stage of BHTF – Chinese medical students at Fudan University – it would be relatively easy to build a comprehensive user profile. Then, for example, we would be able to decide if, for a certain medical term, its Chinese translations would be more helpful to the Chinese student than would the English definition, or vice versa. As medical students, these users are expected to be equipped with a greater knowledge of terms than non-professionals, so that when they look up a certain English medical term, it is likely to be because they want to read its definitions in English and to see examples illustrating its actual usages. In other words, medical students are very likely to use BHTF for productive purposes as well as receptive ones, though the original English version is designed for meeting decoding needs. If a semi-technical term is looked up, it may suggest that the student's level of scientific English is below average, and therefore it is best to direct the user immediately to the Chinese translations which can solve their decoding problems more efficiently.

This kind of hypothesizing is based on what Tarp calls “function-related needs”; needs identified as objective and in an extra-lexicographic situation, and differing from “usage-related needs” which occur only during the actual consultation process (2009: 283). For instance, it may happen that when a certain esoteric medical terminology is looked up by a user and they are offered its English definition and examples, they always move straight on to its Chinese translations. This could imply that the term in question is completely new to most medical students who look it up on BHTF, and

that its English definitions may not be clear enough for them after all. It is equally possible that a semi-technical term is looked up more often for its English than its Chinese parts, which could mean that this term is familiar to most users and more likely to pose difficulty in encoding rather than decoding tasks. In that case, what is required is another adaptive system called “log file action analysis” (Kwary, 2011: 37) which saves the users’ search actions for different terms. The initial setup for that particular term may be automatically changed after a number of such actions being recorded.

The second stage of BHTF would present a more complex scenario, when is the resource is made accessible to the general community, composed of mostly Chinese citizens with on average a very limited command of English. Yet, over the years, scholars interested in future dictionaries have discussed and predicted the possibility of individualization or customization of dictionaries (Dodd, 1989; Atkins, 1996; Whitelock & Edmonds, 2000; de Schryver, 2003; Lew & de Schryver, 2014); each discussion more daring and confident than the previous. Indeed, given the speed and scale of technological development, one has every reason to feel confident about the advent of ever more advanced adaptive hypermedia software in the near future which could save and process search tasks performed by individual IP addresses and adapt the system to any particular user’s needs.

7. Conclusion

Online Health TermFinder, which is currently under construction at Macquarie University, is a completely user-oriented, nonprofit, digitalized lexicographical project aimed at providing linguistic and factual information on medical terms with open access on the internet. HTF targets users are either health professionals speaking English as a second language or native-English speakers with low literacy levels, and serves predominantly decoding purposes. Its bilingualized version, BHTF, to be constructed at Fudan University, will primarily target Chinese medical students with both decoding and encoding needs.

With the current development focused on the field of breast cancer, directions for expansion are already under consideration within the Macquarie research team. Other types of cancer and major medical and health problems such as orthopedics and mental health also call for online information from well-designed termbanks. Meanwhile, after consulting the Medical Faculty, the Fudan BHTF team has nominated priority areas to align with the structure of medical training, which include the field of various cancers, respiratory diseases, and paediatrics. Once a new area for development has been identified, online materials in English in the relevant field will be sought and collected to build databases and thus a new round of TermFinder construction will begin.

8. References

- Atkins, B. T. S. (1996). Bilingual Dictionaries: Past, Present and Future. In M. Gellerstam et al. (eds.) *Euralex '96 Proceedings I-II, Papers submitted to the Seventh EURALEX International Congress on Lexicography in Goteborg, Sweden*. Gothenburg: Department of Swedish, Goteborg University, 515-546.
- Brusilovsky, P. & Millan, E. (2007). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In P. Brusilovsky, A. Kobsa & W. Nejdl (eds.) *The Adaptive Web*. Berlin: Springer-Verlag, 3-53.
- Chen, Bangxian. (1984). *History of Chinese Medicine*. Shanghai: Shanghai Bookstore.
- Dodd, W. S. (1989). Lexicomputing and the Dictionary of the Future. In G. James (ed.), *Lexicographers and Their Works. (Exeter Linguistic Studies 14.)* Exeter: Exeter University Press, 83-89.
- De Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic Dictionary Age. *International Journal of Lexicography*, 16(2), pp. 143-199.
- Fu, Weikang. (1990). *History of Chinese Medicine*. Shanghai: Shanghai Chinese Medicine University Press.
- Kwary, D. A. (2011). Adaptive Hypermedia and User-Oriented Data for Online Dictionaries: A Case Study on an English Dictionary of Finance for Indonesian Students. *International Journal of Lexicography*, 25 (1), pp. 30-49.
- Lemke, J. (1998). Multiplying Meaning: Visual and Verbal Semiotics. In J. Martin & R. Veel (eds.) *Reading Science*. London: Routledge. Available at:
<http://academic.brooklyn.cuny.edu/education/jlemke/papers/mxm-syd.htm>
- Lew, R. & de Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography*, 27(4), pp. 341-359.
- Li, D. & Chen, W. (2006). The English-Chinese Translation of Medical Terms. *Chinese Translators Journal*, 6, pp. 60-64.
- Ma, Boying, Gao, Xi & Hong, Zhongli. (1993). *History of Medical Culture Communication between China and the World*. Shanghai: Wenhui Publish House.
- Riggs, F. W. (1989). Terminology and Lexicography: Their Complementarity. *International Journal of Lexicography*, 2 (2), pp. 89-110.
- Sun, Zhuo. (2010). The Creation of Modern Medical Terms: A Case Study of Benjamin Hobson and His A Medical Vocabulary in English and Chinese. *Studies of Natural Science History*. vol. 4. Available at:
http://www.cssn.cn/st/st_xsplc/201404/t20140410_1063151.shtml
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Niemeyer: Tübingen.
- Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexicos*, 19, pp. 275-296.
- Whitelock, P. & Edmonds, P. (2000). The Sharp Intelligent Dictionary. In U. Heid et al. (eds.), *Proceedings of the Euralex International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th, 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, pp. 871-876.

Wu, L. T., & Wong, C. (1932). *History of Chinese Medicine*. Tientsin: Tientsin Press. Ltd.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Appendix 1: Contents of the databases

<i>document</i>	<i>words</i>	<i>target audience</i>
Cancer Council NSW	117738	public
Cancer Council Australia	10265	public
National Cancer Prevention and Early Detection Policy	11270	public
Cancer Council Victoria brochures	37993	public
Cancer Australia website	87993	public
BC in men	6556	(men) public
Clinical best practice and info for health professionals	497760	professional
Cancer Australia pamphlets	66830	public
Breast cancer risk factors: a review of the evidence 2009	38899	professional
All BCI pamphlets in word doc	58308	public
Breast Cancer network Australia website	166486	public
Information for health professionals	876	professional
BCNA pamphlets	114102	public
National Breast Cancer Foundation_part of website	3309	public
ABC Health & Wellbeing - Breast Cancer	22494	public
Pink Hope	16863	public
pink hope pamphlets	16630	public
Life After early Breast Cancer	20606	public
Breast Cancer and Axillary Lymph Nodes	644	public
BRCA Genes and Breast Cancer	622	public
TOTAL	1296244	

Appendix 2: Comparative rankings of top 24 words and terms in the two databases

Professional			Public		
		514830 wds			521232 wds
rank	term	frequency	rank	term	frequency
1	breast	11204	1	cancer	10730
2	cancer	10403	2	breast	9565
3	women	4724	3	women	5167
4	risk	2437	4	treatment	2932
5	clinical	2424	5	information	2011
6	treatment	2017	6	risk	1863
7	patients	1687	7	help	1550
8	study	1492	8	care	1447
9	evidence	1387	9	health	1447
10	practice	1286	10	surgery	1275
11	management	1269	11	people	1172
12	information	1227	12	reconstruction	1108
13	biopsy	1188	13	support	1101
14	guidelines	1138	14	time	1083
15	imaging	1137	15	research	1034
16	national	1116	16	pain	1012
17	Australia	1111	17	family	975
18	carcinoma	1067	18	chemotherapy	963
19	diagnosis	1050	19	find	931
20	care	1027	20	Australia	917
21	early	1007	21	feel	914
22	studies	991	22	side	811
23	health	937	23	effects	800
24	screening	924	24	doctor	790

Towards the enrichment of terminological resources by scientific corpora analysis

Izabella Thomas, Iana Atanassova

Research Centre in Linguistics and in Natural Language Processing Lucien Tesnière,
University of Franche-Comté, Besançon 25030, France
E-mail: izabella.thomas@univ-fcomte.fr, iana.atanassova@univ-fcomte.fr

Abstract

The research presented in this paper explores the possibility of enriching terminological databases through the analysis of recent scientific publications. Our main concern is to evaluate how useful automatic term extraction can be to a human expert. To carry out our experiment, we constructed two corpora of recent scientific papers in two different sub-domains of the bio-medical sciences. Then we proceeded with three steps: automatic term extraction and ranking from a set of corpora of scientific papers; evaluation of the overlap of the candidate terms (CTs) extracted from the corpora and those present in the multidisciplinary terminology portal TermSciences; and evaluation by domain experts of the three sets of the top 200 CTs extracted from the different corpora. To extract terms we used the Sensunique Platform, a web based platform for building terminological resources. Our results show that only about 10% of the extracted CTs are present in the TermSciences resource, which means that many of the extracted CTs, if validated, could potentially be used to enrich the terminological database. Furthermore, the expert evaluation of the top 200 terms for each sub-corpus shows clearly that about 75% of these CTs are correct terms in the respective domains. This validates our ranking algorithm.

Keywords: terminology; term acquisition; term extraction; term recognition; scientific papers

1. Introduction

The research presented in this paper aims to explore the possibility of enriching terminological databases through the analysis of recent scientific publications. The analysis is intended to be representative of a typical situation of a terminologist at work; therefore, it is constrained by the size of the corpora and the number of candidate terms (CTs) to be managed by an analyst. One can imagine two applicative scenarios: enriching an existing resource or building a new terminological resource from scratch, as can be the case for some institutions. Our main concern is to evaluate the usefulness of automatic term extraction for human experts, i.e. the relevance of automatically constructed lists of CTs compared to the given terminological resource. More precisely, we investigate the improvement of the strategy of filtering of CTs proposed by automatic term extractors in order to organize better the work of domain experts by ordering the list of CTs according to their termhood probability.

An interest in automatic term acquisition from corpora has been developing since the

1990s (Jacquemin & Bourigault, 2003). The task consists of the automatic recognition and extraction of terminological units from different domain-specific text collections. Resulting CTs can be used in more complex applications such as Information Extraction and Retrieval, ontology construction, document indexing etc. Building and enriching domain-specific vocabularies by the analysis of corpora constitutes one of the major applications in this domain. Its objective is to help domain experts find the best term candidates from corpora, taking into consideration the type of resource to be constructed (Bourigault & Jacquemin, 2000; Bourigault et al., 2004). Since the 1990s, numerous automatic tools, mostly term extractors, have been developed based essentially on two types of approaches: statistic or linguistic, or a hybrid of these two methods¹. Some of these tools have been developed, or can be used, for the French language, for example ANA (Enguehard & Pantera, 1995), Acabit (Daille, 1995), Lexter (Bourigault, 1993), TermoStat (Drouin, 2002), YaTeA (Aubin et al., 2006). The term extractors are considered mature technology nowadays (Cerbah et al., 2006), but this affirmation depends on the objective of the terminology acquisition: Information Retrieval or terminological mono- or multilingual resource building, requires higher quality results. In this context, the main problems concerning term extractors are the distinction between terms and non-terms, the quantity of noise in the results and the omission of relevant terms (silence). To improve the quality of the results, the task of term extraction is completed by CT scoring and ranking with the aim of classifying the extracted CTs according to their termhood probability, i.e. an evaluation of how likely it is that a particular CT is a term.

Scientific papers are used to construct domain specific corpora, sometimes along with other types of texts, such as technical documents, instruction manuals, web pages, sometimes as the only sort of documents included in the corpus (for example Kim et al., 2003; QasemiZadeh, 2014). Often, scientific corpora are used to study the inter-disciplinary scientific language or the structure of scientific discourse (Bertin et al., 2015). For the terminological purpose, the construction of the corpus depends generally on the objective of the terminological task and varies in several parameters, among which: domain and degree of specialization, reliability of sources, type of sources, and type of resources to be constructed (Cabr e, 2007). We choose scientific publications to construct our corpus because they are considered good sources of terminology, and they reflect the up-to-date state of scientific terminology. We work with peer-reviewed open access journals, to guarantee the quality and validity of the text as well as its accessibility. By comparing the specialist vocabularies that are actually used in texts with existing terminological dictionaries, we can identify novel terms that are commonly used among specialists but have not yet appeared in the online terminological databases.

The originality of our work lies in the choice to investigate the specific, human expert-oriented terminological task. First, we query relatively small corpora. Even if

¹ For a synthesis of the methods see for example Cabr e et al. (2001) and Drouin (2002).

nowadays the tendency is to use large corpora, we are interested in small text collections (about 20,000 words). The reason for this is that an expert has to build a new corpus for any new terminological project and this is not a trivial task. The small size of the corpora requires an accurate estimation of their degree of specialization: they should not concern too large a domain, but rather pertain to specific sub-domains. Even if the concepts of domain and sub-domain are rather naive and not formally defined, they are useful considerations for terminologists (Kageura, 1999). The other problem with large corpora is the number of CTs proposed by automatic extractors. For example, for the corpus of European patents concerning pharmacology, which comprises 2,500,000 words, 303,648 CTs were proposed (Mondary et al., 2013). Any new term added to a terminological database should be necessarily validated by a human expert. It is hardly imaginable (and not necessary) to humanly manage hundreds of thousands of CTs extracted from large text collections in a specific domain. Therefore, automatic strategies of filtering are necessary.

Our previous experience with a public French institution (Etablissement Français du Sang [National Blood Bank Organization], Bourgogne/Franche-Comté, France) revealed that some organizations do not hold large text collections (Plaisantin Alecu et al., 2012). This is confirmed by Drouin (2002), who used corpora, of sizes comparable to ours provided by a private company and described as representative of their terminological work, to test his term extractor. The disadvantages of using small corpora could be the lower efficiency of statistical measures and frequencies in automatic extraction of CTs, which could influence the quality of the extracted CTs.

We investigate the overlap of the CT sets extracted from scientific corpora with existing terminological databases, in particular with the objective of identifying novel terms for the enrichment of these resources. It is commonly admitted that there is a gap between the terminology used in texts and that used in existing terminological resources. This can be explained by the fact that terminological activity has been defined by what is called the general theory of terminology, established by Wüster and the Vienna Circle. This theory prescribes the onomasiological top-down approach to terminology: from concept to term. Therefore, the real usage of terms in context has been neglected in the process of establishing terminological dictionaries.

The overlap between terminological resources and specialized vocabularies extracted from corpora can serve different objectives; for example, evaluating the results of the term extractors. Other studies evaluate the relationship between a corpus and a terminological resource in terms of ‘lexical coverage’, a sort of adequacy between a corpus and a resource in order to match the most relevant resource to a given corpus (Ninova et al., 2005). Our approach is slightly different: for a given corpus and a given resource, we want to propose the most relevant terms from the texts that do not exist in the resource.

2. Methods

To extract terms from the corpora, we use three previously mentioned term extractors that are part of the Sensunique Platform² (Thomas et al., 2014): YaTeA (Aubin et al., 2006), Termostat (Drouin, 2002) and Acabit (Daille, 1995). The Sensunique Platform compiles the results proposed by each extractor into a unique list of CTs. The Platform is also linked to web services from an external resource: TermSciences³, a multidisciplinary terminology portal developed by CNRS-INIST (France). This allows us to check automatically which of the extracted CTs exist in this resource.

In the Platform, the termhood probability score is obtained by a weight assignment algorithm which takes into account two features: the number of extractors that propose the same term (which we call ‘multi-extraction’ and which is a sort of a ‘voting system’ for extractors) and whether or not a CT is present in the TermSciences (see more details in section 2.2). We hypothesize that the weighted sum of these features can provide an efficient ranking criterion for the extracted CTs in terms of their termhood probability.

This methodology has already been used for the task of establishing the lexicon of a Controlled Language (Thomas et al., 2015): the Sensunique Platform was developed towards this particular objective. One of the aims of the current research is to verify its suitability to more classical terminological tasks. It is important to know that the platform is analyst-oriented, i.e. it includes a CT management interface with numerous functionalities facilitating the analysis and validation of the extracted CTs (visualization of CTs in their corpus of origin, search and filters of the list of CTs, advanced concordancer for searching in the corpus of origin etc.).

2.1 Protocol

Our main study questions are a) whether scientific papers can be used to enrich the existing terminological databases, and b) how the ranking of automatically acquired lists of CTs could facilitate the task of term validation for a human analyst. More precisely, we want to estimate how many of the best ranked CTs will be validated as terms by a human expert. To answer these questions, we proceed with three steps:

- 1) automatic term extraction and ranking from a set of corpora of scientific papers using Sensunique Platform;
- 2) evaluation of the overlap of the CTs extracted from the corpora and those present in the TermSciences resource;

² Station Sensunique, <http://www.station-sensunique.fr/>

³ TermSciences, <http://www.termosciences.fr/>

- 3) evaluation of the top 200 CTs proposed by the platform for different corpora by domain experts.

To complete this research we also evaluate how the variability of corpora influences the automatic extraction results. Some additional results (performance of each extractor, distribution of termhood probability scores) are provided to facilitate discussion of the relevance of the features that are used to rank CTs.

2.2 Corpora and resources

To carry out our experiment, we constructed two corpora in two different sub-domains of the bio-medical sciences: *Mesenchymal stem cells* (C1) and *Vaccination* (C2). Each corpus consists of recent scientific papers taken from the chosen thematic issues (respectively 2011 and 2007) of the French specialized online medical revue *Médecine/Sciences*⁴. This journal is peer-reviewed and available in open access. The fact that the issues are thematic guarantees the homogeneity of the corpora. All the articles are written in French.

Each of the two initial corpora was used to obtain three different sub-corpora in the following way: for each sub-corpus one third of the papers were replaced by other papers from the same sub-domain. As a result, each pair of sub-corpora contains two thirds of common papers and one third of papers which are specific to each sub-corpus. This allows us to study the stability of the extracted CT sets with respect to variations in the corpus.

All the sub-corpora have similar sizes. The number of words in each of the six resulting sub-corpora is given in the Table 1.

Corpus	C1			C2		
	Mesenchymal stem cells			Vaccination		
Sub-corpus	C1a	C1b	C1c	C2a	C2b	C2c
Total number of words	17,213	17,839	17,266	21,042	21,244	21,075

Table 1: Corpus size

TermSciences is a multi-lingual and multi-purpose terminological database assembling vocabularies produced by major French research institutions (Khayari et al., 2006). Currently, it contains 650,000 terms related to 190,000 concepts. TermSciences includes three biomedical terminology resources: the French translation by the Institut National de la Santé et de la Recherche Médicale (INSERM) of the MeSH thesaurus from the US National Library of Medicine, the public health thesaurus of the Banque de Données de Santé Publique (BDSP) and the dictionary of human and mammal

⁴ <http://www.medecinesciences.org>

reproduction biotechnology of the Institut National de la Recherche Agronomique (INRA). It is difficult to know the number of terms that each of these resources contains, since such detailed information is not available on the website of TermSciences. According to the INSERM website⁵, the French version of MesH 2014 contains 83,399 terms distributed into 16 themes. The public health thesaurus of the Banque de Données de Santé Publique (BDSP) version 4 contains 12,825 terms⁶ and the paper version of the dictionary of human and mammal reproduction biotechnology of the Institut National de la Recherche Agronomique (INRA) contains over 200 terms (Bouroche-Lacomb, 2011).

The choice of the TermSciences terminological database was motivated by several factors: it has a large coverage of different subjects in bio-medicine, it combines several other terminology resources and it is the biggest multi-domain resource in France. For these reasons, we expect that terms from the two specific sub-domains of our corpora, *Mesenchymal stem cells* and *Vaccination*, are present in the TermSciences database.

2.3 Termhood probability scoring

Terms extracted from each corpus were ranked using the same weight assignment algorithm. For the needs of our experimentation, we used the following two criteria:

1. the number of extractors proposing a CT: the highest score is attributed to the CTs extracted simultaneously by the three extractors, then to those extracted by two of them, and finally to those extracted by only one extractor; this procedure, called multi-extraction (Plaisantin Alecu et al., 2012), has proved to give better results than using only one term extractor (21% higher recall and 9% higher precision values compared to the use of only one extractor). The results of the multi-extraction (on much bigger corpora and with a larger number of extractors) are also judged relevant by Mondary et al. (2013).
2. the presence of a CT in the external resource (TermSciences): the Platform verifies if a CT is already present in TermSciences; for the composed CTs, three types of attestations are looked for (with decreasing score attributed): a) the whole composed CT, b) its head **and** modifier separately, i.e. occurring in two different entries in TermSciences c) its head **or** modifier separately, i.e. either the head or the modifier occurring in TermSciences. For example, for the CT *cellules souches (stem cells)*, if the whole CT is not present in TermSciences, the Platform will look for its head (*cellules*) and/or its modifier (*souches*) separately. This procedure is motivated by the hypothesis that a composed CT containing an already attested terminological element is more likely to be a term than a CT without any terminological constituent.

⁵ Accessed at: <http://mesh.inserm.fr/mesh/presentation.htm> (20/05/2015).

⁶ Accessed at: <http://asp.bdsp.ehesp.fr/Thesaurus> (20/05/2015).

The combination of these different criteria results in a termhood probability score, ranked as shown in Table 2. The best termhood probability score (rank 1) is obtained by the CTs proposed simultaneously by three extractors and attested as a whole term in TermSciences. The second best score (rank 2) is given to the CTs proposed by two extractors and attested in TermSciences etc. The lowest termhood probability score (rank 12) is attributed to the CTs proposed by only one extractor without any attestation in TermSciences.

TERMHOOD PROBABILITY RANK	CRITERIA					
	Number of extractors			Attestation in TermSciences		
	1	2	3	whole CT	head and modifier	head or modifier
1			x	x		
2		x		x		
3	x			x		
4			x		x	
5			x			x
6			x			
7		x			x	
8		x				x
9	x				x	
10		x				
11	x					x
12	x					

Table 2: Termhood probability score

2.4 Evaluation

To evaluate the quality of the extracted CTs for each sub-corpus we proceeded as follows. We considered the terms which are present in TermSciences as valid terms and therefore we did not need to evaluate them by human experts. We can directly observe the number of these terms for each sub-corpus. For the rest of the terms, which have been extracted by the Sensunique Platform but are not present as a whole term in TermSciences (and therefore have termhood probability ranks below 3), we considered the top 200 terms. Two highly qualified human experts in the domain (professors of immunology) were consulted for the evaluation. Each expert was presented with a list of extracted terms and asked whether the CT corresponds to a term in the domain. The possible answers were: *yes*, *no* and *possibly* (for the cases that need deeper analysis or additional information).

Additionally, we measured the overlaps between the sets of CTs extracted from each sub-corpus. This gives us an indication of the stability of the extracted lists of CTs depending on modifications of the corpus within the same domain.

3. Results and Discussion

3.1 General results

Tables 3 and 4 present the general results of the analysis of each sub-corpus in terms of the number of CTs proposed per extractor and the number of CTs attested in TermSciences (any type of attestation).

	C1a	% total CTs extracted	C1b	% total CTs extracted	C1c	% total CTs extracted
Total words	17,213		17,839		17,266	
Total CTs extracted	5,173		5,072		5,242	
YaTeA	3,390	65.53%	3,379	66.62%	3,434	65.51%
Acabit	2,204	42.61%	2,146	42.31%	2,261	43.13%
TermoStat	1,489	28.78%	1,445	28.49%	1,481	28.25%
Total CTs present in TermSciences	<i>4,022</i>	<i>77.75%</i>	<i>3,935</i>	<i>77.58%</i>	<i>4,001</i>	<i>76.33%</i>

Table 3: General results for C1

	C2a	% total CTs extracted	C2b	% total CTs extracted	C2c	% total CTs extracted
Total words	21,042		21,244		21,075	
Total CTs extracted	5,894		5,655		5,586	
YaTeA	3,784	64.20%	3,592	63.52%	3,675	65.79%
Acabit	2,586	43.88%	2,516	44.49%	2,370	42.43%
TermoStat	1,583	26.86%	1,458	25.78%	1,535	27.48%
Total CTs present in TermSciences	<i>4,365</i>	<i>74.06%</i>	<i>4,215</i>	<i>74.54%</i>	<i>4,100</i>	<i>73.40%</i>

Table 4: General results for C2

The sum of the CTs extracted by the extractors is not equal to 100% of all the CTs extracted, because some CTs are extracted by several extractors; in these statistics they are counted separately for each extractor.

In general, the number of CTs extracted from each sub-corpus remains relatively stable, which means that this number varies little with small changes of the papers in the corpus. The percentage of CTs proposed by each extractor is also stable across the sub-corpora and moreover across the different corpora. YaTeA is the most prolific term extractor: it extracts between 63.52% and 66.62% of all extracted CTs; the results of TermoStat vary between 25.78% and 28.78% of all extracted CTs.

The number of the CTs present in TermSciences is stable across the sub-corpora and seems rather high (more than 73% for each sub-corpus). However, this result is to be handled with care, since all types of attestations are taken into consideration, even if only a part of a CT is found. Consequently, not all of the CTs attested will be finally validated as terms.

3.2 Distribution of termhood probability score and ratio of CTs attested in TermSciences

Tables 5 and 6 present for each corpus the ratio of the CTs extracted per specific termhood probability (TP) rank.

TP rank	C1a	% total CTs extracted	C1b	% total CTs extracted	C1c	% total CTs extracted
1	54	1.04%	52	1.03%	54	1.03%
2	165	3.19%	141	2.78%	153	2.92%
3	320	6.19%	308	6.07%	295	5.63%
<i>Total of CTs present in TermSciences as terms</i>	<i>539</i>	<i>10.42%</i>	<i>501</i>	<i>9.88%</i>	<i>502</i>	<i>9.58%</i>
4	105	2.03%	99	1.95%	108	2.06%
5	243	4.70%	232	4.57%	226	4.31%
6	12	0.23%	11	0.22%	12	0.23%
7	114	2.20%	124	2.44%	116	2.21%
8	719	13.90%	747	14.73%	775	14.78%
9	152	2.94%	172	3.39%	154	2.94%
10	84	1.62%	98	1.93%	90	1.72%
11	2,150	41.56%	2,060	40.62%	2,120	40.44%
12	1,055	20.39%	1,028	20.27%	1,139	21.73%
Total CTs extracted	5,173	100.00%	5,072	100.00%	5,242	100.00%

Table 5: Detailed results for C1 ratio of CTs per TP

TP rank	C2a	% total CTs extracted	C2b	% total CTs extracted	C2c	% total CTs extracted
1	55	0.93%	44	0.78%	57	1.02%
2	140	2.38%	147	2.60%	143	2.56%
3	306	5.19%	313	5.53%	309	5.53%
<i>Total CTs present in TermSciences as terms</i>	<i>501</i>	<i>8.50%</i>	<i>504</i>	<i>8.91%</i>	<i>509</i>	<i>9.11%</i>
4	111	1.88%	108	1.91%	118	2.11%
5	257	4.36%	230	4.07%	246	4.40%
6	13	0.22%	10	0.18%	15	0.27%
7	124	2.10%	118	2.09%	117	2.09%
8	803	13.62%	761	13.46%	745	13.34%
9	191	3.24%	186	3.29%	169	3.03%
10	120	2.04%	101	1.79%	117	2.09%
11	2,378	40.35%	2,308	40.81%	2,196	39.31%
12	1,396	23.69%	1,329	23.50%	1,354	24.24%
Total CTs extracted	5,894	100.00%	5,655	100.00%	5,586	100.00%

Table 6: Detailed results for C2: ratio of CTs per TP

It is also worth noting that for the two corpora, over 60% of the CTs extracted have the two lowest TP scores, i.e. they are rank 11 (extracted by one extractor and having a head or a modifier attested in TermSciences) and rank 12 (extracted by one extractor). This means that for the majority of CTs there is no agreement between different extractors as to what should be considered a term. To exemplify this fact, Table 7 presents the number of CTs extracted by two or three extractors and the number of CTs extracted by only one extractor, for C1.

Corpus	C1a		C1b		C1c	
Extractors	Number of CTs	% total CTs extracted	Number of CTs	% total CTs extracted	Number of CTs	%total CTs extracted
Acabit and YaTeA and TermoStat	414	8.00%	394	7.77%	400	7.63%
Acabit and YaTeA	392	7.58%	410	8.08%	426	8.13%
Acabit and TermoStat	95	1.84%	111	2.19%	116	2.21%
YaTeA and TermoStat	595	11.50%	589	11.61%	592	11.29%
Acabit	1,303	25.19%	1,231	24.27%	1,319	25.16%
YaTeA	1,989	38.45%	1,986	39.16%	2,016	38.46%
TermoStat	385	7.44%	351	6.92%	373	7.12%
Total CTs extracted	5,173	100.00%	5,072	100.00%	5,242	100.00%

Table 7: Multi-extraction for C1

The fact that the majority of CTs is extracted by only one extractor can be explained by the differences in the methods used by each extractor. Consequently, the number of CTs proposed by each extractor is different, as can be seen in Tables 3 and 4. Nevertheless, we make the hypothesis that being proposed by several extractors is a good indicator for a CT to be a term (see section 3.4 Expert evaluation).

The total of CTs attested as terms in TermSciences (ranks 1, 2 and 3) varies by 0.84% for C1 (from 9.58% to 10.42%, Table 5). This ratio is similar for C2 (from 9.66% to 9.90%, Table 6). We can therefore assume that the average ratio of attested terms in different corpora is around 9.50% of all extracted CTs.

3.3 Analysis of the performance of the extractors

To obtain a first evaluation of the performance of extractors, we tested the results against the terms present in the terminological database, i.e. the CTs attested in TermSciences as whole terms and extracted at least by one extractor. For each sub-corpus and each extractor, we calculated the precision (P) relative to the TermSciences terminological database, i.e., the ratio of the extracted CTs and attested in TermSciences as whole terms divided by the total number of the extracted CTs.

Tables 8 and 9 present the results of this evaluation for the two corpora. The first column (T) gives the number of CTs attested as a whole term in TermSciences for each extractor.

Extractor	C1a sub-corpus		C1b sub-corpus		C1c sub-corpus	
	T	P	T	P	T	P
Acabit	128	5.81%	118	5.50%	123	5.44%
YaTeA	466	13.75%	430	12.73%	429	12.49%
TermoStat	218	14.64%	198	13.70%	211	14.25%

Table 8: Evaluation of the extractors for C1

Extractor	C2a sub-corpus		C2b sub-corpus		C2c sub-corpus	
	T	P	T	P	T	P
Acabit	129	4.99%	130	5.17%	132	5.57%
YaTeA	444	11.73%	443	12.33%	451	12.27%
TermoStat	178	11.24%	166	11.39%	183	11.92%

Table 9: Evaluation of the extractors for C2

The results are constant between the sub-corpora and corpora. Acabit is the worst scored term extractor; its precision is significantly lower than that of the other extractors. Yatea and TermoStat receive similar precisions, but TermoStat performs slightly better for C1 and YaTeA for C2.

This first evaluation shows that each separate extractor proposes a high number of CTs, most of which are not present in the terminological database. These CTs can be potentially good term candidates to enrich the terminological database, but they have to be validated by human experts. It means that, for example, 83.25% of the CTs proposed by YaTeA (100%-13.75%, Table 8, C1a sub-corpus), namely 2840 CTs, have to be validated manually. In a previous study on similar corpora using the multi-extraction method (Plasaintin-Alecu, 2012), we demonstrated that when considering the whole set of CTs extracted by two or more extractors, the best precision is around 37%. Consequently, we can roughly estimate that about 60% would not be valid terms if we consider the entire list of extracted CTs. For this reason, it is useful to propose a ranking algorithm which assigns weights to the CTs and puts the best candidates at the top of the list. In order to validate the ranking algorithm that we propose, in the next section we present the results of the evaluation by human experts of the top 200 CTs, ranked by our algorithm (see section 3 Termhood probability scoring).

3.4 Expert evaluation

For each sub-corpus, we created a list of the top 200 best scored CTs which are not present as whole terms in TermSciences. These CTs correspond to rank 4 (proposed by three extractors and whose head **and** modifier are attested in TermSciences) and rank 5 (proposed by three extractors and whose head **or** modifier are attested in TermSciences). They were submitted to the experts for evaluation. Table 10 shows the distribution of these CTs per rank for each corpus.

TP rank	C1a	C1b	C1c	C2a	C2b	C2c
4	105	99	108	111	108	118
5	95	101	92	89	92	82
Total CTs	200	200	200	200	200	200

Table 10: Top 200 CTs for C1 et C2 not present in TermSciences as whole terms

The six different sets of 200 terms overlap, and as a result, a total of 595 unique CTs had to be evaluated by the experts: 332 unique terms in C1 and 269 unique terms in C2. To evaluate the stability of the extracted CTs depending on the choice of the papers in the corpus, we observe the overlap between the sets of CTs extracted from each pair of sub-corpora. Table 11 presents these results.

Corpus	Number of extracted CTs	% (of the total 595)
C1: C1a, C1b & C1c	14	2.35%
C1a & C1b	80	13.45%
C1a & C1c	82	13.78%
C1b & C1c	78	13.11%
Only C1a	24	4.03%
Only C1b	28	4.71%
Only C1c	26	4.37%
C1 (any sub-corpus)	332	55.80%
C2: C2a, C2b & C2c	84	14.12%
C2a & C2b	51	8.57%
C2a & C2c	62	10.42%
C2b & C2c	50	8.40%
Only C2a	3	0.50%
Only C2b	15	2.52%
Only C2c	4	0.67%
C2 (any sub-corpus)	269	45.21%
C1 & C2	6	0.01%

Table 11: Overlap between the sets of extracted CTs for the top 200 of CTs extracted from each sub-corpus

We observe that in C1 there is relatively little overlap between the three sub-corpora: only 14 CTs were extracted in total, while for C2 this number is 84. This means that the papers in the C2 corpus seem to be more homogeneous and replacing one third of the corpus has a very low impact on the sets of extracted terms. For the C1 corpus, the majority of CTs are shared between two sub-corpora, and each sub-corpus contributes with around 26 CTs (from 24 to 28).

Another important observation is the number of CTs that were extracted from both C1 and C2. These terms are only six in number and we can hypothesize that this is due to the fact that the two corpora contain articles on two different subjects (*Mesenchymal stem cells* and *Vaccination*) that use different terminologies. We can therefore suppose that the majority of extracted CTs are closely related to the subjects

of the corpora. Table 12 presents the six CTs extracted from both C1 and C2.

CTs extracted from both C1 and C2 (in French)	English translation
<i>cellules dendritiques</i>	<i>dendritic cells</i>
<i>diabète de type</i>	<i>diabetes type</i>
<i>efficacité clinique</i>	<i>clinical effectiveness</i>
<i>mécanismes régulateurs</i>	<i>regulating mechanisms</i>
<i>passages successifs</i>	<i>successive passages</i>
<i>réponse immunitaire</i>	<i>immune response</i>

Table 12: CTs extracted from both C1 and C2

Each CT was evaluated by one expert, who was asked whether they consider this CT as a valid term in the domain. The experts had a choice of three possible answers: *yes*, *no* and *possibly*. Five of the six terms from Table 12 were positively evaluated by the experts (with the answer *yes*), and the candidate term *diabète de type* was evaluated with the answer *no*. Table 13 presents the results for all sets extracted from the corpora.

Answer	C1a	C1b	C1c	Total C1	C2a	C2b	C2c	Total C2
<i>yes</i>	154	136	148	240	154	152	151	203
<i>possibly</i>	15	26	16	34	18	21	23	29
<i>no</i>	31	38	36	58	28	27	26	37
Total CTs	200	200	200	332	200	200	200	269

Table 13: Expert evaluation of the top 200 extracted CTs not present in TermSciences

This table shows that a large majority of extracted CTs were positively evaluated by the experts. Using these results we calculate the precision among the top 200 extracted CTs ranked by the Sensunique platform in two ways:

1. Strict evaluation: only the CTs evaluated with *yes* considered as correct;
2. Loose evaluation: CTs evaluated with either *yes* or *possibly* considered as correct.

	C1a	C1b	C1c	Total C1	C2a	C2b	C2c	Total C2
Strict evaluation	77.00%	68.00%	74.00%	72.29%	77.00%	76.00%	75.50%	75.46%
Loose evaluation	84.50%	81.00%	82.00%	82.53%	86.00%	86.50%	87.00%	86.25%

Table 14: Precision for the top 200 extracted CTs for each corpus

Table 14 presents the precision values for this evaluation. These results are very promising. In fact, we can see from Table 14 that for all sub-corpora the precision for the strict evaluation is above 68%, and for five out of six sub-corpora it exceeds 74% and an average of about 75% of the CTs were evaluated as correct. Furthermore, the precision is above 81% for the loose evaluation. This means that the criteria that we have considered allow us to perform ranking with little noise among the top results. At

the same time, as shown in Tables 8 and 9, the results of the three extractors have little overlap with the TermSciences database. This means that the extraction from scientific corpora is an adequate approach for the enrichment of terminological databases.

We work only with the top 200 extracted CTs which are not present in TermSciences, and thus this evaluation concerns only the criteria corresponding to ranks 4 and 5, as the CTs with higher ranks feature much further down the list. The evaluation of all ranks can be carried out but it is very expensive because of the large number of extracted CTs.

4. Conclusions

Using the multi-extraction method implemented in the Sensunique platform, we have carried out the extraction of terms working with relatively small corpora of about 20,000 words. The number of candidate terms extracted from each corpus is very large, about 6,000 (single word terms or multiword terms) which makes the results difficult to use by the experts. The reason for this high number of CTs is that the multi-extraction method combines the results of three different extractors. In this context it is important to consider ranking algorithms that order the lists of extracted CTs by relevance. In our study we considered two major ranking criteria based on an external terminological resource and on votes by several extractors.

The main objective of our study was to propose new strategies for the enrichment of existing terminological resources using scientific corpora. In general, language evolves quickly and there is little overlap between terms found in terminological databases and terms actually used in scientific writing. For example, our results (Tables 5 and 6) show that only about 10% of the extracted CTs are present in the TermSciences resource, which means that many of the extracted CTs, if validated, could potentially be used to enrich this terminological database. Furthermore, the expert evaluation of the top 200 terms for each sub-corpus shows clearly that the majority of these CTs are correct terms in the respective domains. We can therefore conclude that scientific corpora constitute suitable sources for terminological extractions.

In general, the quality of the results of extractors reduces for smaller sized corpora. For example, working with small corpora we have previously found (Plaisantin Alecu et al., 2012) that the best extractor, YaTeA, reaches 58% of recall and the best precision value for a single extractor, Termostat, to be 28%. For this reason, it is interesting to consider the multi-extraction method as it proposes more relevant results in terms of recall. The disadvantage of the multi-extraction, i.e. a larger number of CTs compared to the results of only one extractor, can be compensated using ranking criteria for the extracted CTs. The ranking algorithm that we propose allows us to obtain high precision among the top results, i.e. 75% of the best ranked CTs can be used to enrich the terminological database. Consequently, we have shown that we can produce good results, even if we work with relatively small corpora.

5. Acknowledgements

The authors thank professors Estelle Seillès (Etablissement Français du Sang (National Blood Bank Organization), Bourgogne/Franche-Comté) and Dominique A. Vuitton (Research Federation « Cell and Tissue Biology and Engineering » FED 4234, University of Franche-Comté) for their expert assistance.

6. References

- Aubin, S., & Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, Springer, pp. 380–387.
- Bertin, M., Atanassova, I., Larivière, V. & Gingras, Y. (2015). The Invariant Distribution of References in Scientific Papers. *Journal of the Association for Information Science and Technology (JASIST)*, doi: 10.1002/asi.23367.
- Bourigault, D., Aussenac-Gilles, N. & Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes. Un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle* 18 (1), pp. 87–110.
- Bourigault, D. & Jacquemin, C. (2000). Construction de ressources terminologiques. In J.-M. Pierrel (ed.) *Industrie des langues*. Hermès, Paris, 2000, pp. 215-233.
- Bourigault, D. (1994). *Lexter: Un Logiciel d'EXtraction de TERminologie: Application à l'acquisition des connaissances à partir de textes.* EHESS, Paris.
- Bouroche-Lacomb, A. (2001), *Biotechnologies de la reproduction chez les mammifères et l'homme : vocabulaire français-anglais*, INRA Editions.
- Cabré, M.T. (2007). Constituer un corpus de textes de spécialité. *Cahier du CIEL*, pp. 37-56.
- Cabré, M.T., Estopà, R. & Vivaldi, J. (2001). Automatic term detection: A review of current systems. In D. Bourigault D., C. Jacquemin & M.-C. L'Homme (eds.) *Recent Advances in Computational Terminology*. Amsterdam / Philadelphie, John Benjamins, pp. 53–87.
- Cerbah, F., & Daille B. (2006). Une Architecture de Services Pour Mieux Spécialiser Les Processus D'acquisition Terminologique. *Traitement Automatique Des Langues (TAL)* 47(3), pp. 39–61.
- Enguehard, Ch., & Pantera L. (1995). Automatic Natural Acquisition of a Terminology. *Journal of Quantitative Linguistics* 2 (1), pp. 27–32.
- Daille, B. (1995). Repérage et Extraction de Terminologie Par Une Approche Mixte Statistique et Linguistique. *TAL. Traitement Automatique Des Langues* 36 (1-2), pp. 101–118.
- Drouin, P. (2002). *Acquisition Automatique Termes: L'utilisation Des Pivots Lexicaux Spécialisés*. PhD thesis, Université de Montréal, 2002.
- Drouin, P. (2003). Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology* 9(1), pp. 99–115.
- Ibekwe-Sanjuan, F. (2006). Repérage et annotation d'indices de nouveautés dans les écrits scientifiques. In *Indice, index, indexation. Actes du colloque international, Université Lille-3*, pp. 1-11.

- Jacquemin, Ch. & Bourigault, D. (2003). Term extraction and Automatic Indexing. In Mitkov R. (ed.) *Oxford Handbook of Computational Linguistics*. Oxford University Press.
- Kageura, K. (1999). Theories of terminology: A quest for a framework for the study of term formation ». *Terminology* 5(1), pp. 21-40.
- Khayari, M., Schneider, S., Kramer, I., & Romary, L. (2006). Unification of multi-lingual scientific terminological resources using the ISO 16642 standard. The TermSciences initiative. *arXiv preprint cs/0604027*.
- Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1), pp. 180–182.
- Mondary, T, Nazarenko, A., Zargayouna, H., & Barreaux, S. (2013). Aide À L'enrichissement D'un Référentiel Terminologique: Propositions et Expérimentations. In *20e Conférence Sur Le Traitement Automatique Des Langues Naturelles (TALN'2013)*, pp. 779–786.
- Ninova, G., Nazarenko A., Hamon T. & Szulman S. (2005). Comment Mesurer La Couverture D'une Ressource Terminologique Pour Un Corpus. *TALN 2005*, 2005.
- QasemiZadeh, B. & Handschuh, S. (2014). The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *COLING 2014: 4th International Workshop on Computational Terminology*.
- Plaisantin Alecu, B., Thomas, I., & Renahy, J. (2012). La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques, Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, ATALA/AFCP, pp. 511-518.
- Thomas I., Plaisantin Alecu B., Germain B. & Betbeder M.-L. (2014). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources. In A. Abel et al. (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: EURAC research, Volume: II*, pp. 707-726.
- Thomas, I., Laroche, L., Plaisantin-Alecu, B., Betbeder, M.-L., Seillès, E., Renahy, J., Blagoskonov, O. & Vuitton, D.-A. (2015). Computerization of a 'controlled language' to write medical standard operating procedures (SOPs). In *Proceedings of Conference on Health and Social Care Information Systems and Technologies, HCist 2015 October 7-9, 2015*, Procedia Computer Science, Elsevier (to appear).

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



medialatinitas.eu. Towards Shallow Integration of Lexical, Textual and Encyclopaedic Resources for Latin

Krzysztof Nowak¹, Bruno Bon²

¹ Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland

² Institut de recherche et d'histoire des textes, CNRS, Paris, France

E-mail: krzysztof@ijp-pan.krakow.pl, bruno.bon@irht.cnrs.fr

Abstract

medialatinitas.eu is a lightweight web application which integrates dictionaries, corpora and encyclopaedic resources for Latin. The integration takes place principally on the level of the user-friendly interface, so no explicit links between resources are provided. The main objectives of *medialatinitas.eu* are: improving access to distributed data; challenging separation of linguistic and encyclopaedic information in lexicographic description; compensating for deficiencies of existing lexicographic resources; building a community of users who apply computational methods in their study of Latin texts.

As for the architecture, *medialatinitas.eu* is implemented as a mashup application: the user's query (as of now, only lemma search is supported) is processed and despatched to both local and distant services (RESTful APIs, SPARQL endpoints); the results are subsequently returned and displayed on the main page as a set of separate widgets. The widgets may contain short concordance lines and tables, but special attention has been given to alternative ways of content presentation, namely charts and visualisations. The widgets are provided with rich graphical hints and hold together thanks to such narrative devices as interpretative notes or explicative commentary. As a whole the widgets contribute to extensive description of Latin lemmas according to their grammatical, semantic and cultural properties.

Keywords: lexicographic mashup; data reuse and integration; visualisation;
dictionary-corpus interface; Medieval Latin

1. Introduction

Latin was one of the most widely used languages in European history. In its spoken and written form it was the language of daily communication, law, literature, and science for over fifteen centuries on the territory stretching from Spain through Germany to Poland and from Sweden through Croatia to Italy. This geographical, chronological and functional variation is reflected in a large number of texts which, in turn, gave rise to a vast body of secondary literature of which dictionaries form an essential part.

The multifarious resources, even if partly digitised by now, remain still widely dispersed and do not easily lend themselves to integrated search. Moreover, separate electronic text collections usually cover only small proportion of the texts preserved to our times and do not have any pretensions to representativeness. Often, they would also be available only through an interface that does not allow for any subtler

query. As for the electronic dictionaries, their selective spatio-temporal coverage, multilingual definitions, and differing editorial styles make that they cannot be said to account for Latin development in any systematic way if consulted separately.

medialatinitas.eu is a web application which aims for a meaningful integration of textual, lexicographic and encyclopaedic resources for Latin through a user-friendly and attractive interface. It is also an attempt to generate a coherent narrative from incomplete data despite variety of technologies in use. The integration is said to be shallow, since the heterogeneous content (dictionaries, encyclopaedias and corpora) has been linked only to the degree needed for its unified query and retrieval. It takes place, then, at the level of the web interface which, thus, constitutes presentational layer and a point of access to the services running in the background. At the moment, *medialatinitas.eu* is intended in particular for academic audience (lexicographers, linguists, historians etc.), but teachers and students of the medieval literature should also find it useful.

2. Data, Goals, Design

2.1 What to integrate: data

medialatinitas.eu makes extensive use of the existing digital resources for the Latin language and culture. The data which are going to be integrated within the web application may be roughly divided into three groups (Figure 1):

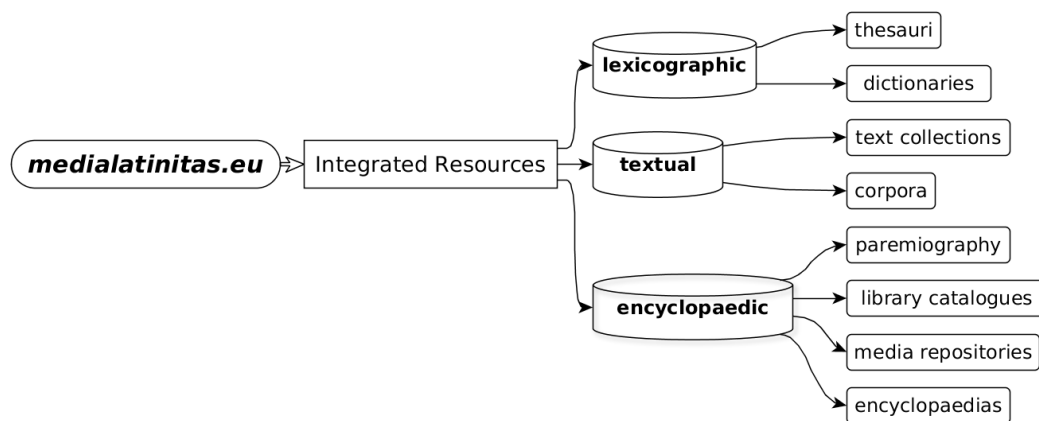


Figure 1: *medialatinitas.eu* resources: general outlook.

1) lexicographic resources: dictionaries of Classical, Medieval and Modern Latin, both academic (e.g. *Novum Glossarium Mediae Latinitatis*, *Lexicon Mediae et Infimae Latinitatis Polonorum*) and community-based (e.g. *Latin Wiktionary*); dictionaries and thesauri of ancient and medieval placenames, gazetteers (*Pelagios Project*, *Orbis Latinus*, *Getty Thesaurus of Geographic Names*, *GeoNames*);

2) corpora (e.g. *Fontes. Corpus of Polish Medieval Latin*, *Croatiae Auctores Latini*) and text collections (e.g. *Perseus Project*, *Patrologia Latina* etc.);

3) encyclopaedic resources: encyclopaedias (*Wikipedia*, in particular its Latin version), paremiological resources (*Latin Wiktionary*), document and image repositories (*Europeana*), library catalogues (*Internet Archive*, *Open Library*), lists of medieval authors (e.g. *Novum Glossarium*, *VIAF*), hybrid resources (e.g. *BabelNet*).

Regarding their origin, the vast majority of resources were created by external institutions and only very few are the result of in-house projects (*Novum Glossarium*, *LMILP*, *Fontes*). As one may suppose, this and the format in which the data are generated, imply different strategies of access and reuse, and contribute to the complexity of the integration task, as the resources are mostly exploited “as they are”.¹ In-house dictionaries come originally as TEI-conformant XML files based on a shared encoding scheme. Both external and in-house corpora were delivered as XML files containing lightweight document mark-up for meta-data and structural features of the text. Each corpus text was tokenised and annotated with part-of-speech (PoS) and lemma labels. The annotation was performed using the *TreeTagger* (Schmid 1994). The Latin parameter file that the tagger requires was based mostly on the texts from the *Perseus Digital Library* and the *Index Thomisticus*; however, this is likely to be changed in the nearest future, once the work on the Medieval Latin parameter file comes to an end (*Omnia Project TreeTagger*).

The majority of external resources are exploited through their public RESTful APIs or SPARQL endpoints; so the *medialatinitas.eu* remains to some degree agnostic of the original data formats or encoding schemes (Figure 2). Regardless of their origin, even the locally hosted data are exposed to the web application through the APIs:

- dictionaries deployed in an *eXist-db* instance are exposed through respective RESTful API;
- textual corpora are deployed in a *CQPWeb* (Hardie, 2012) instance; since for the moment *CQPWeb* does not offer a web API, it is used only as an advanced corpus research tool;
- OCR texts and less-structured text collections are stored in *eXist-db* Lucene-based indexes and exposed through a RESTful API.

Yet, the role of locally running services is by no means limited to only exposing data, since they also serve to enrich, compute on and prepare data for subsequent display:

- the *WikiLexicographica* (Bon & Nowak, 2013), an implementation of the *Semantic MediaWiki* (Krötzsch et al., 2006), combines dictionaries with geographical and chronological dimension, thus enabling rich data representation;
- an *R* (R Core Team, 2015) session is exposed to the web application through the *OpenCPU API* (Ooms, 2014) and permits computation on corpus and lexicography

¹ For explanation, see below.

resources; *rcqp* package (Desgraupes & Loiseau, 2012) is used to connect to the *CQP* engine; A. Guerreau’s scripts for lexical statistics (*Medialatinitas Github*) allow to find co-occurrences of the lemma in the corpus, while S. Evert’s *wordspace* package (Evert, 2014) is employed to calculate word similarities based on their distributional features.

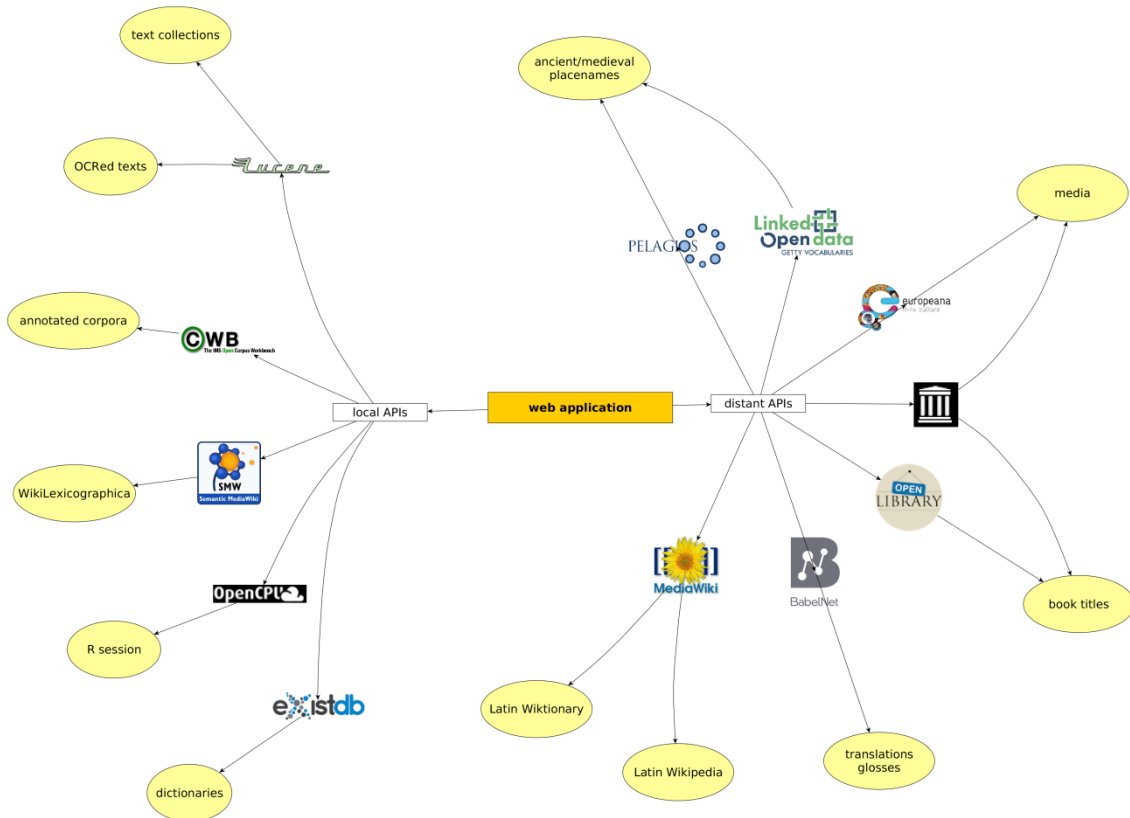


Figure 2: *medialatinitas.eu*: exploited APIs.

2.2 Why integrate: objectives?

medialatinitas.eu was created in order to:

- 1) improve access to distributed resources and facilitate the dictionary writing process;
- 2) stimulate research on Medieval Latin vocabulary through linked resources and popularise an innovative approach to the study of Latin text;
- 3) integrate a community of experienced and early-stage researchers who want to apply computational methods in Latin philology, history and linguistics.

2.2.1 User's commodity

On the most intuitive level, handling scattered resources results in losing time and energy. This primarily stems from the very fact that the data are stored in different locations. Not only do the users have to consult multiple web pages, but they can never guarantee the accessibility of the resource of their choice, as its availability depends entirely on whether the distant service is actually running. Even if it is, each service or repository the user has to consult forces them to respectively adapt their search strategy, remind of the query syntax or verify the integrity of the data. The latter, in particular, may often be difficult to assess, as many databases or text collections still lack appropriate documentation which would explain text or dictionary origin, its scope, data encoding scheme adopted and so on. In the worst scenario, a scarcity of information would increase the disadvantages inherent in many non-research-driven web resources, such as a lack of quality control, unclear or dubious choice principles, fragmentary and subjective character.

2.2.2 Answering old and asking new research questions

Yet, the user convenience, albeit important, is not the main objective of the *medialatinitas.eu* project. The principle that underpins the design of the present web application is to challenge the separation of knowledge components that should effectively cooperate in comprehension of the Medieval Latin text and culture. To achieve this goal and to compensate for the deficiencies each separate resource presents, *medialatinitas.eu* enables their concurrent, yet meaningful retrieval, and intertwines them in order to construct a coherent account of a word's meaning potential, its grammatical and syntactical properties and cultural function. At the same time, *medialatinitas.eu* promotes alternative forms of access to linguistic data (charts, maps etc.) and their reuse in new research contexts.

At a more specific level, *medialatinitas.eu* builds upon its linguistic content by addressing those issues which are either typical of a lexicography description in general or which affect Medieval Latin dictionaries in particular, namely:

- 1) limited account of variation of the Latin vocabulary;
- 2) limited or inadequate frequency information which is based mainly on manual excerption of the linguistic evidence;
- 3) purely linguistic approach to sense definition.

Numerous benefits that come from closer integration of lexicographic and corpus resources need not be enumerated here. Within the main interface of the *medialatinitas.eu* corpus, data are used to shed more light on the distributional properties of the Latin vocabulary. These are handled unsatisfactorily in the Medieval Latin dictionaries which did not adopt any coherent system of marking, for example, word frequency, except for such imprecise labels as 'more often' or 'very

often'. This, in turn, makes distinguishing between widespread and limited phenomena often a challenging task, as the latter (*hapax legomena* included) are being traditionally given relatively more space than high-frequency lemmas. Moreover, existing evaluation of the frequency of word or grammatical/syntactical pattern is far from ideal, since it is based on evidence which was manually retrieved from the sources (Guerreau-Jalabert & Bon, 2010; Bon, in print). The dictionaries also often fail to provide an adequate account of the diachronic, diatopic and genological features of the word use. On the one hand, they would often overestimate stability of semantic or grammatical patterns through the ages, while neglecting their changing function and dynamic distribution across the text genres. On the other hand, the available dictionaries (some of them still in progress) cover neither all periods nor all geographical zones of Latin development. Targeted corpus query may compensate for their shortcomings in this regard.

Equally important are reasons for closer integration of encyclopaedic data. *medialatinitas.eu* draws on the research of modern linguistic theory which demonstrates that the distinction between linguistic competence and real-world knowledge is not as clear-cut as the lexicographic practice shows (Geeraerts, 2000). *medialatinitas.eu* searches, then, for a compromise between the rigour of purely linguistic definition and the fact that the users of historical dictionary usually need more information when trying to understand ancient text, as the amount of the shared cultural background is necessarily significantly limited. This is the more remarkable, as Medieval Latin was for centuries the language of scientific, theological and philosophical writing, so exhaustive dictionaries (as the majority of those currently in print are) inevitably have to deal with this terminological richness². Although medieval terminology calls for different sense defining strategy than one applied in general lexicography, one often comes across definitions that, due to their purely linguistic character, are virtually void of any explicative potential. Meaningful reuse of encyclopaedic resources in the *medialatinitas.eu* application will help to tackle such specific cases and enrich dictionary content in general.

Finally, a closer integration of encyclopaedic and lexicographic data is desired for practical reasons. A good example in that respect might be proper names which are traditionally excluded from Medieval Latin general dictionaries. Yet, the correct decoding of place or personal names is crucial for understanding ancient text and constructing its referential layer. As a result, the readers of a medieval author will often find themselves consulting dictionaries and encyclopaedias at the same time. Aside from user convenience, however, including proper names will be of benefit, for instance, when describing common nouns if the latter are motivated by the former or vice versa (e.g. *aqua* 'water' is a component of many place names), etc.

2.2.3 Building a community of users and developers

² It makes some researchers claim that the Medieval Latin language was practically a special language (Bon, 2013).

Finally, the present work aims to integrate a community of developers and researchers. Although there now exists an active community of digital medievalists and the number of researchers who apply computational methods in their work on medieval texts has been steadily growing, until now no large-scale effort has been undertaken in order to integrate distributed data or to help developers embed their code snippets into a larger application. The same is true of pedagogical resources: despite numerous individual initiatives that have been launched (e.g. on-line bibliographies etc.), researchers willing to exploit automatic methods in their work cannot refer to any set of guidelines which would be appropriate for Latin text processing and query. This is why *medialatinitas.eu* will enable users to contribute their widgets³ as R and JavaScript code snippets responsible for single, yet self-contained functionality. Finally, the knowledge base that will constitute an important part of the *medialatinitas.eu*⁴ will provide users, on the one hand, with a curated collection of guidelines, showcases and links, and, on the other hand, with a complete description of digital medievalists' workflow - from the OCR to the corpus query.

2.3 How to integrate: application design and architecture

In the current development stage, *medialatinitas.eu* sticks with an integration model that could be characterised as ‘shallow’. The word is, however, used in the pregnant sense, as it is meant to describe implementation which is shallow, lightweight and agile at the same time.

The present integration model is called *shallow*, firstly because the data are not provided any explicit links and integration takes place principally in the application user interface (UI). As for now, virtually no effort has been put into harmonising different classes of resources, also same-class data are stored or dynamically queried “as they are”. Dictionaries, corpora and encyclopaedias do not refer to any common system of identifiers; therefore, for example, there is no formal connection established between the dictionary headword AQUA ‘water’, the lemma AQUA in the annotated corpus, and the *Latin Wikipedia* article for AQUA. As was already said, the same applies for same-class data, so, for instance, there is no inherent mapping between the entry AQUA in the DuCange’s *Glossarium mediae et infimae latinitatis* and its equivalent in the *LMILP*; similarly, there exists no explicit link between two identical lemma labels in separate corpora, if they have been annotated with different lemma sets. *Ad-hoc* equivalence between two dictionary headwords or lemmatised word forms is established if they have an identical orthographic form and share a PoS label. Other resources are currently retrieved based on a simple full-text query.

Secondly, *medialatinitas.eu* is designed as a three-level deep application (Figure 3) which offers the user to look up, for each lemma: 1) a general overlook; 2) an extended view; 3) an advanced view.

³ See below.

⁴ The knowledge base is beyond the scope of the present paper.

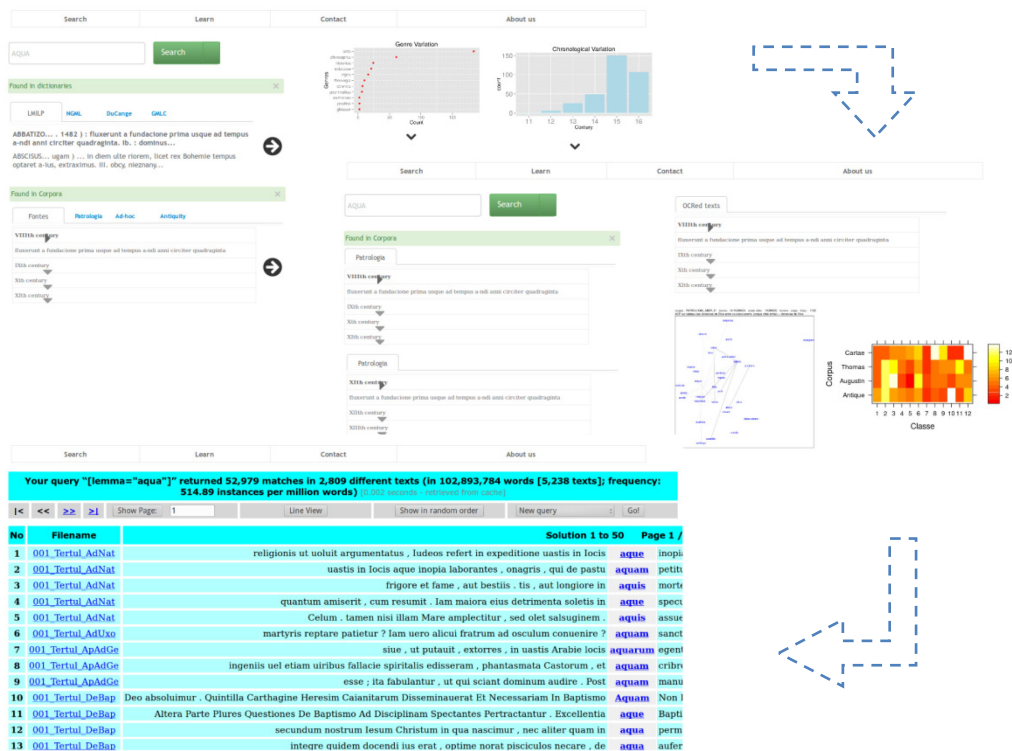


Figure 3: Three levels of the *medialatinitas.eu* application: 1) general view; 2) extended view; 3) native application (here, *CQPweb*).

1) When visiting the main page, the user initially comes across a simple search form. Once the query phrase is specified (currently only lemma search is supported), it is next despatched to locally and remotely running services and APIs. The returned results are processed and, subsequently, displayed on the same page.⁵ Its layout is built around a grid system and consists of a series of separate widgets, each responsible for displaying some portion of information about the word in question. As a whole, the widgets contribute to a general, yet varied outlook of the word meaning, its linguistic properties and distribution. The widgets that have been implemented so far present, for instance: 1) short excerpts from definitions of the Classical and Medieval Latin dictionaries; 2) short extracts from corpora concordances; 3) selected morpho-syntactic properties (inflectional type, gender, tense or case endings etc.) of the word (retrieved from the electronic dictionaries); 4) distribution of word forms in the corpora; 5) diachronic and genological distribution of the lemma; 6) co-occurring terms in selected corpora; 7) similar words in selected corpora; 8) translations and similar terms (retrieved from the *BabelNet*); 9) links to the *Latin Wikipedia* pages whose text contains the word in question; 10) list of quotations which contain the searched lemma (retrieved from the *Latin Wiktionary*); 11) list of titles of literary works which contain the lemma (retrieved from the *Internet Archive*); 12) list of images (Figure 4) whose description contains the lemma (retrieved from the *Europeana*); 13) map of the place names (Figure 5) that contain the lemma (retrieved from the *Pelagios Project*, *Getty Thesaurus of Geographic Names* and *GeoNames*).

⁵ Single Page Application (SPA) model of web application design is adopted here.

medialatinitas.eu employs various forms of data display; widgets are, thus, implemented as tables (1–3, 8) or lists (9–12), but also as charts, visualisations (4–7) and maps (13).⁰

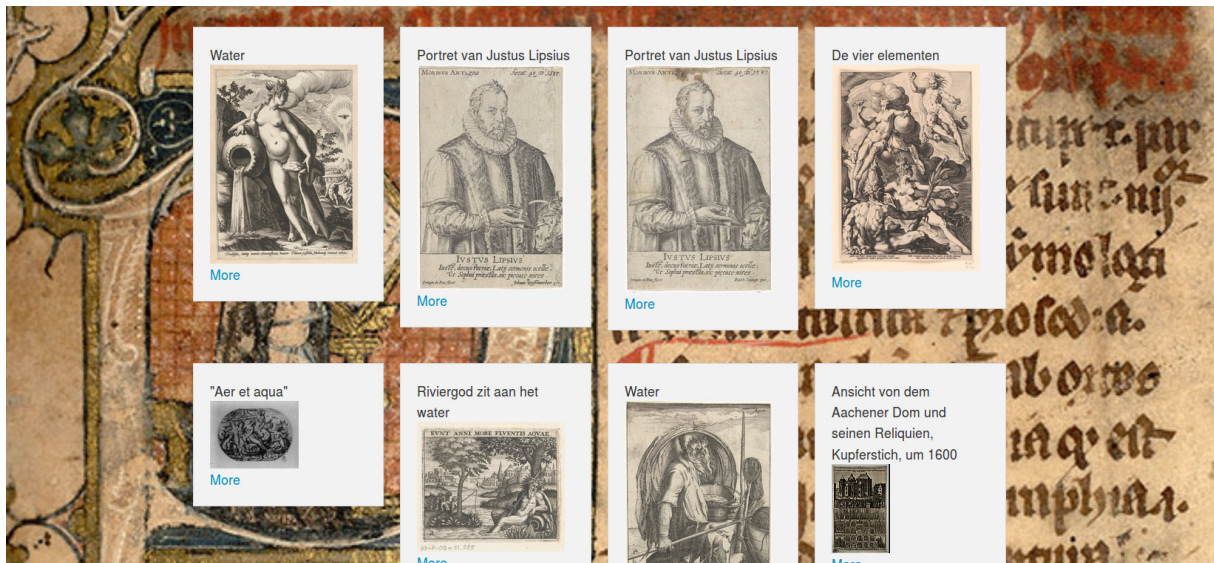


Figure 4: Media widget: images whose description matches the string *aqua* ‘water’ (fetched from the *Europeana*).

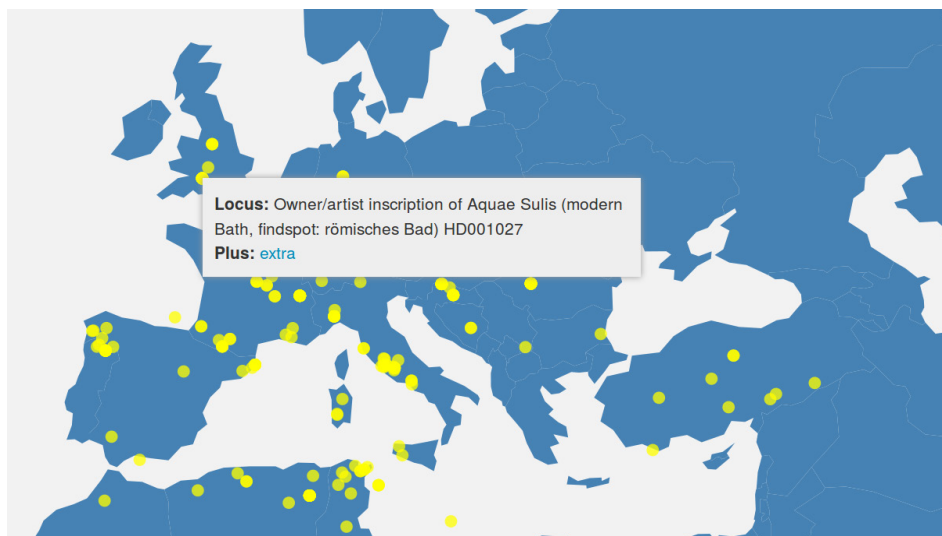


Figure 5: Map widget: yellow points represent ancient place names composed of the lemma *aqua* ‘water’ (geographical coordinates and labels are fetched from the *Pelagios Project API* and visualised with the *d3.js* library).

II) the extended view, which is beyond the scope of the present paper, is accessible upon clicking on any of the main page widgets and offers a more detailed and focused perspective on the selected properties of the lemma. For the moment, only *shiny*-based (Chang et al., 2015) dashboard for lexical statistics has been developed.

III) the native application (*CQPweb*, *eXist-db*, *R shiny* interface etc.) is accessible

from the extended view and constitutes the deepest layer of the *medialatinitas.eu* web interface.

The present application can be considered as *lightweight*, as there is no intention for it to become a complete virtual research environment. It is conceived as a modular platform that allows rapidly plugging in new widgets and testing dynamically alternative modes of linguistic data representation. Because it eventually always refers the users to a native application, there is no ambition to replace any existing software, as it is believed that such mature tools as, for instance, *CQPweb*, already offer an exceptional set of features that one can most effectively build on. As a result, *medialatinitas.eu* is *agile*⁶, since it is open to further expansion and should change according to the research interest and needs of its users and developers.

3. Discussion

3.1 *medialatinitas.eu* as a mashup application

In its design principle, *medialatinitas.eu* most resembles a mashup, which is defined as “a composite application developed starting from reusable data, application logic, and/or user interfaces typically, but not mandatorily, sourced from the Web” (Daniel & Matera, 2014: 3). It is ‘composite’, as it integrates data from more than one web service, each of which is a full-blown web application or a SPARQL endpoint. Following Daniel and Matera’s classification, *medialatinitas.eu* may be further described:

- regarding its “composition”, as a hybrid mashup, for the integration takes place both in the application logic and in the UI layer;
- regarding its “domain” or purpose, as a scientific, discovery-driven mashup;
- regarding “environment” or “deployment context”, as a web mashup in which logic layer is distributed over client and server: whenever a small data portion is involved, the client application written in AngularJS is responsible for processing Ajax calls, computing on their results and presenting them; however, once larger datasets come into play, especially when the user switches from a general view to the more specific one or when heavy calculation is to be applied, the burden of processing shifts towards the server and the client need only consider the visualisation of the returned data.

The user’s lemma query is passed to a mediator which subsequently transfers it to a series of wrappers. These, in turn, execute API calls and return back the results. The mediator, then, tackles the syntactic heterogeneity of the data, while wrappers deal with the idiosyncrasies of each source, thus resolving schematic heterogeneity. The

⁶ The use of this term is distantly inspired by the notion of *agile software development*.

problem of semantic heterogeneity of the data remains, as aforementioned, unresolved and this needs to be addressed in the nearest future by compiling a canonical list of lemmas that could be used to harmonise headwords of dictionaries, corpus annotations and encyclopaedic entities.

3.2 Meaningfulness, narrative and reproducible research

Rather than only assemble pieces of information in one place, *medialatinitas.eu* aims to provide whenever possible a relatively exhaustive and coherent narrative of each lemma. As for the exhaustiveness, the variety of the resources employed assures that no crucial level of word description is omitted. Dictionaries, apart from the obvious semantic information, provide also morphological, orthographical, syntactical and pragmatical information. Corpora contribute to the description of frequency, collocational features and computed meaning of a word. They are also a valuable source of knowledge about diachronic evolution of the lemma. Finally, the cultural component is covered by the use of paremiological resources, iconographic evidence (which helps to trace down allegorical sense), thesauri (for example, plant names) etc.

medialatinitas.eu should provide its users with a coherent and meaningful narrative for three reasons. Firstly, it is a reaction to the growing popularity of automatically compiled on-line content aggregators in which the very fact of juxtaposing multiple resources seems often to suffice as their *raison d'être*. Such a seemingly objective form of data presentation, at the same time, obscures the fact that the composition itself is already an interpretation. Secondly, the presence of contextualising, explicative or interpretive commentary seems to be what may distinguish human-oriented research applications from the popular, yet mainly machine-oriented resources, such as *WordNet* or *BabelNet*. Thirdly, *medialatinitas.eu* is also an exercise of a new form of lexicographic discourse in the era of linked linguistic data.

At the most basic level, the narrative “glue” is generated in the form of short introductory phrases which precede each widget or widget group. Being functionally equivalent to the headers, they do not add any substantial information; instead they, first, enable users to get instant insight into what linguistic or cultural phenomenon is represented in a specific section of the page and, second, make possible reading the whole page as a continuous text.

Aside from that, the narrative is built across the page by means of three other devices:

- 1) graphical and textual hints;
- 2) explicative and interpretative passages;
- 3) dynamically generated reports.

Graphical and narrative hints that the users find all over the interface indicate quality, scope and completeness of the presented data. Since *medialatinitas.eu* is to be a research tool, the user needs to be able to assess, first of all, whether they may safely draw conclusions from the gathered resources, and, secondly, whether insights offered in visuals, such as maps or charts, are of more than decorative value. To this goal, graphical signs and corresponding labels have been employed throughout the page, which signal:

- whether a widget was built on a resource of high, low or unknown quality;
- which chronological and geographical dimension a specific resource represents and
- whether it covers some phenomenon fully or only partially.

In a practical case of an excerpt from the *LMILP*, the quality, scope and coverage would be set *resp.* as “high (academic)”, “10-15th c., Poland”, and “full”, whereas in the case of an OCR-ised text they would be specified as “low (OCR)”, “6-12th c., Europe” and “partial”.

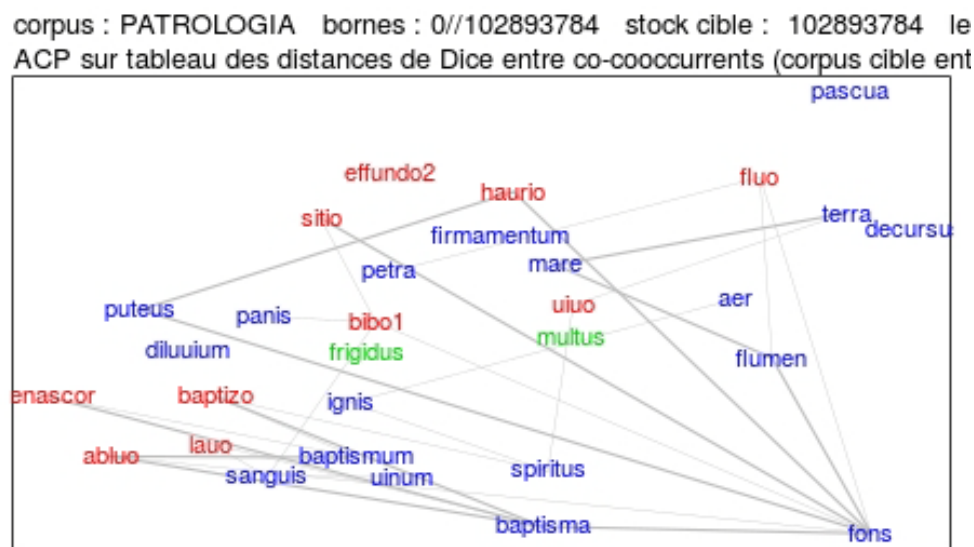


Figure 6: PCA chart: computed co-occurrences of the lemma *aqua* ‘water’ in the *Patrologia Latina* corpus (generated with A. Guerreau’s R script).

In the *medialatinitas.eu* visualisation widgets are accompanied by a short passage whose role is to explain what procedures have been applied to yield the results and to help with their interpretation. There are at least two reasons for providing such an explanation. First, *medialatinitas.eu* sticks with the reproducible research paradigm. At any time, the user may learn not only how specific visualisation was generated, but also explore its theoretical background. Second, some less standard forms of data presentation simply do not suffice without commentary text, if they are to be more than a decorative device. Whereas a barplot illustrating diachronic distribution of a specific word is relatively self-explanatory, the same cannot be said about the

boxplots, PCA charts (Figure 6) or co-occurrence barplots (Figure 7) which should be accompanied by a supplementary text if they are not to overwhelm a less advanced user.

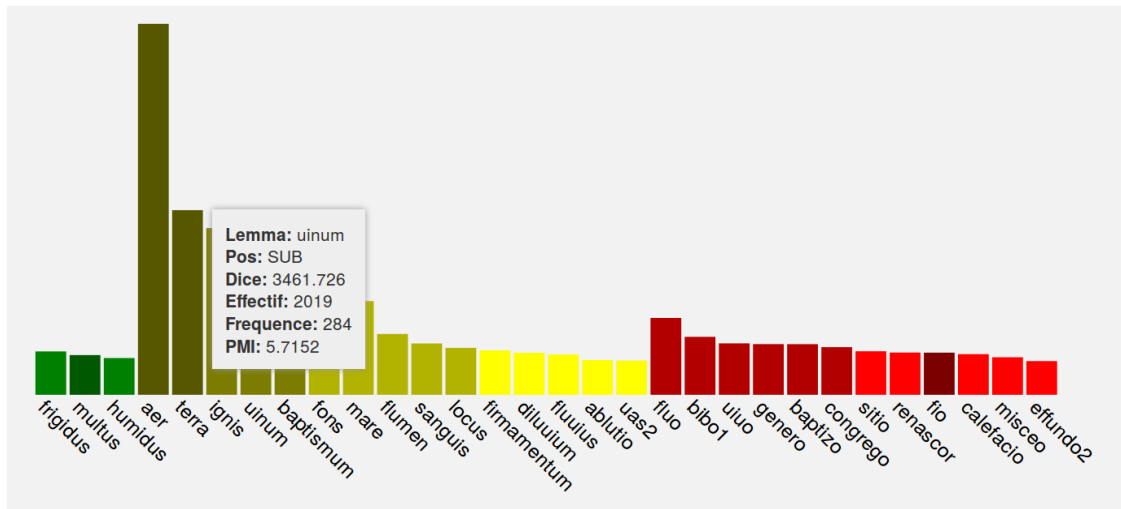


Figure 7: Barplot representing computed co-occurrences of the lemma *aqua* ‘water’ in the *Patrologia Latina* corpus (data fetched from an R session exposed with OpenCPU API; the chart generated with the help of the *d3.js* library).

Hints are, therefore, provided as to how one can interpret the geometric properties of the chart, such as distance between the points, width of the boxplot, and so on. In the case of the co-occurrence barplot, for instance, apart from the information provided in the legend, one may learn that the selected coefficient promotes some type of collocates or that the intensity of the colour of the bar corresponds to the absolute frequency of the co-occurring word in the specific corpus.

Finally, the paradigm of reproducible research is further promoted by enabling users to download complete reports from their queries. Since the reports are available not from the main, but from the extended, view page built with *shiny R* and *OpenCPU*, they will not be explained here in more detail.

4. Conclusion

4.1 Previous research

The integrative, holistic approach to word meaning has been a central idea of philology since its Greek origins, but is also accentuated in modern, cognitive lexical semantics.⁷ The architecture of the application presented in this paper, which is a hybrid tool (Granger, 2012), benefits from research on mashups and content aggregation (Daniel & Matera, 2014). *medialatinitas.eu* makes heavy use of visualisation techniques and other alternative ways of lexicographic data

⁷ Geeraerts (1988; 2009) is one of few researchers to notice the link between the historical-philological and cognitive semantics.

representation and in that respect it builds on recent research into linguistic data visualisation (Theron & Fontanillo, 2013). The notion of reproducible research in scientific computing has recently gained interest, as easy to use R packages such as *knitr* (Xie, 2014) were made available. Although, at the current stage, *medialatinitas.eu* adopts a lightweight, UI-based model of data integration, further work will certainly focus on closer data integration following the LLOD model (Chiarcos et al., 2013). Already in its present form, though, the application exploits existing Semantic Web resources⁸, such as *BabelNet*, *Europeana*, *Getty Thesaurus of Geographical Names*, *Pelagios Project* etc. Integration of the lexicographic resources is promoted and stimulated to an unprecedented extent within the *European Network of e-Lexicography (ENeL)* of which the authors of the present paper are members.

Interest in dictionary applications that follow the aggregator or mashup model seems to be rapidly growing in recent years, with such eminent examples as *Dictionary.com*, *FreeDictionary*, or *Wordnik*.⁹ Regardless of the reasons for it, the aforementioned resources usually offer aggregation of the dictionary content within a user-friendly interface equipped with an efficient query engine. Yet, in the majority of cases, they reuse popular general dictionaries, do not offer any further commentary concerning fetched data and in particular they do not inform about the credibility of the resources. This makes them hardly usable as research tools. The situation is slightly different when one takes into account such aggregators as *Dictionnaire vivant de la langue française*. One will find here juxtaposed the excerpts from renowned lexicographic works (e.g. *TLFi*), but also a selection of corpus and web quotations, as well as charts illustrating the changing frequency of the word. The *DVLF* seems to adopt the same design as that of *Logeion* which, apart from aggregating Latin dictionaries, presents additional information for each lemma based on the *Perseus Digital Library*: a list of authors who frequently use the word and a small selection of co-occurring terms. *medialatinitas.eu* differs from the websites mentioned above not only in the general architecture or scope of the integration, but also in the resources employed, use of encyclopaedic data, implementation of complex statistics, visualisation techniques etc. The same properties distinguish *medialatinitas.eu* also from more general oriented text analysis frameworks such as the *Perseus Digital Library* which collects a large number of lemmatised Latin and Greek texts of Classical Antiquity and Renaissance. Apart from the already mentioned differences, *medialatinitas.eu* is principally lemma-, not text-, oriented; therefore, it is expected to be used as a tool for exhaustive analysis of the vocabulary and not as a reading environment. Moreover, *medialatinitas.eu* employs graphical hints, rich visualisations and mapping; exploits modern academic works rather than older dictionaries or text

⁸ The extensive list of dictionary APIs can be found on the *ProgrammableWeb* website. Accessed at: <http://www.programmableweb.com/category/all/apis?keyword=dictionary>. (23 May 2015)

⁹ According to the *Alexa* website ranking, among 20 most viewed dictionary pages there are at least three resources of this kind: *The FreeDictionary* (the 380th most popular page on the web and the second most popular dictionary after *WordReference*), *SpanishDict* (1827th) and *Your Dictionary* (2116th).

editions; goes beyond in-house resources and uses transparent co-occurrence and frequency measures. Unlike the *Perseus*, it makes a clear distinction between text collection and linguistic corpus and contains a great deal of medieval texts. Contrary to the *Perseus*, which has not seemed to evolve much over the last few years, *medialatinitas.eu* is conceived as a modular, open to extension, lightweight application.

4.2 Further development

The future development of the *medialatinitas.eu* will focus on four main objectives. First, a more appropriate model of linguistic data integration needs to be adopted in order to better deal with the diachronic evolution of Latin vocabulary and with conflicting annotations of linguistic resources. Apart from a faster and more direct search, closer integration should also lead to more sophisticated processing of the user's input. Currently, the search is limited to lemmas only and as such it requires the user to have a rather good knowledge of the Latin language. Secondly, more data should be hosted locally which should help to lower the query and page display times. It is also desirable, because the external APIs (the *BabelNet HTTP API* is one example) often limit the number of queries that can be sent from a single IP address. Thirdly, new widgets should be added and the existing ones need to be constantly improved. The system of graphical hints should be refined and more techniques of data representation and computation should be suggested. Finally, a community of users and content providers needs to be expanded.

5. Acknowledgements

Work on the present paper has been funded by:

- 1) the “Young Researcher Grant” of the Polish Ministry of Science and Higher Education attributed to Krzysztof Nowak by the Institute of Polish Language (Polish Academy of Sciences);
- 2) the “Soutien à la mobilité internationale” grant attributed to Bruno Bon by the Institut des Sciences Humaines et Sociales (Centre National de la Recherche Scientifique).

The paper has greatly benefited from the discussions and workshops of the *ENeL* COST Action in which both authors have had the opportunity to participate.

The authors would also like to thank Renaud Alexandre (IRHT CNRS) for his remarks.

6. References

- Alexa*. Accessed at: <http://www.alexa.com/>. (23 May 2015)
- BabelNet*. Accessed at: <http://babelnet.org/>. (23 May 2015)
- Blatt, F. & Lefèvre, Y. & Monfrin, J. & Dolbeau, F. & Guerreau-Jalabert, A. (eds.). (1957-2011). *Novum Glossarium Mediae Latinitatis*. Copenhagen/Bruxelles/Genève. Available at: <http://www.glossaria.eu/ngml>.
- Bon, B. (2013). Le vocabulaire technique en latin médiéval, entre mythe et réalité. In H. Leithe-Jasper & M.-L. Weber (eds.) *Fachsprache(n) im mittelalterlichen Latein / Technical Language(s) in the Latin Middle Ages / Langage(s) technique(s) au moyen âge latin, Tagungsakten der fünften internationalen mittellateinischen Lexikographentagung (München, 12.-15. September 2012)*. *Archivum Latinitatis Medii Aevi*, 71, pp. 355-375.
- Bon, B. (in print). Histoire et perspectives du ‘Novum Glossarium Mediae Latinitatis’. *Proceedings of the 7th International Conference on Historical Lexicography and Lexicology (ICHLL 2014)*. Bern/New York: Peter Lang.
- Bon, B. & Nowak, K. (2013). WikiLexicographica: Linking Medieval Latin Dictionaries with Semantic MediaWiki. In I. Kosem & J. Kallas & P. Gantar & S. Krek & M. Langements & M. Tuulik (eds.) *Electronic Lexicography in the 21st century, Thinking outside the paper: Proceedings of the eLex 2013 Conference*. Tallinn-Ljubljana: Trojina, Institute for Applied Slovene Studies; Eesti Keele Instituut, pp. 407-420. Available at: <http://eki.ee/elex2013/proceedings>.
- Chang, W. & Cheng, J. & Allaire, JJ. & Xie, Y. & McPherson, J. (2015). *shiny: Web Application Framework for R*. Available at: <http://CRAN.R-project.org/package=shiny>.
- Chiarcos, C. & McCrae, J. & Cimiano, Ph. & Fellbaum, Ch. (2013). Towards Open Data for Linguistics: Linguistic Linked Data. In A. Oltramari & P. Vossen & L. Qin & E. Hovy (eds.) *Theory and Applications of Natural Language Processing*. Berlin/Heidelberg: Springer, pp. 7–25.
- Corpus Thomisticum*. Accessed at: <http://www.corpusthomisticum.org/it/index.age>. (23 May 2015)
- Croatiae Auctores Latini*. Accessed at: <http://www.ffzg.unizg.hr/klafil/croala/>. (23 May 2015)
- d3.js*. Accessed at: <http://d3js.org/>. (23 May 2015)
- Daniel, F. & Matera, M. (2014). *Mashups: concepts, models and architectures*, New York: Springer.
- Desgraupes, B. & Loiseau, S. (2012). *rcqp: Interface to the Corpus Query Protocol*. Available at: <http://CRAN.R-project.org/package=rcqp>.
- Dictionary.com*. Accessed at: <http://dictionary.reference.com/>. (23 May 2015)
- ENeL. European Network of e-Lexicography*. Accessed at: <http://www.elexicography.eu/>. (23 May 2015)
- Europeana*. Accessed at: <http://www.europeana.eu/portal/>. (23 May 2015)
- Evert, S. (2014). Distributional Semantics in R with the wordspace Package. In

- Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin: Dublin City University/Association for Computational Linguistics, pp. 110–114. Available at: <http://anthology.aclweb.org/C/C14/C14-2024.pdf>.
- eXist-db*. Accessed at: <http://www.glossaria.eu/treetagger/>. (23 May 2015)
- Fontes. Corpus of Polish Medieval Latin*. Accessed at: <http://scriptores.pl/fontes>. (23 May 2015)
- Geeraerts, D. (1988). Cognitive Grammar and the History of Lexical Semantics. In B. Rudzka-Ostyn (ed.) *Topics in Cognitive Linguistics*. Amsterdam: John Benjamins Publishing Company, pp. 647-677.
- Geeraerts, D. (2009). *Theories of Lexical Semantics*. Oxford: Oxford University Press.
- Geonames*. Accessed at: <http://www.geonames.org/>. (23 May 2015)
- Getty Thesaurus of Geographic Names*. Accessed at: <http://www.getty.edu/research/tools/vocabularies/tgn/>. (23 May 2015)
- Glossarium mediae et infimae latinitatis*. Accessed at: <http://ducange.enc.sorbonne.fr/>. (23 May 2015)
- Granger, S. (2012). Introduction: Electronic lexicography - from challenge to opportunity. In S. Granger & M. Paquot (eds.). *Electronic lexicography*. Oxford: Oxford University Press, pp. 1–11.
- Guerreau-Jalabert, A. & Bon, B. (2010). Le trésor au Moyen âge: étude lexicale. In L. Burkart & al. (eds.) *Le trésor au Moyen âge*. Firenze: Sismel, pp. 11-31.
- Hardie, A. (2012). CQPweb — combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17 (3), pp. 380–409.
- Internet Archive*. Accessed at: <https://archive.org/>. (23 May 2015)
- Krötzsch, M. & Vrandečić, D. & Völkel, M. & Haller, H. & Studer, R. (2007). Semantic Wikipedia. *Journal of Web Semantics* 5 (4), pp. 251–61.
- Latin Wiktionary*. Accessed at: http://la.wiktionary.org/wiki/Pagina_prima. (23 May 2015)
- Le Dictionnaire vivant de la langue française*. Accessed at: <http://dvlf.uchicago.edu/>. (23 May 2015)
- LMILP: eLexicon Mediae et Infimae Latinitatis Polonorum*. Accessed at: <http://scriptores.pl/elexicon>. (23 May 2015)
- Logeion*. Accessed at: <http://logeion.uchicago.edu/>. (23 May 2015)
- Medialatinitas Github*. Accessed at: <https://github.com/medialatinitas/>. (23 May 2015)
- Nowak, K. (2014). The eLexicon Mediae et Infimae Latinitatis Polonorum, Electronic Dictionary of Polish Medieval Latin. In A. Abel & C. Vettori & N. Ralli (eds.) *The User in Focus: Proceedings of the XVI EuraLex International Congress*. Bolzano/Bozen, pp. 793-806. Available at: <http://euralex2014.eurac.edu>.
- Omnia Project Treetagger*. Accessed at: <http://www.glossaria.eu/treetagger/>. (23 May 2015)
- Ooms, J. (2014). The OpenCPU System: Towards a Universal Interface for Scientific

- Computing through Separation of Concerns. *ArXiv e-prints*. Available at: <http://arxiv.org/pdf/1406.4806v1.pdf>.
- Open Library*. Accessed at: <https://openlibrary.org/>. (23 May 2015)
- Orbis Latinus*. Accessed at: <http://olo.rigeo.net/>. (23 May 2015)
- Pelagios Project*. Accessed at: <http://pelagios.dme.ait.ac.at/api>. (23 May 2015)
- Perseus Digital Library*. Accessed at: <http://www.perseus.tufts.edu/hopper/>. (23 May 2015)
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. Available at: <http://www.R-project.org/>.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester. Available at: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf>.
- Semantic Mediawiki*. Accessed at: <https://semantic-mediawiki.org/>. (23 May 2015)
- SpanishDict*. Accessed at: <http://www.spanishdict.com/>. (23 May 2015)
- The Free Dictionary*. Accessed at: <http://www.thefreedictionary.com/>. (23 May 2015)
- Theron, R. & Fontanillo, L. (2013). Diachronic-Information Visualization in Historical Dictionaries. *Information Visualization* 14 (2), pp. 111–36.
- TLFi: Le Trésor de la Langue Française Informatisé*. Accessed at: <http://atilf.atilf.fr/tlf.htm>. (23 May 2015)
- VIAF*. Accessed at: <http://viaf.org/>. (23 May 2015)
- Wordnik*. Accessed at: <https://wordnik.com/>. (23 May 2015)
- Xie, Y. (2014). knitr: A Comprehensive Tool for Reproducible Research in R. In V. Stodden, F. Leisch, & R. D. Peng (eds.) *Implementing reproducible research*. Boca Raton: Chapman and Hall/CRC, pp. 3-32.
- YourDictionary*. Accessed at: <http://www.yourdictionary.com>. (23 May 2015)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Discovering hidden collocations in a bilingual Spanish–English dictionary

Margarita Alonso Ramos

Universidade da Coruña, Campus de Zapateira s/n, 15071 CORUÑA (SPAIN)

E-mail: lxalonso@udc.es

Abstract

This paper addresses the problem of how to exploit the collocational information included in an online Spanish–English dictionary. Even though collocations are not identified as such in this dictionary, abundant collocational information is used as a means of distinguishing senses. Given that this information is structured in XML markup, the conversion into a bilingual collocation database seems viable in order to obtain the germ of a first Spanish–English collocation dictionary. The concept of collocation used here comes from the Explanatory and Combinatorial Lexicology (Mel’čuk, 2012). In this framework, collocations are understood as recurrent phrases composed of two lexical units, one of which, the *base*, is selected according to its meaning, while the selection of the other, the *collocate*, is determined by the base. The methodology I propose consists of reorganizing the links between words in such a way that the bilingual collocational correspondence is included in the entry for the base. The lexical tool obtained as a result of this reorganization could be exploited for different applications in natural language processing, ranging from machine translation to computer assisted language learning systems.

Keywords: collocations; bilingual dictionary; reusability of lexical resources

1 Introduction

Collocations are usually not especially well treated in bilingual dictionaries, irrespective of the language pair concerned¹. This can be attributed to the fact that bilingual dictionaries tend to put more emphasis on comprehension than on language production, whereas collocations are mainly *idioms of encoding* (Makkai, 1972). Such is the case of the online bilingual *Oxford Spanish–English Dictionary* (OSED <http://www.oxforddictionaries.com/es/traducir/espanol-ingles/>). This dictionary provides answers for an L2 Spanish user who wants to understand the meaning of a word, but gives a more complicated access to an L1 Spanish user aiming to produce a collocation in English. For instance, an L1 Spanish user who wants to know how to say *coger una enfermedad* ‘to catch an illness’ in English will not find the answer in the entry for the noun, but in the entry for the verb, after scrolling through a rather long article in order to find the translation *to catch an illness*. However, if the information

¹ For an overview of the treatment of collocations in the French–Spanish Larousse dictionary, see Alonso Ramos (2001). As far as the collocations in Spanish–English electronic dictionaries, see Corpas Pastor (in press).

is included under the entry for the noun *enfermedad*, access would be easier, because this is the point of departure: the user wants to speak about an illness, the *base* of the collocation.

The concept of collocation used here comes from the Explanatory and Combinatorial Lexicology (Mel'čuk, 2012). This concept does not differ substantially from that used in the *Oxford Collocations Dictionary* (OCD). In this framework, collocations are understood as recurrent phrases composed of two lexical units one of which, the *base*, is selected according to its meaning, while the selection of other, the *collocate*, is determined by the base; in the above example, the collocate *coger* is lexically context dependent on the base *enfermedad*. Both elements of the collocation are selected in different ways. The lexical selection of the base is semantically driven, whereas the selection of the collocate is lexically driven. For instance, if a speaker wants to name the meteorological phenomenon consisting of water falling onto Earth in drops, the selection in English of the noun *rain* is semantically driven, whereas the selection of *heavy* to express that the rain is intense is lexically driven. In contrast, in Spanish or in French, it is not possible to translate *heavy rain* as *lluvia pesada* (Sp.) or *pluie lourde* (Fr.) with the literal translation of *heavy*. The correct translations are *fuerte lluvia* and *forte pluie* lit. 'strong rain'. In English you can say *a strong wind*, but not *a strong rain*, in contrast to Spanish and French, which use the adjective *fuerte* or *fort* in both cases. In this case we have three collocations where the base is a N and the collocate is an Adj.

The grammatical patterns displayed by collocations include also relations between: 1) V and N, the N being the subject or the object of the V; 2) V and Adv; and 3) N and N. See the following table:

Language	Base	Collocate	Gram.Pattern
En.	rain	<i>heavy</i>	N-Adj
Es.	<i>lluvia</i>	<i>fuerte</i>	N-Adj
Fr.	<i>pluie</i>	<i>forte</i>	N-Adj
En.	to rain	<i>cats and dogs</i>	V-Adv
Sp.	<i>llover</i>	<i>a cántaros</i>	V-Adv
Fr.	<i>pleuvoir</i>	<i>des cordes</i>	V-Adv
En.	walk	<i>take</i>	V-Obj
Es.	<i>paseo</i>	<i>dar</i>	V-Obj
Fr.	<i>promenade</i>	<i>faire</i>	V-Obj
En.	secret	<i>lies in</i>	V-Subj
Es.	<i>secreto</i>	<i>estriba en</i>	V-Subj
Fr.	<i>secret</i>	<i>resides dans</i>	V-Subj
En.	chocolate	<i>square</i>	N-N
Es.	<i>chocolate</i>	<i>onza</i>	N-N
Fr.	<i>chocolate</i>	<i>carré</i>	N-N

Table 1: Collocational equivalences following different grammatical patterns

Collocations are especially problematic for production, but not so much for comprehension. If a user of the dictionary needs an adjective expressing 'intense' when speaking about *rain*, he needs to find that *rain* combines with *heavy* in the entry for

rain. This is the normal procedure in collocation dictionaries such as the OCD: to provide the information under the entry of the base; i.e. the noun entry in the case of verbal, nominal and adjectival collocates (*rain*, *secret*, *chocolate*), and the verb entry in the case of adverbial collocates (*to rain*). However, in bilingual dictionaries, even if collocations are included, they are not identified as such, but are presented as a means of distinguishing senses, as I will show in the next section².

This poor arrangement of collocational information can be found in printed bilingual as well as in electronic dictionaries, since the latter, at least those compiled by mainstream publishers, have inherited the problems already present in printed versions. Nevertheless, electronic dictionaries allow us to retrieve hidden information more easily. Almost two decades ago, Fontenelle (1997) built a bilingual collocational database from a bilingual dictionary, although limited by the information contained in a machine readable dictionary. Nowadays, when online dictionaries rely on structured information in XML markup, the idea of “turning” a dictionary into a database is even more compelling.

This paper addresses the problem of how to exploit the collocational information included in the OSED, trying to take the first steps to fill the gap left by the absence of a Spanish–English dictionary of collocations³. As a result of the reorganization of the collocational information, it is possible to obtain lexical data for the germ of a Spanish–English collocation dictionary. These data can be used to compile either a dictionary in the strict sense of the term, or an online lexical tool to be exploited by platforms involved in machine translation or other applications. In the next section, I present how collocations are offered in the OSED in the part Spanish–English, and the different problems of accessibility that this display poses. Section 3 elaborates on a possible strategy to obtain a Spanish–English collocation dictionary by establishing different links between the XML tags. Section 4 focuses on the difficulties that this task presents in relation to the selection of the potential bases and to the selection of collocates in English. Finally, Section 5 draws some conclusions and presents an estimation of the viability of the final output.

2 Treatment of collocational information in the OSED

Putting combinatorial information under the collocate entry (instead of under the base)

² Within the Explanatory and Combinatorial Lexicology, a different conception of bilingual dictionary of collocations is claimed: a bilingual part aimed at selecting the translation equivalent of the base of the collocation, and a monolingual part where the collocation of the target language is described. See Alonso Ramos (2001), Meyer (1990) and Iordanskaja & Mel’čuk (1997).

³ According to Ferrando (2012), the appearance of bilingual dictionaries of collocations is recent. This author mentions 1958 as the date of the publication of an English–Japanese dictionary. Nowadays, it is possible to find some bilingual dictionaries of collocations for other pairs of languages; for example, English–Russian (Benson & Benson, 1993), German–French (Ilgenfritz et al., 1989), German–Italian (Konecny & Autelli, 2014).

makes these entries very long and user-unfriendly to look at. The user has to scroll down long stretches of text in order to find the translation of a collocation, such as *poner atención* ‘to pay attention’, for example. This problem can be solved if the combinatorial information is placed under the entry for the base, in this case, the noun *atención*.

In what follows, I will present the different displays of collocational information in the OSED. There are three main strategies to present collocational information under the collocate entries:

- As an example, sometimes without a translation equivalent. For instance, under the entry for the adverb *encarecidamente*, we can find the collocation *pedir encarecidamente*. Note that no translation equivalent for the adverb is provided. See:
 - 1) a. *le pido encarecidamente que haga lo posible por ayudarlo*
b. I **ur**ge o [formal] **beg** you to do whatever you can to help him
- As an equivalent construction, in a lemmatized form. For instance, under the entry for the adverb *perdidamente*, a translation equivalent, *hopelessly*, is provided and after that, the equivalent constructions are presented. See:
 - 2) a. *estar perdidamente enamorado de alguien*
b. to be **hopelessly in love** with somebody
- By providing the Spanish base in brackets. There are two main distinctions: when the Spanish collocation has the syntactic pattern “N *de* N”, the base is introduced with the preposition *de*. For instance, under the entry for the noun *grano*, different translation equivalents are supplied according to the different bases included in brackets. See:
 - 3) (*de trigo, arroz*) grain; (*de café*) bean; (*de mostaza*) seed

With all other syntactic patterns, the base is included in brackets⁴: a noun in brackets in the entries for adjectives or verbs, on the one hand, and a verb in the entry for adverbs, on the other hand. For instance, under the entry for the adjective *acérrimo*, two translations are provided depending on the noun included in brackets. See:

⁴ Atkins and Rundell (2008: 217) refer to these sense indicators as *collocators*. In the jargon used in OUP, these words in brackets are called *collocates*, following the Sinclairian approach to collocations, whereby both elements of a collocation can be considered collocates, since no directionality in the relation is postulated. I would rather avoid this confusing terminology and will limit the term *collocate* to the lexical unit selected by the base. Corpas Pastor (in press) uses the term *collocational sets* for the series of potential collocates of a given base and/or the series of potential bases for a given collocate. However, in this dictionary only series of bases for a given collocate are displayed in this way.

4) (*partidario/defensor*) staunch; (*enemigo*) bitter

In a similar way, in the entry for the verb *cometer*, we find different translations associated with different nouns. See:

5) (*crimen/delito*) to commit; (*error/falta*) to make; (*pecado*) to commit

In this case, the noun acts as the grammatical object of the verb, but it also can be its grammatical subject. See for instance the entry for the verb *estallar*:

6) (*guerra/revuelta*) to break out; (*t tormenta*) to break

The same procedure is used with collocations following the pattern “V+Adv”, but not systematically. Thus, in the entry for the adverb *bulliciosamente*, we find two translations associated with different verbs. See:

7) (*festejar/protestar*) noisily; (*jugar*) boisterously

However, an adverbial collocate is not always treated in the same way. Sometimes the translation is given without any information about the base; for instance, under the entry for *radicalmente*, only the translation *radically* is found irrespective of the base. The possible explanation is that in Spanish as well as in English this adverbial collocate is selected by the verb *cambiar* or its equivalent in English *to change*. In other cases, a translation equivalent is provided, but different translations appear in the examples. This is the case of the adverb *definitivamente*. See:

8) (*resolver/rechazar*) once and for all

- a. *el texto quedó terminado **definitivamente** en la sesión de ayer*
the text was finalized at yesterday's meeting (no translation)
*the **final** o **definitive** version of the text was drawn up at yesterday's meeting*
- b. *mientras se resuelve **definitivamente** el problema*
*while waiting for a **final** o **definitive** solution to the problem*

None of these strategies have been devised to introduce collocational information, but rather to try to provide semantic cues in order to choose the best translation equivalent in the context of a given base.

Although it is not very frequent, it is also possible to find collocational information under the entry for the bases, especially for collocations following the syntactic pattern “V+N” or “N+V”. This is done by means of examples. For instance, in the entry for the noun *guerra* (‘war’), we find different verbal collocations in the examples. See:

9) a. *nos declararon la guerra*
b. *they declared war on us*

10) a. *están en guerra*

b. *they are at war*

11) a. *cuando estalló la guerra*

b. *when war broke out*

A further source of collocational information is what this dictionary calls *compounds*⁵. For instance, under the *café* entry, we find *café americano* ('large black coffee'), *café con leche* ('white coffee'), *café cortado* ('coffee with a dash of milk'), etc.

In sum, the procedures for including collocational information do not favour the use of the dictionary in terms of production. As stated in the introduction, an L1 Spanish user who wants to know how to say *coger una enfermedad* 'to catch an illness' in English will not find the answer in the entry for the noun, but in the entry for the verb, after scrolling through the rather long entry of *coger* in search of the translation *to catch an illness*. This procedure yields long entries highly difficult to look up. For instance, the entry for the verb *coger* offers 68 translations including senses and examples. With the removal of the translations linked to collocations, the entry would contain only 22 translations and would be, therefore, considerably more accessible. Some headwords functioning only as collocates could remain with the single role of providing part of speech or any other morphological information, but they would not need a whole entry. In the case of the adjective *mortal*, out of the four senses included in this entry, only the first one should remain, since the other three are collocates that should be given in the entry for the nouns in brackets. See:

12. (*ser*) mortal; (*herida*) fatal mortal; (*dosis*) fatal, lethal; (*odio/enemigo*) mortal; (*aburrimiento*) *fue un aburrimiento mortal – it was lethally (inglés norteamericano) o (inglés británico) deadly boring*

The inclusion of the collocational information under the collocate entry does not favour the use of the dictionary for comprehension either, due to the length of the entry and the lack of organisation in the microstructure. If an L1 English user wants to know what *coger* means with *enfermedad*, it is possible to devise an option consisting of launching a query which goes through the whole dictionary. In this way, entries for collocates will only be the result of a query⁶.

After this overview of the treatment of collocational information in the OSED, the main conclusion is that it contains abundant information, but this is not appropriately

⁵ The distinction between compounds and collocations is not trivial. As an illustration, in the Spanish part, the collocation *diente de ajo* ('clove of garlic') is treated as a compound, but in the English part, it is treated as other collocations: under the entry for the collocate *clove*, we find: "(of garlic) *diente*". For an overview of the distinction between compounds and collocations in Spanish, see Alonso Ramos (2009).

⁶ Queries of this kind are already available, although some refinements would be necessary, since now they return not only collocations. See the query for COGER:
<http://www.oxforddictionaries.com/search/spanish-english/?q=COGER&multi=1>.

organized nor displayed. In the next section, I put forward a proposal to build a bilingual collocational database with this information.

3 Taking advantage of implicit collocational information

The fact that the OSED relies on structured information with XML markup makes possible the retrieval of collocational information. Two tags are used to indicate special co-occurrences. These tags are <cs> and <co>. The first one is used to mark the noun acting as the typical subject of a given verb. For example, a typical subject of the verb *contagiarse* ‘to spread’ is the noun *enfermedad*. This information appears in the entry for the verb:

13. CONTAGIARSE_V

[<cs enfermedad> (‘illness’)] to spread, be transmitted

The tag <co> is more frequently used because it covers different relations: verb and object, noun and modifying adjective and finally, verb and adverb.

14. COGER_V

[<co enfermedad> (‘illness’)] to catch; [<co insolación> (‘sunstroke’)] to get

15. GRAVE_{ADJ}

[<co enfermedad> (‘illness’)] serious; [<co voz> (‘voice’)] deep

16. AUTOMATICAMENTE_{ADV}

[<co abrirse/cerrarse (‘to open/ to close’)] automatically

For the collocations following the pattern “N de N” as *grano de café* (‘coffee bean’), a further tag is used: <ind>. This tag is also employed to introduce quasi-synonyms of the headword and, therefore, its automatic exploitation in retrieving collocational information is more complicated. Retrieving the collocations contained in the examples is not trivial either. All examples are tagged with the tag <ex> irrespective of whether or not they include collocations. For instance, under the entry for *pegar*, we find an example including a collocation and another including an idiom:

17)a.<ex no te acerques, que te **pego la gripe**_don't come near me, I'll give you my flu>

b.<ex la verdad es que **la pegamos** con su regalo__we really were dead on o spot on with her gift>

Therefore, this study will be limited to the information which can be more easily exploited automatically, the collocational pairs tagged as <co> and <cs>. After

extracting all the words tagged with <co/cs> and the headwords in the Spanish–English dictionary, I obtained a file with 21,358 pairs consisting of a noun linked with an adjective, a verb, and much less frequently, a verb with an adverb by means of the tags <co/cs>. The nouns appear in singular and in plural, and in some occasions with the article (see the entry for *romper* where we find <un amigo> or <un novio>). After the lemmatisation, there are 3024 words with the tag <co>, 140 of which are verbs and 2880 nouns; and 889 words with the tag <cs>, all of which are nouns, since this tag covers the relation between a noun as grammatical subject and the verb. The intersection between <cs> and <co> is 729 words. The total number of words disregarding the distinction between <co> and <cs> is 3184. This means that the bilingual collocational dictionary could have about 3184 bases for the Spanish part. By way of example, the verb *vivir* (‘to live’), which appears tagged as <co> in the entry for the adverb *despreocupadamente* (‘in a carefree way’), or the noun *zapato* ‘shoe’, which appears tagged as <co> in the entry for the adjective *plano* (‘flat’) and for the verb *acordonar* (‘to lace’) or as <cs> in the entry for the verb *apretar* (‘to be too tight’) are presented in an Excel file in the following way:

vivir	co	despreocupadamente
zapato	co	plano
zapatos	co	acordonar
zapatos	cs	apretar

Table 2: Sample of potential Spanish collocations

From this point, the procedure to be followed in order to build a collocational tool can be synthesized in the following steps:

- 1) Obtaining the English translation related to the tag <cs/co> from the entry for the Spanish headword. For example, in the XML markup entry for ATACAR and in the entry for CONTAGIAR :

18) ATACAR

```
<trg ><cs >virus/enfermedad</cs><tr>to attack</tr></trg>
```

19) CONTAGIAR

```
<trg ><cs >enfermedad</cs><tr>to spread</tr> <tr> to be transmitted</tr></trg>
```

- 2) Aligning the Spanish headword with the English translation in order to have the translation of collocates. For example:

20) ATACAR –ATTACK

21) CONTAGIAR –TO SPREAD, TO BE TRANSMITTED

- 3) Aligning the Spanish and English collocates with the word tagged as <co/cs>. For example:

BASE	SyntRel	COLL-ES	COLL-EN
enfermedad	co	ARRASTRAR	DRAG ON
enfermedad	cs	ATACAR	ATTACK
enfermedad	co	ATAJAR	KEEP IN CHECK
enfermedad	co	BENIGNO	BENIGN

Table 3: Sample of bilingual collocational database

This file can be seen as a germ of a collocational dictionary since we have turned a file of headwords and the values of the tags into what can be a starting point of a bilingual collocational database consisting of a potential base, a syntactic relation and the collocate in both languages⁷. Not all words tagged as <co/cs> are equally productive: out of the total, only 214 are used 20 or more times; among them, there is the noun *persona* ('person'), which appears as <co/cs> in 1261 entries, and the noun *resultado* ('result'), which appears in 24 entries. After an exploration of the data, we can see different cases: highly productive values, as *persona* (1261), *ropa* 'cloth' (129), *animal* 'animal' (109), or *situación* 'situation' (103), and much less productive ones, such as *acceso* 'access' (3), *abanico* 'fan' (2) or *abeja* 'bee' (1). The four former nouns are the most frequently used, but note the difference between the first and the second noun: from 1261 to 129 entries. About 1600 words are used only in one entry. However, between the highly productive words and the very unproductive ones, there is a significant number of words that can become the bases of a collocational entry with 30 or 40 collocates in average. For instance, the noun *enfermedad* ('illness') will contain 42 bilingual collocations; the entry for *acuerdo* ('agreement') will contain about 25 collocates, etc. The entries for these nouns in some collocational dictionaries in the respective languages are much longer (for instance, the entry for *agreement* in OCD contains 56 collocates and the entry for *acuerdo* in DCP 179). However, since a bilingual Spanish–English collocation dictionary does not yet exist, poor entries are

⁷ Note that in this way we do not obtain the translation of the base. This translation should follow another strategy based on semantic grounds to be described in the bilingual dictionary, rather than in the collocational bilingual dictionary. For instance, translating *enfermedad* as *sickness*, *disease* or *illness* does not depend on which are its collocates, but on semantic differences existing between these three English equivalents. See the help note that appears under the CSED dictionary: <http://www.collinsdictionary.com/dictionary/spanish-english/enfermedad?showCookiePolicy=true#footnote 1>.

better than no entries. This file is merely a starting point because it also needs to be filtered. Some distinctions should be established among the words tagged as <co/cs>, which will result in that many pairs will not be part of the collocational tool. In what follows, I will focus on the difficulties or challenges regarding the selection of bases and the selection of the translation of collocates.

4 Filtering Spanish bases and English collocates

In order to arrive at the situation depicted in Table 3, it is necessary, first, to be sure that the relation between the word tagged as <co/cs> and the headword is a collocational relation. Secondly, it is necessary to identify with precision which is the translation equivalent, since, in many cases, the OSDE does not propose any and gives only an example.

4.1 Selection of bases: semantic and lexical tags

As I have pointed out, the purpose of the words in brackets is to help to find the translation of the headwords in combination with these words, not necessarily to give collocational information. For this reason, the words tagged as <co/cs> sometimes represent meanings and sometimes stand for lexical units. In the first case, I will call them *semantic tags*, and in the second case, *lexical tags*. Words are used as semantic tags when their role is to provide a semantic restriction on the nouns that can instantiate the object of a verb⁸. For instance, under the entry of *coger*, we can find:

22) [trabajo ('work')/casa ('house')] to take

The example provided for that sense is:

23) *no puedo coger más clases – I can't take on any more classes*

The nouns <[trabajo/casa]> restrict semantically what could be the object of *coger* when it means 'to accept', but it is possible to use the verb *coger* without these nouns as well, as illustrated with the example: *no puedo coger más clases* ('I can't take on any more classes'). Here we do not have the word *trabajo* ('work'), but the meaning 'trabajo', which can be associated to the meaning of (dar) *clases* 'to teach'.

In contrast, most occurrences of <cs/co> are lexical tags. By lexical tag, I mean the specific word or lexical unit that is combined with the headword. For instance, again in the entry of *coger*, we find:

24) [tren ('train')/autobus ('bus')/taxi] to catch, take

The three nouns in brackets are given to provide the translation of the collocations

⁸ Regarding the role of selectional restrictions and collocations as markers of senses in the dictionaries, see Atkins and Rundell (2008: 302).

resulting from combining *coger* with any of these nouns, as *coger un tren* ('to catch a train').

The problem is that it is not always clear for the user when the tag is used as a semantic restriction, i.e. as a semantic cue to help find the correct meaning of the headword, and when it is used as a lexical tag, i.e. when it specifies the base of a specific collocation which serves to give the translation of this collocation. This ambiguity will make the automatic treatment difficult. For instance, under the entry for the verb *acometer* 'to undertake', we can find:

25) [empresa ('undertaking')/proyecto ('project')] to undertake, tackle

With this information, it is not possible to know with certainty when the word tagged as <co/cs> is representative of a semantic group and when it is only a specific combination. For instance, the noun *tarea* ('task') inherits the collocate *acometer*, because *tarea* can be considered a hyponym of *empresa* or *proyecto*, but it is not explicitly indicated. In the case of semantic tags with this hyperonymic role, it would be useful to study the possibility of automatically deriving collocations by means of some formalism establishing paradigmatic relationships such as Eurowordnet (Vossen, 1998). For instance, if under the entry of the verb *abandonar* ('to abandon'), the noun *actividad* ('activity') is treated as a <co>, all nouns which are considered activities could inherit the collocate *abandonar*: *estudios* ('studies'), *lucha* ('fight'), *curso* ('course'), etc. Therefore, by using some formalism which serves to infer relationships, the initial collocational database could be enriched with new information. However, the formalism should also have the possibility of blocking the inheritance for tags as *persona* 'person' which most of the time represents a semantic restriction and can be eliminated as a potential base to be included in a collocational tool. Thus, in the entry for *abandonar*, it is possible to find *persona* as <co>, as in the following examples of the OSED:

26) a. *abandonó a su familia* – he abandoned ◦ deserted his family
b. *abandonó al bebé en la puerta del hospital*– she abandoned ◦ left the baby at the entrance to the hospital

Nevertheless, the combinations *abandon his family/the baby* are not collocations. Here, the tag <persona> is used to outline the meaning of *abandonar*, but *abandonar* is not a lexical unit selected by the nouns *familia* or *bebé*. Therefore, pairs such as "persona-abandonar" should be eliminated of the collocational database.

In sum, from the initial file, some potential bases should be eliminated, such as *persona* because it is mostly used as a semantic restriction, but some others could be added by using some formalism handling inheritance relationships.

4.2 Selection of the translation of collocates

The policy of the OSED is not very systematic with respect to the way of providing

translation equivalents for collocates. In the ideal situation, we would have a translation equivalent of the collocate with an example in both languages. Thus, under *alcanzar*, we find:

- 27) (acuerdo) to reach
los acuerdos alcanzados en materia de desarme
the agreements reached in the field of disarmament

This information could be easily turned into a bilingual collocation entry:

BASE	SyntRel	COLL-ES	COLL-EN
acuerdo	co	ALCANZAR	TO REACH

Table 4: Bilingual collocation entry

However, the OSED does not always provide a translation equivalent and sometimes gives only an example. In these cases, several possibilities exist:

1. The translation equivalent is recoverable from the example. We have two parallel collocations in the two languages. See the entry for *levantar*:

- 28) (ojos)
*me contestó sin **levantar los ojos** del libro*
*she answered me without looking up o without **lifting her eyes** from her book*

From the example, the following equivalence could be established, through an automatic syntactic parsing:

BASE	SyntRel	COLL-ES	COLL-EN
ojos	co	LEVANTAR	TO LIFT

Table 5: Bilingual collocation entry

2. The translation equivalent represents a different construction in English. This kind of mismatch is very frequent when comparing collocations in different languages (see Mel'čuk & Wanner, 2001). For instance, under the entry of *arder* ('to burn'), we find:

- 29) (estómago)
me arde el estómago
I've got heartburn

In Spanish, the noun *estómago* ('stomach') is the subject of the verb *arder* 'to burn', but the English noun *heartburn* is not the translation of *estómago*: this noun expresses

the meaning expressed by the verb *arder* in Spanish. In this case, the correspondence between both collocations is more difficult to be derived automatically, because the following mapping is wrong:

30) estómago arder to have got

When the meaning of the collocation is distributed between the base and the collocate in different ways in both languages, it is necessary to give the translation of the base (see footnote 6).

Another example, similar to the previous one, could be the mismatch between a light verb construction in Spanish and a single verb in English. In Spanish, it is possible to express the meaning *golpe* ('blow') by the suffix *-azo*, as in *codazo* 'blow given with the elbow'. Any noun created in this way selects a light verb such as *dar* 'give' or in Mexico *arrimar*. In contrast, English uses a single verb *to elbow*. In the entry for *arrimar*, we find:

31) (golpe)
me arrimó un codazo – he elbowed me

In this case, the correspondence is between a collocation and a single verb.

3. In some occasions, lexical gaps prevent a translation. Consequently, the OSED provides a paraphrase of the Spanish collocation. This is the case of *habitación interior*:

32) (habitación/piso) (*with windows facing onto a central staircase or patio*)

5 Conclusion

This paper has described the process of construction of a bilingual collocation database from information already included in an online bilingual dictionary. The approach of reusing existing resources was frequently used in the beginning of the 1990s, but even though nowadays NLP applications tend to rely on big corpora by extracting linguistic knowledge from statistical regularities, I believe that the lexicon is still necessary; especially a lexicon which has been informed by lexicographers. The construction of lexicons from scratch continues to be time-consuming and costly, as in the time when Fontenelle (1997) proposed his collocational database. For this reason I consider that it is worth the effort to reuse the collocational information included in the OSED. This approach of reusing previous lexicographic work can be complemented with current techniques of extracting collocational information from a parallel Spanish–English corpus, especially to provide frequency information. In this way the bilingual collocation dictionary would be corpus-based, not corpus-driven, because the collocations have been established previously in the OSED, not induced from a corpus. Nonetheless, if the final goal is to build a comprehensive bilingual

collocation dictionary, the information extracted from the OSED should be complemented by corpus-induced combinatorial information.

The work presented here only concerned the Spanish–English part of the OSED, but it can be assumed that a similar XML encoding is used in all other bilingual dictionaries from this publisher. Therefore, the potential of bilingual collocational databases is big. As pointed out earlier, the bases and the translations in the database need to be filtered by lexicographers, but according to my estimates this task is not especially time-demanding. In order to obtain a definitive collocational database, technological and lexicographical skills are needed. First, it is necessary to implement a program which automatically establishes the new links between the words involved in collocations. Second, collocational relationships need to be verified by expert lexicographers.

As a possible future line of research, the bilingual collocation database could also be enriched with the *lexical functions* (Mel'čuk, 1996). The apparatus of lexical functions is used in the dictionaries issued from the Explanatory and Combinatorial Lexicology to describe semantically and syntactically collocations:

IncepOper₁(enfermedad) = coger, pillar

IncepOper₁(illness) = to catch

The role of *interlingua* played by the lexical functions could be exploited for search engines involved in machine translation or in information retrieval since they can be used for sense disambiguation. Finally, collocations tagged with lexical functions are unquestionably useful in the field of second language learning.

6 Acknowledgements

The work presented this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the FEDER Funds of the European Commission under the contract number FFI2011-30219-C02-01. I would like to express my gratitude for the help given by the team working in the Global Division of Oxford University Press during my stay in 2014. I would also like to thank Marcos García Salido and Orsolya Vincze for their careful reading and fruitful comments.

7 References

- Alonso Ramos, M. (2001). Construction d'une base de données des collocations bilingue français-espagnol. *Langages*, 143, pp. 5-27.
- Alonso Ramos, M. (2009). Delimitando la intersección entre composición y fraseología. *Lingüística española actual*, 31(2), pp. 5-37.

- Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Benson, M. & Benson, E. (1993). *Russian English Dictionary of Verbal Collocations*, Amsterdam/Philadelphia: John Benjamins.
- Corpas Pastor, G. (in press). Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues". In S. Torner & E. Bernal (eds.). *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches*. Chicago, IL: Ohio State University Press.
- CSED: Collins Spanish-English Dictionary (On-line version). Accessed at: www.collinsdictionary.com/dictionary/spanish-english/ (23 May 2015)
- DCP: Bosque, I. (dir.) (2006). *Diccionario combinatorio práctico del español contemporáneo*. Madrid: SM.
- Ferrando, V. (2012). *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*. PhD Dissertation. Tarragona: Universitat Rovira i Virgili.
- Fontenelle, T. (1997). *Turning a Bilingual Dictionary into a Lexical Semantic Database*. Tübingen: Max Niemeyer Verlag.
- Ilgenfritz, P., Stephan-Gabinel, N. & Schneider, G. (1989). *Langenscheidts Kontextwörterbuch Französisch-Deutsch*, Berlin/München: Langenscheidt.
- Iordanskaja L. & Mel'čuk, I. (1997). Le corps humain en russe et en français. Vers un Dictionnaire explicatif et combinatoire bilingue. *Cahiers de Lexicologie*, 70(1), pp. 103-135.
- Konecny, C. & Autelli, E. (2014). *Kollokationen Italienisch-Deutsch*. Hamburg: Helmut Buske.
- Makkai, A. (1972) *Idioms Structure in English*. The Hague: Mouton.
- Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon". In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: John Benjamins, pp. 37-102
- Mel'čuk, I. (2012). Phraseology in the Language, in the Dictionary, and in the Computer. *Yearbook of Phraseology*, 3 (1), pp. 31–56.
- Mel'čuk, I. & Wanner, L. (2001). Towards a Lexicographic Approach to Lexical Transfer in Machine Translation (Illustrated by the German-Russian Language Pair). *Machine Translation*, 16(1), pp. 21-87.
- Meyer I. (1990). Interlingual Meaning-Text Lexicography: Towards a New Type of Dictionary for Translation. In J. Steele (éd.), *Meaning-Text Theory: Linguistics, Lexicography, and Applications*, Ottawa: University of Ottawa, Ottawa, pp. 175-270.
- OCDSE: *Oxford Collocations Dictionary for Students of English*. (2009). 2nd edition. Oxford: Oxford University Press.
- OSD: Oxford Spanish-English Dictionary (On-line version). Accessed at www.oxforddictionaries.com/translate/spanish-english/ (23 May 2015).

Vossen. P. (1998). Eurowordnet. A multilingual database with lexical semantic networks for European Languages. Dordrecht: Kluwer Academic Publishers.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Management and exploitation of conceptual data and information in technical termbases: the electrotechnical vocabulary

Laura Giacomini

Department of Translation and Interpreting, University of Heidelberg, Plöck 57a, 69117
Heidelberg (Germany)
E-mail: laura.giacomini@iued.uni-heidelberg.de

Abstract

This paper addresses the lexicographic challenges related to the management and exploitation of conceptual data and information by examining the example of electrotechnical vocabulary. Four online tools with different source, typology and reference language will be presented and compared from the point of view of the user's needs. By focusing first on the conceptualization level of the underlying database and then taking into account how this interfaces with the terminological component, the paper will progressively provide specific insights into data availability, ease of access and consistency, and will hint at possible ways to improve conceptual representation in LSP e-lexicography.

Keywords: e-lexicography; term; concept; termbase; technical domain

1. Introduction

In order to evaluate the potential of e-lexicographic tools concerning the quality of data representation for the end user, usability tests are required to highlight the level of effectiveness, efficiency and user satisfaction that a specific tool can achieve (Heid, 2012; Giacomini, 2014). In particular, a satisfactory level of effectiveness, i.e. degree of task completion, and efficiency, i.e. the amount of time needed to perform a task, largely rely upon formal and content-related coherence of the underlying termbases.

With reference to current database and knowledge management theories (Alwert & Hoffmann, 2003; Halpin & Morgan, 2010; Pratt & Adamski, 2011), data are defined as raw lexical and conceptual items, which can be classified, condensed and contextualized to obtain conceptual and terminological information. This implies that two dichotomies need to be taken into account at the same time in this study: on the one hand, the dichotomy between conceptual and terminological items, and on the other hand, that between cognitively unprocessed data and information conveyed by data during consultation.

Starting from a quite comprehensive definition of e-lexicographic tools as information tools of a lexicographic kind (Leroyer, 2012), which can be referred to as, for instance,

dictionaries, glossaries or wiki tools, this paper addresses the challenges related to the treatment of conceptual data and information in terminology databases that serve as a lexicographic basis. The final goal is to explore the extent to which structured and consistent management of conceptual items goes hand in hand with their direct exploitation by dictionary users, which results in increased effectiveness and efficiency of the tool. The paper aims to illustrate this topic through an examination of electrotechnical vocabulary. Section 2 describes the set of example resources that have been taken into consideration in this study, the ideal user that is addressed in the analysis and the method employed in the study. A comparative analysis of the different resources and its results are presented in Section 3, while Section 4 contains some final observations.

2. Representative tools, the applied method and the addressed user

Online resources with different distributions of source and reference language have been selected to make a comparison from the point of view of a user’s needs (Tarp, 2008; Koplenig, 2011). This selection is not intended to be exhaustive and should be seen as a way of exemplifying the procedure and drawing first conclusions on the correlation between management and exploitation of conceptual items from a lexicographic point of view. The representativeness of these tools lies in the fact that they exhibit some of the most widespread LSP e-lexicographic structures and cover the prevalent types of sources consulted by translators as the user group addressed in this study. By considering the Function Theory of lexicography (cf. Tarp, 2008) as the theoretical basis of this analysis, the ideal target user group of these lexicographic resources has been, in fact, identified as professional translators performing a passive translation task or producing a specialised text in their native language (Mayer, 1998). The concrete usage situations are primarily of a communicative kind, but consultation for cognitive purposes, i.e. for knowledge acquisition (Tarp, 2008), is also contemplated, especially in the case of monolingual tools. Table 1 illustrates the combination of source and languages in each of the tools. Specific content-related and formal features on the macrostructural and microstructural level will be introduced and discussed in the next section.

TOOL	SOURCE	LANGUAGE(S)
International Electrotechnical Vocabulary (IEV, or Electropedia) and IEC Glossary	standardization organization	multilingual
IATE database (Electronics and electrical engineering section)	institution: EU	multilingual
Open Energy Information Glossary (OpenEI)	open source wiki	English
Electrical Glossary (Fluke Electronics)	company-internal terminology	English

Table 1: The selected tools

This choice allows for a broad assessment of knowledge representation and of the extent to which its formalisms affect the consultation performance in terms of task completion and time investment (cf. Schärfe et al., 2006).

In order to reach this goal a three-step procedure has been applied, which will now be introduced. By focusing first on the conceptualization level and then taking into account how this interfaces with the terminological component, the paper will progressively provide specific insights into

- a) conceptual structure availability (presence and depth/granularity of conceptual networks, e.g. the one including ELECTRIC CURRENT, ALTERNATING CURRENT, DIRECT CURRENT, etc.)¹,
- b) ease of access (degree of transparency of term-related concepts, e.g. to what extent the user can retrieve, view and consult the conceptual network of ELECTRIC CURRENT by directly accessing the conceptual layer of the database or while performing a term search) and
- c) consistency (regular and logical correspondence between concepts and terminological designations, e.g. between ELECTRIC CURRENT and related simple terms, multiword terms, abbreviations, acronyms and their variants: *ampere*, *ampère*, *amp*, *A*, *ampere-hour*, *ampere-hour meter*, etc.).

3. Analysing the management and exploitation of conceptual data in the selected resources

3.1 Conceptual structure availability and properties

The availability of conceptual structures has been assessed on the grounds of the macrostructural properties of the selected tools and is summarised in Table 2, with special focus on the example of the concept alternating current. Not all principles of terminology management proposed by the German terminology association DTT (2014) can be properly evaluated by taking into consideration the only surface features of the lexicographic resources. This paper concentrates on the depth of the conceptual structures, the available relations involving the lower conceptual level (the bottom level, which can be taken into account independently of the typology of the superordinate structures) and the presence of conceptual networks as the criteria that apply particularly to the treatment of the conceptual layer. An important, initial assumption is that correlations between concepts and terms can be multivocal in both directions: a concept may be verbalized by means of more than one term, and a term may designate more than one concept.

¹ Concepts are written in small caps, terms in italics.

TOOL	DEPTH OF CONCEPTUAL STRUCTURES	CONCEPTUAL RELATIONS AT THE LOWER LEVEL
IEV Online	subject area > section > specific concept / term: e.g. CIRCUIT THEORY > GENERAL > ALTERNATING CURRENT / <i>alternating current</i>	a) multivocal lower – higher level: e.g. ALTERNATING CURRENT < CIRCUIT THEORY / ROTATING MACHINERY / INDUSTRIAL ELECTROHEAT / ... b) no relation lower – lower level: e.g. ALTERNATING CURRENT ? DIRECT CURRENT
IATE	domain > specific concept / term: e.g. ELECTRONICS AND ELECTRICAL ENGINEERING > ALTERNATING CURRENT / <i>alternating current</i>	a) multivocal lower – higher level: e.g. ALTERNATING CURRENT < ELECTRONICS AND ELECTRICAL ENGINEERING / ELECTRICAL INDUSTRY / TOWN PLANNING b) multivocal lower – lower level: e.g. ALTERNATING CURRENT – DIRECT CURRENT (antonym) / PULSATING CURRENT (related)
OpenEI / Fluke Glossary	no conceptual structure: e.g. Ø > ALTERNATING CURRENT / <i>alternating current</i>	a) no relation lower – higher level b) univocal lower – lower level: e.g. ACTIVE POWER – AMPÈRE-HOUR (Fluke Glossary)

Table 2: Conceptual structures availability and properties

IEV Online, or Electropedia, is an electrical and electronic terminology database comprising around 20,000 terms. Created by the standardization organization IEC (International Electrotechnical Commission), it has definitions provided in English and French, and equivalents in several other languages. From a macrostructural viewpoint, IEV Online can be classified as a resource with a complex, not fully developed organizational system and a primarily systematic arrangement. The prevalent conceptual criterion in IEV Online is a classification in which the main subject areas of the electrotechnical field are recorded and further subdivided into more specific sections, eventually leading to final-level concepts and terms. The tool displays a multivocal directionality of conceptual relations: a final-level concept may be attributed to more than one superordinate section or subject area, e.g. ALTERNATING CURRENT can refer to six different subject areas, taking the form of 11 different terminological entries, among which *alternating current machine* or *capacitor fed alternating current track circuit* (the same term is never recorded under more than one category). However, despite the clear hierarchical categorization, no definite relations (e.g. co-hyponymy) can be identified among the large number of final-level items. For instance, in the structure Area: CIRCUIT THEORY > Section: GENERAL no relation can be established between CIRCUIT : ELECTRIC CIRCUIT : MAGNETIC CIRCUIT : ... CIRCUIT or between DIRECT CURRENT : ALTERNATING CURRENT : ACTIVE

CURRENT : INDUCTIVE CURRENT : ... CURRENT. This aspect can be traced back to insufficient granularity in the lower level of the conceptual structure.

The second resource, the multilingual IATE database, is a comprehensive institutional resource recording terms from a broad range of disciplines, including an electronics and electrical engineering section (domain no. 6826). This domain directly includes a number of concepts/terms with no further groupings (further categories can only be found in the external EuroVoc at <http://eurovoc.europa.eu>), and, like in the case of IEV Online, final-level items may belong to more than one domain (multivocal relations). Different from the previous resource, the IATE database contains relations between items at the lower conceptual level and labels them accordingly (e.g. antonyms). Although its macrostructure can be defined as fully systematic, IATE's degree of granularity and its conceptual development are clearly unsatisfactory, and no conceptual network is available. The fact that the database covers several different domains may be one of the main causes.

The Open Energy Information Glossary (OpenEI) is an open source monolingual wiki that records data related to the topic of energy. This resource shows a simple and form-determined (i.e. alphabetical) macrostructure and avoids a conceptual organizational system, so that final-level conceptual relations are also not available. If relevant, concepts/terms only seem to be hypertextually linked to each other by means of entry-internal, non-systematic lists labelled "Related Terms" (univocal relations). The same macrostructural type and an analogous kind of approach to conceptually related items can be found in the last resource, the monolingual Electrical Glossary provided by Fluke Electronics, an example of a lexicographic resource reflecting a company-specific view of the domain and its terminology.

3.2 Ease of access to conceptual data

As it has been observed by Giacomini (2015), macrostructural features of LSP e-lexicographic tools, in particular the presence of systematic relations among concepts/terms, may be generally less discernible to the user, since they can often be only partially noticed during consultation. In this section, the actual access to conceptual relations via the microstructure and/or the conceptual structure will be taken into consideration (Table 3).

It has emerged in the previous section that none of the tools contain a structured ontology but, at the most, a conceptual structure based on a closed set of subjects. This structure is available in IEV Online and IATE. The former allows for an external access to its subject areas, which are listed in a hierarchical structure linked to further category and detail pages. The user can choose between performing a term search via a search mask and browsing the available subject areas. In the second case, the search for a specific term may require a longer amount of time, unless the user is already familiar with the recorded subdisciplines and also has an operational knowledge of the

previously mentioned categorization criteria. On the one hand, a term is never recorded under more than one category, whereas combinations of a term may appear under different categorical labels (*alternating current* itself only belongs to the general section of the CIRCUIT THEORY area); on the other hand a concept may be related to different areas (ALTERNATING CURRENT is related to CIRCUIT THEORY / ROTATING MACHINERY / GENERATION, TRANSMISSION AND DISTRIBUTION OF ELECTRICITY / SWITCHING AND SIGNALLING IN TELECOMMUNICATIONS / SIGNALLING AND SECURITY APPARATUS FOR RAILWAYS / INDUSTRIAL ELECTROHEAT, cf. Table 2).

TOOL	ACCESS TO CONCEPTUAL RELATIONS VIA THE CONCEPTUAL STRUCTURE	ACCESS TO CONCEPTUAL RELATIONS VIA THE MICROSTRUCTURE
IEV Online	direct access	direct access; non-specified relations; totally accessible; systematic; hyperlinked
IATE	no access, only filtering function	direct access; specified relations; totally accessible; systematic; not hyperlinked
OpenEI	∅	indirect access via related terms; totally accessible; systematic; hyperlinked
Fluke Glossary	∅	indirect access via related terms; partially accessible; non-systematic; not hyperlinked

Table 3: Access to conceptual data

IATE's users can only employ available domain categories as search filters during term search and cannot directly consult these categories. This results in a necessity for the user to perform rather specific queries and the impossibility to retrieve all terms belonging to the same domain. Moreover, besides the absence of a subdomain categorization (cf. Table 2), only one domain can be selected as a filter at once, which makes it quite difficult for the user to identify terminological cross-references between different disciplines. In comparison to IEV Online, IATE does not clearly highlight the terms containing the search term itself, so the user is often compelled to analyse long lists of search results in order to look for relevant conceptual/terminological combinations. The search for *alternating current*, for instance, produces, among others, results such as *alternating current generation system*, *single-phase alternating current*, *indirect alternating current converter*, *alternating current supply*, etc., which are indistinctly put together.

From the microstructural perspective, the first two resources share another important feature, which is direct access to conceptual relations: the IEV Online microstructure offers non-specified relations, whereas the IATE entries name the multivocal lower–lower level relations already mentioned in Table 2, even though this does not

seem to happen systematically. These kinds of properties are also present in the other two resources, but, as they do not rely on an underlying conceptual structure, they are far less developed. For this reason, the users of OpenEI and the Fluke Glossary can only retrieve indirect information regarding conceptual relations through article-internal cross-references to other terms. Table 3 highlights the following characteristics of the microstructural access to conceptual relations: availability of a direct vs. indirect access (i.e. access to conceptual items vs. access via terminological items), total/partial access, access systematicity, presence of specified relations (i.e. relations which have been attributed a type), access through hyperlinked data. The results show different possible combinations of these characteristics, which can be summarized and evaluated in the following categorisation proposal:

1) access via the conceptual structure (if such structure is available):

1.1) direct access (direct access enables the user to retrieve information concerning the conceptual relations independently of the consultation of the terminological layer, and can thus actively support consultation for cognitive purposes)

1.2) no access

2) access via the microstructure:

2.1) type of access:

2.1.1) no access

2.1.2) direct access (the user can directly access conceptual information while looking up a term. If this condition is given, the type of conceptual relation between the term and other concepts/terms can either be specified or not)

2.1.2.1) specified relations (as a result of this feature, users should be able to look for single types of relations and identify clusters of concepts such as synonyms, hyponyms, troponyms etc.)

2.1.2.2) non-specified relations

2.1.3) indirect access via terminological items

Moreover, 2.1.2 (direct access) and 2.1.3 (indirect access) can be described in terms of:
2.2) total/partial access (the user can access the same conceptual information by looking up any involved term or only some of the involved terms)

2.3) access systematicity (access to conceptual information is coherently implemented for all terms)

2.4) availability of hyperlinked data

3.3 Consistency of concept-term correspondences

This section concentrates on the degree of consistency in the correspondences between concepts and terms in the analysed e-lexicographic resources, an aspect that is closely related to their mediostructural properties. In the ideal case, a resource should contemplate a coherent and recognizable mediostructure, independent of the depth of its conceptual structure and of the access to conceptual relations it provides. This paper leans on a conception of concepts and terms according to ISO 1087-1:2000. This norm, dealing with the vocabulary of terminology, defines a concept as a unit of

knowledge created by a unique combination of characteristics, whereas a term is a verbal designation of a general concept in a specific subject field. Terms can be instances of different kinds, such as simple terms, complex terms (e.g. collocations or compounds), symbols and formulae.

In order to test the consistency of the selected resources despite their heterogeneity, the example of the general concept ELECTRIC CURRENT and of related terms will be taken into consideration. Specific tasks have been accomplished that aim to assess consistency:

- search for the terms related to ELECTRIC CURRENT by accessing the conceptual structure
- search for the terms related to ELECTRIC CURRENT by looking up *electric current*
- is a correlation between ELECTRIC CURRENT and the corresponding terms coherently represented?
- if yes, is it present in both directions, i.e. when moving from the term to the concept and vice versa?

TOOL	CONCEPTUAL RELATIONS ACCORDING TO THE TYPE OF QUERY & DEGREE OF TERMINOLOGICAL COVERAGE		CONSISTENT REFERENCES CONCEPTS-TERMS
IEV Online	search by subject area: leads to different relations (hyponymy, meronymy etc.)	it is not possible to identify ELECTRIC CURRENT if not by browsing the content of all or selected subject areas	no references
	search by term: leads to hyponyms	<i>electric current</i> includes 5 hyponymical terms belonging to 3 subject areas	yes; the hypernym is referenced to the hyponyms, but not vice versa
IATE	search by term: leads to hyponyms	<i>electric current</i> includes 3 hyponymical terms and 1 synonymous term (<i>current</i>)	yes; the hypernym is referenced to the hyponyms, but not vice versa
OpenEI	search by term: leads to a specific term only	<i>electric current</i> is related to 6 other terms (non-specified relations)	no: consistency in cross-references is not always given
Fluke Glossary	search by term: leads to a specific term only	<i>electric current</i> is not among the glossary terms, but it is referred to in the article of <i>Ampère</i>	no consistency in cross-references

Table 4: Consistency of conceptual data

Table 4 summarizes the results of the analysis by presenting information concerning the types of relations the user can find according to the kind of query he/she performs, the corresponding degree of terminological coverage and a general evaluation of the consistency of mediostructural correlations between concepts and terms.

As these results show, taxonomical relations are better captured by a cross-referencing system, and are therefore likely to be rendered in a coherent way. Other types of relations tend to be widely underrepresented even in resources with a well-developed conceptual structure like IEV Online, which hints at the fact that underlying termbases do not reach a sufficiently deep level of ontological coverage when it comes to the identification of all available relations among concepts.

Semantic word-families are another interesting aspect in terminology which seems to be largely neglected in these resources. By observing the term *ampere*, which is connected to the concept ELECTRIC CURRENT through the relation “unit of measurement” (ampere is the unit of electric current according to SI) and its semantic word-family includes both orthographic variants (*Ampère*), abbreviations (*A*, *amp*) and compounds (*ampere-hour*, *ampere-hour meter*), it becomes clear that all of these terms should be systematically cross-referenced to each other. However, none of the selected e-lexicographic tools offers a satisfying and coherent representation of this cluster of terms: *ampere* is always referenced to the concept ELECTRIC CURRENT through the terminological definition, but the other terms a) are only partially available and b) are not coherently cross-referenced to each other (cf. Table 5).

TOOL	CROSS-REFERENCES
IEV Online	ampere > ELECTRIC CURRENT ampere > A, volt-ampere meter, ampere-hour meter, volt-ampere-hour meter ampere <> ampere-turns
IATE	ampere > ELECTRIC CURRENT ampere > A, amp, ampere-turn/ampere turn, ampere-hour capacity, ampere hour/ampere-hour, amperes per metre, kiloVolt Ampere, metre-kilogram-second ampere, volt-ampere-reactive hour meter, volt-ampere, ...
OpenEI	ampere > ELECTRIC CURRENT ampere > amp
Fluke Glossary	Ampère > ELECTRIC CURRENT Ampère > A Ampère-Hour > Ampère

Table 5: Treatment of semantic word-families: *ampere*

Table 5 reveals an overall lack of data concerning the relations among these concepts/terms. IEV Online only records the compounds of *ampere* that are relevant from a subject-area perspective. IATE lists a larger number of compounds but without

clustering them into conceptually coherent groups and offering no opportunity to proceed in the opposite direction, i.e. from a compound to *ampere* (IATE always moves from a base to its compounds/collocations and not vice versa), which is made possible by IEV Online (cf. *ampere* <> *ampere-turns*), although not systematically. OpenEI and the Fluke Glossary display the most lacking treatment of this semantic word-family. The former only refers to the abbreviation *amp* and does not cover other related terms belonging to the family: this may be seen as a partial but not incoherent representation, since this resource does not specifically focus on the topic of electrical engineering. The Fluke Glossary, on the contrary, has an approach to the treatment of these terms which is clearly partial and inconsistent.

4. Observations and outlook

Assessment of the management and exploitation of conceptual data in the selected resources has pointed out important differences in their approach. The OpenEI and Fluke glossaries do not develop a conceptual structure, which results in a partial and often incoherent conceptual representation. This can be a great disadvantage to users, particularly non-experts. IEV Online and IATE offer more advanced solutions: they are both based on an underlying conceptual structure and offer a much larger amount of data. Evaluation of conceptual data and information carried out in Section 3 defines the minimum requirements a termbase intended for LSP e-lexicographic resources should comply with:

- A) Conceptual structure availability and properties:
 - sufficient (multilevel) depth of conceptual structures
 - multivocal relations (lower–higher level and lower–lower level)
- B) Ease of access to conceptual data:
 - direct access via the conceptual structure, with specified relations
 - direct access via the microstructure, with specified relations
- C) Consistency of concept-term correspondences:
 - consistency of cross-references in the search by concept
 - consistency of cross-references in the search by term.

IEV Online and IATE do not satisfy all these conditions but combine only some of them. The main drawback of these resources is the absence of a conceptual structure in the form of an ontology. Unfortunately, structures of subject-groups can be systematic and coherent but fail to cover the entire range of semantic relations among the concepts of a discipline. What a subject-group structure does not record is partly compensated for by means of the terminological definition (cf. the example of *ampere* and of its relation to the concept ELECTRIC CURRENT). However, a domain-specific ontology would ensure a definitely higher degree of data accessibility and data coherence. Conceptual structures should account for the existence of different types of relations, such as

- 1) semantic fields (broadly intended as clusters of concepts displaying, for instance, taxonomic, meronomic, troponymic, or functional relations) and
- 2) semantic word-families (clusters of concepts/terms with morphological affinity, including abbreviations, orthographic variants and word combinations). The two groups can overlap, but distinctive features should also be taken into consideration to guarantee a possibly comprehensive conceptual representation.

By implementing a method for delivering a detailed description of conceptual data representation in LSP e-lexicographic resources, this study has revealed a series of essential properties and their most effective and efficient combinations. At the same time, new ways to improve terminological representation and exploitation in termbases for lexicographic purposes should be looked for by conducting further investigations on other resources and subdomains, as well as dictionary consultation tests according to specific usage situations (e.g. text production, text reception, active and passive translation).

5. References

- Alwert, K. & Hoffmann, I. (2003). "Knowledge Management Tools". In K. Mertins & P. Heisig & H. Vorbeck (eds) *Knowledge Management. Concepts and Best Practices*. Berlin/Heidelberg/New York: pp. 114-150.
- Giacomini, L. (2014). "Testing user interaction with LSP e-lexicographic tools: A case study on active translation of environmental terms". In Proceedings of Konvens, Hildesheim 2014, October 8-10.
- Giacomini, L. (2015). Macrostructural properties and access structures in LSP e-dictionaries for translation: the technical domain. *Lexicographica* 31.2015 (forthcoming).
- Halpin, T. & Morgan, T. (2010). *Information Modeling and Relational Databases*. Burlington, MA.
- Heid, U. (2012). "Electronic Dictionaries as Tools: Toward an Assessment of Usability". In *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London: pp. 287-304.
- Koplenig, A. (2011). "Understanding How Users Evaluate Innovative features of Online Dictionaries – An Experimental Approach". In Proceedings of eLex 2011: pp. 147-150.
- Leroyer, P. (2012). "Change of Paradigm: From Linguistics to Information Science and from Dictionaries to Lexicographic Information Tools". In *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. Continuum, London: pp. 121-140.
- Mayer, F. (1998): *Eintragsmodelle für terminologische Datenbanken. Ein Beitrag zur übersetzungsorientierten Terminographie*. Tübingen.
- Pratt, P. & Adamski, J. (2011). *Concepts of Database Management*. Boston, MA.
- Schärfe, H. & Hitzler, P. & Øhrstrøm, P. (eds.) (2006). *Conceptual Structures: Inspiration and Application. 14th International Conference on Conceptual Structures ICCS 2006*. Aalborg.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge*.
Tübingen.

Websites:

Fluke Electrical Glossary. Accessed at:

<http://www.fluke.com/fluke/inen/solutions/electrical/electrical%20glossary> (15
May 2015)

IATE. Accessed at: <http://iate.europa.eu>. (15 May 2015)

IEV Online. Accessed at: <http://www.electropedia.org>. (15 May 2015)

OpenEI. Accessed at: <http://en.openei.org/wiki/Glossary>. (15 May 2015)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0
International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography

Peter Meyer¹

¹ Institut für Deutsche Sprache, Mannheim
E-mail: meyer@ids-mannheim.de

Abstract

This paper presents a dictionary writing system developed at the Institute for the German Language in Mannheim (IDS) for an ongoing international lexicographical project that traces the way of German loanwords in the East Slavic languages Russian, Belarusian and Ukrainian that were possibly borrowed via Polish. The results will be published in the *Lehnwortportal Deutsch* (LWP, lwp.ids-mannheim.de), a web portal for loanword dictionaries with German as the common donor language. The system described here is currently in use for excerpting data from a large range of historical and contemporary East Slavic monolingual dictionaries. The paper focuses on the tools that help in merging excerpts that are etymologically related to one and the same Polish etymon. The merging process involves eliminating redundancies and inconsistencies and, above all, mapping word senses of excerpted entries onto a common cross-language set of ‘metasenses’. This mapping may involve literally hundreds of excerpted East Slavic word senses, including quotations, for one ‘underlying’ Polish etymon.

Keywords: dictionary writing system; historical lexicography; word senses

1. Introduction

An ongoing international lexicographical project¹ of the Institute of Slavic Studies at the University of Oldenburg and the Institute for the German Language (IDS, Mannheim) traces the way of German loanwords in Polish – as recorded in the *Dictionary of German Loanwords in Standard and Written Polish* (DGLP) – into the East Slavic languages Russian, Belarusian, and Ukrainian. The results will be published in three separate but interlinked dictionaries alongside the already republished DGLP in the *Lehnwortportal Deutsch* (LWP), a web portal for loanword dictionaries with German as the common donor language.² This endeavor draws on a rich Slavic tradition of historical lexicography; a wealth of partially unpublished

¹ The project is funded by the German Research Foundation (DFG); it started in mid-2013 and will be completed in 2017.

² The LWP aims to provide a uniform access layer to a growing number of heterogeneous lexicographical resources, allowing queries for arbitrarily complex borrowing constellations across all component dictionaries (Meyer, 2013), even in chains of borrowing processes (Meyer, 2014a).

dictionary material is currently being excerpted and analyzed both in Oldenburg and at the editorial offices of those dictionaries that are still works in progress, while the IT architecture development and the integration of the resulting dictionaries with an estimated total of more than 1900 new entries into the LWP is carried out in Mannheim.

Section 2 of the present paper will give a brief sketch of the project's main tasks, the lexicographical process and the resources involved. The focus of the paper is on *wdlpOst*, the dictionary writing system developed at the IDS Mannheim for the specific purposes of the project. A high-level overview of the *wdlpOst* system, its functionality and its data architecture is given in section 3. Section 4 focuses on one of the central advanced features of the system, an editing tool which allows lexicographers to map the widely differing word sense distinctions found in the various East Slavic sources for corresponding headwords onto a common semantic scheme. The closing section 5 gives a brief overview of some further tools of the dictionary writing system.

2. Lexicographical Process: Resources and Workflow

The project's main task consists of extracting and processing lexicographical information on potential Polish-mediated German loanwords in East Slavic from a range of (at present) 15 East Slavic *source dictionaries*, i.e. historical and contemporary monolingual dictionaries of Russian, Ukrainian, and Belarusian. In view of the wealth of data already collected through a number of long-term lexicographical projects and documented in multi-volume dictionaries, no attempt is made to collect new corpus material. The excerpted lexicographical data covers a time span from the eleventh century until the present day and reflects a wide range of lexicographical traditions and approaches. In most cases, the source dictionaries do not indicate the status of words as loans or inherited. Therefore, the excerpted entries must be evaluated in a cross-linguistic perspective in order to formulate hypotheses of possible borrowing pathways. The excerpts are then used to compile entries of the three *target dictionaries* for 'indirect' German loanwords in East Slavic languages that constitute the project's primary scientific outcome and will form part of the loanword dictionary portal LWP.

The project's lexicographical work is directed and mainly carried out at the University of Oldenburg; unpublished parts of four multi-volume historical dictionaries (SRJa11-17, SRJa18, HSBM, SUM16-17) are excerpted from paper slips at the *editorial offices* of these dictionary projects in Moscow (for the SRJa11-17), Saint Petersburg (for the SRJa18), Minsk (for the HSBM), and Lviv (for the SUM16-17).

The project does not intend to perform an exhaustive search for possible German loanwords in the source dictionaries, as this simply would not have been a manageable task for a small three-year project. Instead, the point of departure is defined by the

German loanwords in Polish that are listed in the authoritative dictionary on this topic, the DGLP, whose more than 2400 entries are explicitly restricted to German etyma inherited from Germanic – thus in particular excluding German etyma of Latin or Greek origin – and borrowed directly into written and Standard Polish. The lexicographical process can roughly be divided into four overlapping stages:

- 1. *Exploratory phase* (Oldenburg, editorial offices): All source dictionaries are systematically scanned for *source entries* whose headwords are possible East Slavic cognates of Polish loanwords in the DGLP (including variants and derivatives of these Polish loans). These source entries are tabulated with some basic information in simple spreadsheet tables. No decisions on borrowing pathways, loanword status, etc. are made at this point. This phase is finished and has yielded a total of more than 9000 source entries.
- 2. *Excerption phase* (Oldenburg, editorial offices): Each source entry listed in the spreadsheet tables is turned into an initially almost empty *excerpt* represented as an XML document and stored either in a central database located on an IDS server, or, in the case of the editorial offices where a reliable Internet connection is not always available, in a local computer directory with the option to make periodic backups on the server. The excerpt documents are then filled out using the *wdlpOst* editing system described below in section 3. Excerpts conform to standard practices in historical lexicography and are structured in a similar manner as DGLP entries, listing graphemic and phonemic variants, word senses, and derivatives (including compounds) with their respective variants. Variants and word senses are systematically documented with dated quotations to the extent that such data are available. During the excerption phase, and even afterwards, new candidates for loanwords may be found and subsequently added to the stock of source entries in an iterative process. Such new candidates can sometimes even be looked for in a systematic and extrapolative way by searching for words in an East Slavic language Y that from the point of view of historical phonology (and possibly semantics) closely correspond to known loanwords in another East Slavic language X. A typical example would be the search for Y-correlates of verbal prefixation formations already found for a certain verb stem in X.
- 3. *Compilation phase* (mainly Oldenburg): The often numerous excerpts of source entries on a Russian, Belarusian or Ukrainian lexeme are evaluated philologically and their data is merged into new XML documents, the *target entries* of the newly compiled Russian, Belarusian, and Ukrainian target dictionaries. In this phase, occasional or systematic additional inquiries at the editorial offices are still possible. In some cases, this might include requests for additional information on entries already published, e.g. on first quotations not included in print, but documented on the paper slips. The estimated number of entries will be around 2000. This amalgamation process is far from trivial and is significantly sped up by specific software tools in the *wdlpOst* editor. The most important one of these tools deals with word senses and will be presented below in section 4.

- 4. *Integration phase* (Oldenburg): Target dictionary entries on cognate words from Russian, Belarusian, and Ukrainian are re-examined philologically and from the point of view of historical linguistics; the results are documented as a cross-entry commentary that focuses on the possible and probable borrowing relationships and is supplemented by a visualization of possible borrowing pathways.

3. The Dictionary Writing System *wdlpOst*

For the specific purposes of the project a complex in-house server-based dictionary writing system named *wdlpOst* has been developed at the IDS. *wdlpOst* allows lexicographers to collaboratively edit excerpt documents and compile target entries in the stages 2 to 4 mentioned above. The following is a list of notable features and properties of *wdlpOst*:

- The system is based on a collaborative server/client infrastructure. In the default network mode, a desktop client application (henceforth, the *editor*) communicates via the Internet with a web service that in turn performs create/read/update/delete operations, mainly concerning XML documents, on a relational (Oracle) database management system.
- The web service is protected by strong cryptography (using digital signatures) and takes care of many validations, reporting and backup tasks including a locking mechanism for mutually exclusive access to individual excerpts and target entries.
- The desktop client (editor) operates with an underlying object-oriented data model. XML is used merely for serialization, i.e. for external storage purposes; for details, see Meyer (2014b).
- Client and server software is written in the Java and Groovy programming languages; in particular, this implies that the *wdlpOst* editor is a cross-platform desktop application.
- The client's user interface (GUI) is fully bilingual (German and Russian).
- The *wdlpOst* editor has an offline mode used, as stated above, in the editorial offices to fill out excerpt documents that are stored on the local hard disk. With a mouse click, all data edited so far can be sent to the server whenever Internet connectivity is available.
- For the editor, there are several special 'restricted input modes' that allow student assistants to fill in specific types of information excerpted from dictionaries without the danger of interfering with other entry parts.
- The editor features a live preview and automatic live validation of excerpts and target entries.
- There is a simple server-based source management system that provides a minimum of consistency for abbreviations and dates of quotation sources.

- The date input dialog used for quotations offers sophisticated options to specify ‘fuzzy’ dates where exact data are not available (such as ‘last third of 15th century’) and to distinguish between the dating of a historical source and the dating of the publication a quotation was taken from.
- The editor offers a system of drop-down menus as well as keyboard shortcuts for a large number of special characters of various scripts to be found especially in East Slavic historical dictionaries.
- There are currently three advanced search options available for queries on the project’s data: structured full-text search, XPath-based queries and an interface that presents the totality of the XML documents as a standard relational database with about 40 tables.

The *wdlpOst* system has been in productive use for excerpting data from the source dictionaries since mid-2014.

Figure 1 (below) shows a screenshot of the editor’s main window.

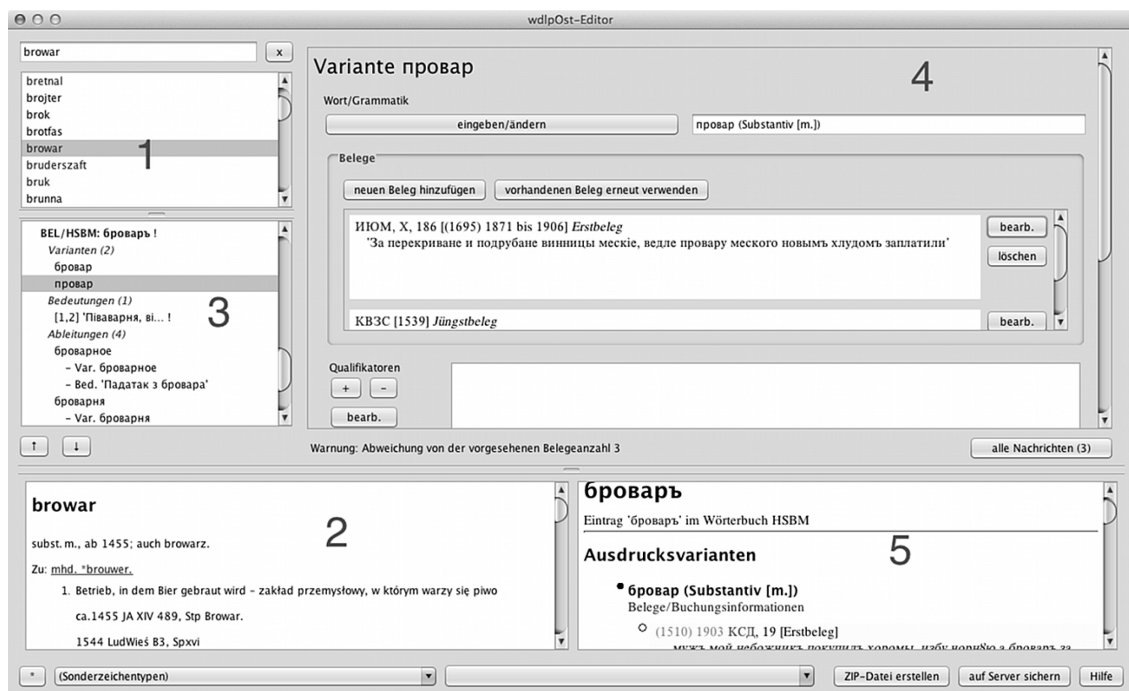


Figure 1: Main editing window of the *wdlpOst* desktop client

The Polish lemmas (and other recorded words such as derivatives as well as their meaning definitions) of the DGLP serve as a common frame of reference for all lexicographical work with the editor. Internally, the editor uses the full XML representation of the DGLP entries for various cross-referencing tasks. As a first step, the working lexicographer must select a Polish headword from the DGLP such as *browar* ‘brewer; brewery’ (from Middle High German *brouwer* ‘brewer’) in an alphabetical lemma list (1). A preview of the corresponding DGLP entry is displayed for quick reference in the main window (2). The central navigational device of the

editor is a list of all excerpts of East Slavic source entries that etymologically ‘belong’ to the DGLP entry selected, i.e., whose lemmas are considered loans from the DLGP lemma or one of its Polish derivatives or at least share their German etymon with it (3). The internal structure of each excerpt is indicated in a tree-like fashion below the headword. Figure 2 shows a part of the navigation list for Polish *browar* ‘brewer(y)’. Two still incomplete excerpts from source entries of different dictionaries can be seen in the image; the upper one concerns the entry *browar* in the Ukrainian historical dictionary SUM16-16 and features two phonologically distinct variants, two word senses, two derivatives (each of them with one graphemic variant and one word sense) and zero competing near-synonyms.

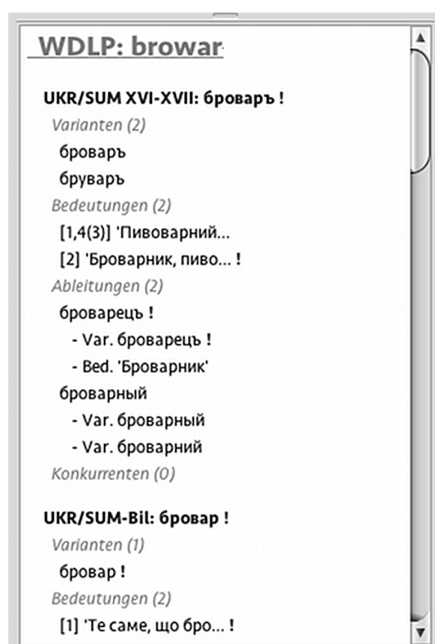


Figure 2: The editor’s navigation tree for a given DGLP headword (here: *browar*)

Clicking on a tree item (e.g., on one of the variant forms) opens the corresponding input panel (4) used for entering all pertinent lexicographical information, including an arbitrary number of records and quotations for a variant or word sense. The excerpt data is presented in a live preview HTML window (5).

4. Merging and Compiling: The Word Sense Mapping Tool

As noted above, the process of merging excerpts of different source entries on the same word during the ‘compilation phase’ is philologically, lexicographically and linguistically difficult: The excerpted source dictionaries (which usually cover different periods of the language) may or may not have different lemmatizations and microstructures, use incompatible word sense distinctions at distant points in the lumping-splitting continuum; there are several differing, partially historical spelling traditions; a lot of diasystematic variation on both the phonological and the morphological level is to be expected; and so on. In addition, there will usually be a lot

of duplicate and sometimes even contradictory information from the various sources. As a consequence, the *wdlpOst* editor includes dedicated tooling for eliminating redundancies and inconsistencies, pruning quotation lists, and other tasks. One of the most important tools, the *metasense editor*, serves to map word senses of excerpted source entries related to one and the same DGLP lemma onto a common cross-language set of ‘metasenses’. These metasenses are the word senses that are actually listed in the target entries for the German loanwords in East Slavic. Each metasense in a target entry is supplemented with the quotations, dates, and definitions of all those word senses in the various dictionaries that have been mapped onto it.

Mapping corresponding word sense information in multiple dictionaries is a well-studied lexicographical problem; cf. Jackson (2002: 91) for a typical textbook example. For the project’s ‘compilation phase’, such mapping is a vital step in operationalizing the investigation of the sometimes involved and even sense-specific borrowing history of words across dictionaries. A German word might have been borrowed multiple times into one or more of the East Slavic languages, each time on a different borrowing pathway (e.g., into Ukrainian either via Polish or via Polish and Russian or directly from German), with correspondingly differing phonological implications and, most importantly, in differing word senses. A careful examination must be based on all available data, i.e. semantics and phonology of all attested variants together with dates of the first and, possibly, last attestations of the different variants.

The need to define, for a set of cognate target dictionary entries, a cross-dictionary spectrum of word senses, is, as a consequence, of a practical nature. The mapping serves a twofold purpose, providing, on the one hand, the word senses of the target entries and, on the other hand, a tool for language contact research. Due to the convoluted history of the contemporary standard East Slavic languages and their common origin in a continuum of closely related dialects (cf. Müller & Wingender, 2013), it is important to be able to identify word senses of cognates across languages. This means that the same set of metasenses should be applied across all three languages.

As a consequence of this ‘instrumentalist’ understanding of the word sense mapping process, well-known important theoretical objections to ‘reifying’ word senses (cf. Hanks, 2000) do not apply in the context of the project described here. On a side note it is not a realistic goal to automate the matching process. There do exist several NLP-based proposals for this task (cf. Ide & Véronis, 1990) but they are geared towards tasks such as optimizing information extraction from multiple dictionaries for the purpose of creating lexical knowledge bases and thus cannot be expected to work well in a multilingual and diachronic setting that requires human philological expertise.

As already indicated in section 3, each (excerpt of) a source entry *E* is linked to a

DGLP entry *P*.³ In the ‘excerpt phase’, the lexicographer specifies, for each word sense *W* given in *E*, which word senses of *P* (if any) match *W* completely and which word senses of *P* (if any) match *W* only partially or potentially. Henceforth, this specification will be called the *DGLP profile* of the excerpted East Slavic word sense definition. Here, matching of an East Slavic word sense *W* with a DGLP word sense *W'* ideally means that the intension related to *W* is included in the intension related to *W'*. In practice, this is a rough and ready method to intuitively and preliminarily classify word sense definitions given the sparse information available. As we shall soon see, the results of this classification are used in the ‘compilation phase’ as a handy heuristic that aids in establishing metasenses.

Figure 3 shows the dialog window used in the editor for the classification procedure. In the hypothetical example shown, the East Slavic word sense definition in question is marked as completely matching sense nr. 1 ‘beer brewery’ and nr. 3 ‘suspicious, unpleasant place’ of the Polish lemma (here, *browar* ‘brewer(y)’) and potentially matching nr. 6 ‘pub’. This DGLP profile is abbreviated as [1,3(6)] throughout the editor. Note that the numbering of the DGLP word senses as well as the German and Polish sense definitions are taken from the original DGLP entries.

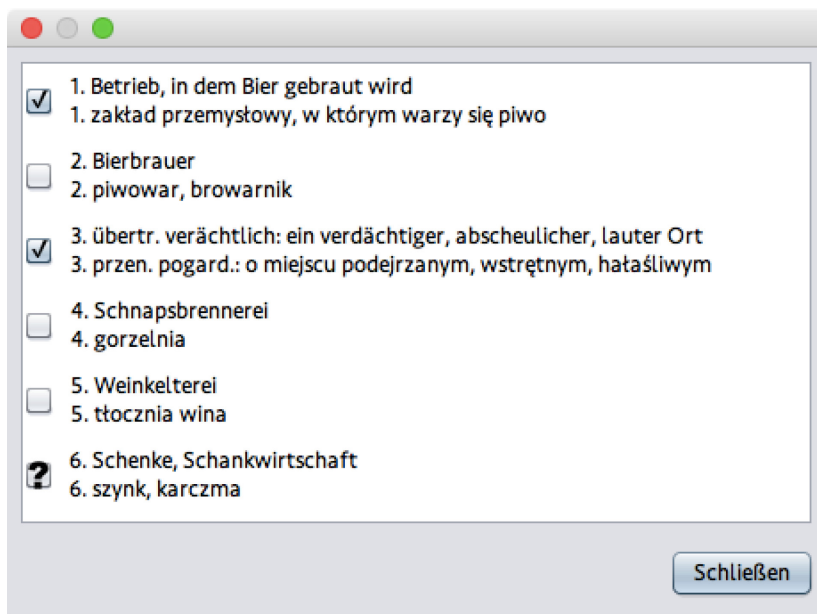


Figure 3: Dialog for assigning a DGLP word sense profile

The metasense editor, to which we now turn, gives the lexicographer a complete overview of all word senses in the excerpted source entries that have been assigned (linked) to a selected DGLP entry. In complicated cases with highly polysemous words there might easily be more than a hundred such word sense definitions, each of which with its own DGLP profile.

³ More precisely, the East Slavic lemma must explicitly be linked to either the lemma or one of its Polish derivatives or compounds as listed in the DGLP entry.

Figure 4 (below) shows the main dialog of the metasense editor, displaying the entirety of excerpted word senses in East Slavic source entries with associated Polish loanword *waga* ‘scales’, which has no less than 24 word senses in the DGLP.

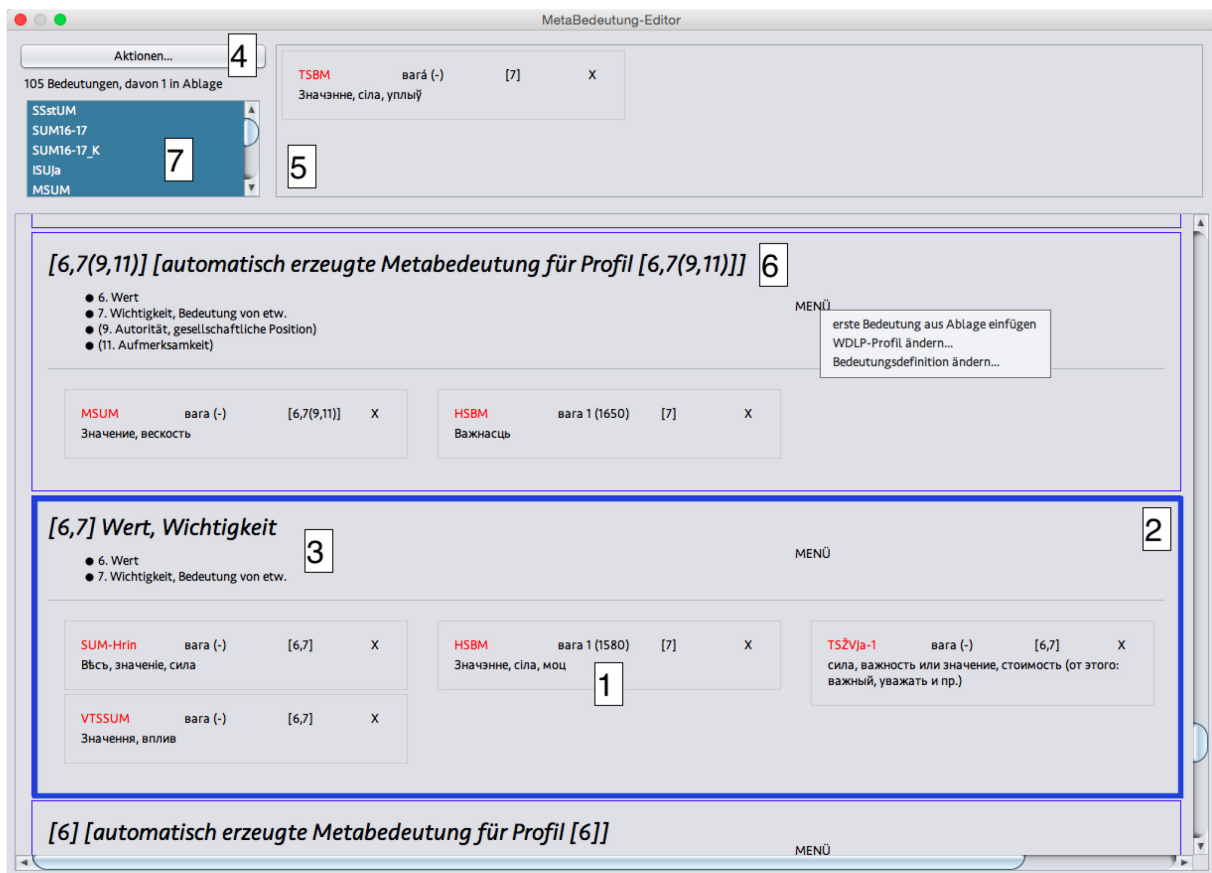


Figure 4: The metasense editor’s main window

Individual word senses as excerpted from source entries are the basic building blocks of the metasense editor. They are visually represented as ‘index cards’ like the one tagged with (1), shown enlarged in Figure 5. The index card contains the complete excerpted definition alongside the conventional abbreviation of the source dictionary, the lemma of the containing source entry in this dictionary, the date of first attestation of the word sense, and the DGLP profile. Double-clicking on the definition opens a window with full information on the word sense excerpt, including quotations and dates.

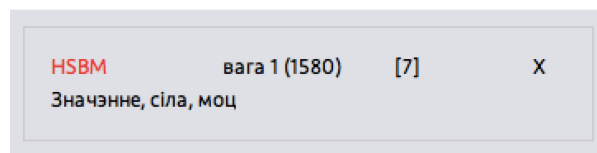


Figure 5: An index card for the word sense ‘meaningfulness, power’ recorded in the source entry *vaga 1* of the dictionary HSBM, with DGLP profile [7]

All index cards that are assigned to a certain metasense are enclosed in an outlined rectangle such as the one indicated in Figure 4 with a broad line (2). They are arranged in three columns according to the object language of the source dictionaries

(from left to right: Ukrainian, Belarusian, Russian) and, per default, sorted by dictionary and first attestation date.

Each metasense rectangle has a caption (3) showing both the (German) definition of the metasense (in the case of (3), ‘value, importance’) and its DGLP profile. Through an action menu (4) the lexicographer can define new metasenses as needed, specifying their definitions and their DGLP profiles. The latter ones will be shown as an additional orientation in the target entries of the loanword dictionary portal LWP. A metasense DGLP profile is independent of the DGLP profiles associated with the index cards belonging to it; in addition, different metasenses may have identical DGLP profiles. In particular, East Slavic loanwords might have word senses not found in the Polish cognate loanword; all such word senses have an ‘empty’ DGLP profile. There is a dedicated action menu button for each metasense that permits users to, amongst other things, reassign all its index cards (excerpted word senses) to another metasense or to simply delete the metasense. The editor will issue a warning whenever two metasenses have overlapping profiles.

At the beginning of the metasense editing process for a given DGLP lemma, only one default rectangle is shown in the editor that does not represent a metasense but simply contains the set of all index cards not yet assigned to any proper metasense. Index cards can be ‘cut’ from their containing metasense rectangle and thereby placed on the clipboard (5), from which they can be reassigned to another metasense by double-clicking on its rectangle’s metasense caption.

The DGLP profiles associated with the excerpted word senses can be used to ‘automatically’ create metasenses for all index cards of a select range of source dictionaries that are not assigned to an already defined metasense yet. This is accomplished by assigning all pertinent index cards with identical DGLP profiles to a newly generated metasense such as (6) having that same DGLP profile and a placeholder definition like ‘automatically created metasense with profile X’. This procedure is one of the main *raison*s *d’être* for the DGLP profiles. The automatic creation process can be initiated through the global actions menu (4) which offers various additional operations such as deleting all metasenses or ‘unassigning’ all of its index cards. It is possible to ‘clone’ an index card and assign the clone to another metasense. This is useful in cases where a word sense definition in an excerpt matches more than one metasense.

During the construction of the metasense spectrum, it is sometimes useful to have the system display only index cards for selected dictionaries (7). In addition, the editor can display which DGLP word senses are not part of any index card or metasense profile yet and optionally create, for any user-selected DGLP word sense x , a corresponding metasense that all index cards with profile $[x]$ are automatically assigned to.

From the above explanations it follows that there is a many-to-many relationship

between excerpted word senses and metasenses. This relationship is not encoded in the excerpts' XML documents but is represented in separate relational database tables. The approach outlined here strives for maximum generality. It would have been much simpler, yet philologically unfeasible, to simply take the DGLP word senses as the *tertium comparationis* for classifying East Slavic word senses: Sometimes the sense distinctions in DGLP loanwords might be too fine-grained, sometimes too coarse for the task at hand.

5. Outlook and conclusion

Several other editor tools for the 'compilation phase' are currently under development. In particular, there will be a 'metavariant editor' that assists the lexicographer, in a fashion similar to the metasense editor, in constructing a cross-dictionary and cross-language system of the graphemic/phonemic variants of all the East Slavic cognates of a Polish loanword in the DGLP. The main purpose of this tool is (a) to abstract from irrelevant spelling variation found in dictionaries of the same language and, additionally, (b) to identify words across Slavic languages that are, from the point of view of diachronic and contact phonology, 'equivalents' of each other (show regular or at least very frequent and typical correspondence patterns for all phonemes), such as Polish *rynek*, Russian *rynok*, Ukrainian *rynok*, Belarusian *rynak*. A similar tool will be available for the derivative forms of East Slavic loanwords.

All of these tools help lexicographers to create synoptic and slightly abstractive representations of certain aspects (lexical semantics, (mor)phonology) of cognate loanwords across the four languages involved. These representations are a useful point of departure for the linguistic assessment of the exact borrowing history of East Slavic loanwords with a German origin. Condensed, tabular versions of these representations will be part of the final target entries; they essentially display, for all four Slavic languages, the dates of the first and – where applicable – last attestation of the metasenses or metavariants at hand. More important, though, is another function of the synopses created by these tools: They make it possible to define the semi-automatic merging process whereby the lexicographical data from a potentially large range of excerpts can be amalgamated to form a target entry. When all synopses are created, the working lexicographer must select those 'metavariants' that he considers to be subsumable under one East Slavic target headword; the *wdlpOst* system can then automatically generate a complete draft version of the target entry, taking into account all metasenses and 'metaderivatives' associated with the metavariants chosen and incorporating all pieces of information from the excerpted dictionaries that are mapped to these meta-items.

This paper has focused on one aspect of the more general conceptual question of how a dictionary writing system can assist in creating cross-linking information between the three layers of lexicographical data involved in the project described here, i.e. the DGLP entries on Polish loans from German; the excerpted data from East Slavic

source dictionaries; and the East Slavic target entries. The intricate lexicographical, linguistic, and technical problems discussed above have let it seem, *pace* de Schryver (2011), unfeasible to simply customize an off-the-shelf dictionary writing system or an XML-editor based software solution; see Meyer (2014a, b) for more detailed argumentation. On the other hand, as is typical of projects in modern electronic lexicography, the in-house software solutions created as a response to this situation also do not lend themselves to easy generalization or abstraction beyond the confines of the very specific project they have been built for.

6. Acknowledgements

I would like to thank Gerd Hentschel and Sabine Ute Anders-Marnowsky (University of Oldenburg) for valuable input and information regarding the philological and lexicographical aspects of the project. Their ideas and thoughts have shaped most aspects of the lexicographical process that is reflected in the software described in this paper.

7. References

- de Schryver, G.-M. (2011). Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters. In K. Akasu & S. Uchida (eds.) *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the Seventh ASIALEX Biennial International Conference, Kyoto, Japan, August 22–24, 2011*. Kyoto: Asian Association for Lexicography, pp. 647-656.
- DGLP: *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts* (2010) [*Dictionary of German Loanwords in Standard and Written Polish*] (edited by de Vincenz, A. & Hentschel, G.; *Studia slavica Oldenburgensia*, vol. 20). Oldenburg: BIS-Verlag. Accessed at: <http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp>. (25 May 2015)
- Hanks, P. (2000). Do Word Meanings Exist? *Computers and the Humanities*, 34, pp. 205-215.
- HSBM: *Historyčny sloŭnik belaruskaj movy [Historical Dictionary of the Belarusian Language]* (1982–). Minsk.
- Ide, N. & Véronis, J. (1990). Mapping dictionaries: A spreading activation approach. In *Proceedings of the 6th Annual Conference of the Centre for the New OED*, University of Waterloo, Canada, pp. 52-64.
- Jackson, H. (2002). *Lexicography. An Introduction*. London/New York: Routledge.
- LWP: *Lehnwortportal Deutsch*. Accessed at: <http://lwp.ids-mannheim.de>. (25 May 2015)
- Meyer, P. (2013). Advanced graph-based searches in an Internet dictionary portal. In I. Kosem, J. Kallas, P. Gantar, P. Krek, M. Langemets & M. Tuulik (eds.)

Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 488-502. Available at:
http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf.

- Meyer, P. (2014a). Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries. In A. Abel, Ch. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014, Bolzano/Bozen*, Bolzano/Bozen: EURAC research, pp. 1135–1144. Available at:
http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_088_p_1135.pdf.
- Meyer, P. (2014b). Entlehnungsketten in einem Internetportal für Lehnwörterbücher. IT-Infrastruktur und computerlexikographischer Prozess in einem Projekt zu polnisch vermittelten Germanismen im Ostslavischen. In M. Mann (ed.) *Digitale Lexikographie. Ein- und mehrsprachige elektronische Wörterbücher mit Deutsch: aktuelle Entwicklungen und Analysen* (= Germanistische Linguistik, 223-224). Hildesheim/Zürich/New York: Georg Olms Verlag, pp. 97-132.
- Müller, D. & Wingender, M. (eds.) (2013). *Typen slavischer Standardsprachen. Theoretische, methodische und empirische Zugänge*. Wiesbaden: Harrassowitz.
- SRJa11-17: *Slovar' russkogo jazyka XI–XVII vv. [Dictionary of the Russian Language from the 11th to the 17th Century]* (1975–). Moskva.
- SRJa18: *Slovar' russkogo jazyka XVIII veka [Dictionary of the Russian Language of the 18th Century]* (1984–). Leningrad/St. Peterburg.
- SUM16-17: *Slovnyk ukraïns'koi movy XVI – peršoi polovyny XVII st. [Dictionary of the Ukrainian Language from the 16th to the First Half of the 17th Century]* (1994–). L'viv.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Using machine learning for semi-automatic expansion
of the *Historical Thesaurus*
of the *Oxford English Dictionary*

James McCracken

Oxford University Press

E-mail: james.mccracken@oup.com

Abstract

The *Historical Thesaurus of the Oxford English Dictionary* (HTOED) provides a highly granular taxonomic classification of the contents of the OED. However, HTOED was based largely on the first edition of the OED (plus supplements), and has not been updated to include content added more recently, or changed content emerging from third-edition revisions. This means that 32% of lexical items in the current OED data set are unclassified.

We use the existing HTOED classifications as training data to classify this ‘missing’ content. The classification system works as a two-stage process. Firstly, for a given input sense, a Bayesian classifier identifies the general topic (high-level thesaurus branch) to which the sense belongs; secondly, a battery of similarity measures identifies possible target nodes within this branch. The system looks for consensus or proximity among the outputs of these methods, in order to pinpoint the optimal node(s) to which the sense should be assigned.

The system is currently able to classify 25% of input senses to the correct node, and a further 40% of input senses to the right neighbourhood (a parent, child, or sibling of the correct node). A web-based UI facilitates the manual checking, approval, and adjustment of proposed classifications.

Keywords: Oxford English Dictionary; Historical Thesaurus; machine learning; lexical ontology; feature extraction

1. Introduction

The *Historical Thesaurus of the Oxford English Dictionary* (HTOED) is a taxonomic classification of the content of the *Oxford English Dictionary* (OED), compiled at the English Language department of the University of Glasgow between 1965 and 2008. The HTOED data were integrated with the OED data in 2010, and now form a core part of OED Online (www.oed.com/thesaurus). The HTOED is also available as a standalone resource at <http://historicalthesaurus.arts.gla.ac.uk/>, and is published as a two-volume book (Kay et al., 2009).

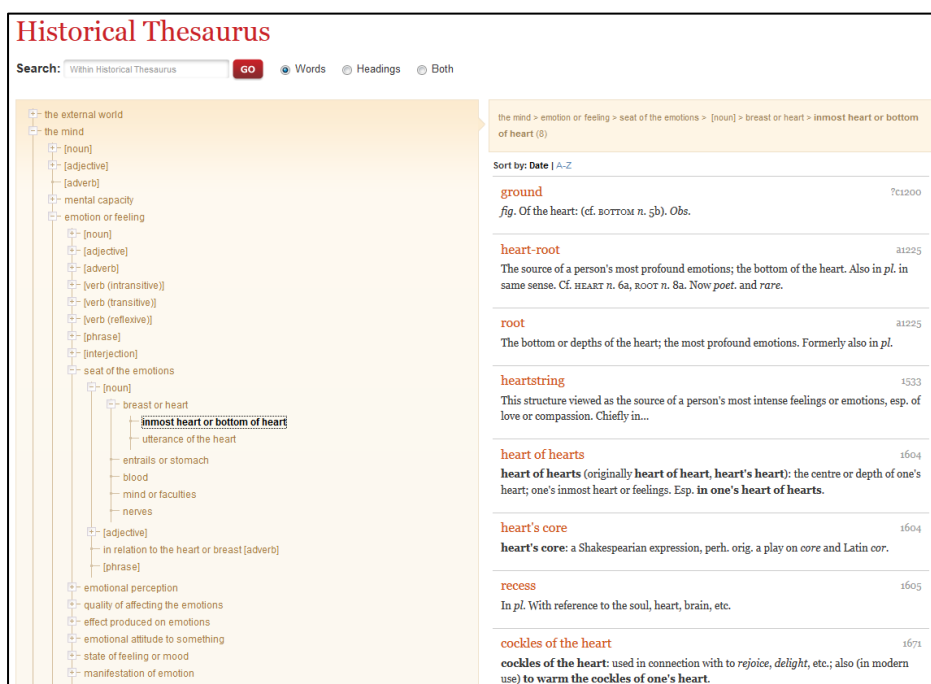


Figure 1: HTOED integrated with OED Online. The taxonomy is shown on the left; the senses in a selected node are shown on the right.

HTOED was based largely on the first edition of the OED (1888–1928) and its supplementary volumes (1933; 1972–1986), and latterly extended to include new material from the OED *Additions* volumes (1993–97).¹ This is now incomplete relative to the current state of the OED, in two main respects:

1. Certain categories of OED material, such as undefined compound lemmas, were systematically omitted from HTOED;
2. HTOED has not been updated to cover new material added to OED since 1997, or new and changed sense distinctions emerging from the Third-edition revision programme which began in 1993.

Consequently, a third of all senses in the current OED data set (264,000 out of 821,000) are not covered in HTOED.²

A project within the OED programme is currently attempting to ‘complete’ the HTOED by assigning an HTOED classification to as many of these 264,000 ‘missing’

¹ It also contains material from several Old English dictionaries, also published separately as *A Thesaurus of Old English* (Roberts & Kay, 1995). Much of this material falls outside the scope of the OED.

² Throughout this paper, I use ‘sense’ to mean any semantically distinct unit of an OED entry, including both senses of main headwords and sublemmas.

senses as possible. This is being done semi-automatically: a supervised machine-learning process uses the existing classifications as training data to classify the input senses (in some cases generating two or three ‘candidate’ classifications); these classifications are then accepted, rejected, or adjusted by human reviewers.

2. Viability of machine learning

On the face of it, this is a very attractive machine-learning task: 557,000 senses manually classified by a team of well-trained researchers within an academic department should make for a very rich and reliable set of training data.

But there are some complicating factors:

1. The HTOED taxonomy is highly granular: for any given input sense, there are over 200,000 candidate labels (i.e. taxonomy nodes), so the amount of training data decreases sharply as you go down a taxonomic branch. The number of training-data senses usually drops to single figures by the fifth or sixth level down.
2. The input senses are not altogether similar to the training-data senses. That is to say, the input senses are not a random subset of the population as a whole. For example, input senses tend (on average) to be more recent, more technical, or more minor than the training-data senses.
3. Although the training data as a whole is very rich, each individual document (dictionary sense) tends to be short and feature-poor. So for a given input sense, it may be difficult to extract a set of features good enough to support comparison with the training-data models.
4. The OED’s unrestricted defining vocabulary means that individual feature values (e.g. a specific word or phrase in a definition) may be very sparse.³

The HTOED taxonomy also presents some problems:

1. The HTOED taxonomy was developed ‘bottom-up’, largely determined by the material that happened to be in the first and second editions of OED (Kay et al., 2009: p.xix). At the very fine-grained level (leaf or near-leaf nodes), the HTOED taxonomy expresses a variety of relations, e.g. meronymy and other associative relations, as well as hypernymy. This fine-grained structure tends to be determined by the specific definitions of the original member senses of each node; hence at this level it becomes harder and harder to determine that an input sense belongs to a given node, and probabilistic approaches break down.

³ This problem is most acute when dealing with superordinates; see section 9.

2. Moreover, a new input sense may represent a concept not currently accounted for by HTOED; so there may be no correct classification in terms of the existing taxonomy.

2.1 Two-stage system

For these reasons, we found that no single machine-learning model was adequate for the task. Instead, we developed a two-stage system:

1. For a given input sense, a naïve Bayes classifier (the `Topic_classifier` module) is used first to identify the probable topic(s), i.e. a relatively high-level branch of the taxonomy;
2. A range of more targeted methods (often with their own Bayesian models) are then applied to determine a specific node within that branch.

These results are collated by a top-level module (the `Central_classifier`) to determine the final classification assigned to the input sense.

Although particular methods may require some parsing and analysis (e.g. to identify superordinates within a definition; see section 9), in general terms this approach is statistical rather than rule-based. That is to say, an input sense is classified by comparing its features to the training data, rather than by any direct attempt to decode its definition. This allows for models that are adaptable to the very variable nature of OED senses.

2.2 Summary of classification methods

The `Central_classifier` first uses the `Topic_classifier` to restrict the search-space within the taxonomy to a particular branch or branches. The following set of methods is then applied to try to find a more specific node or region within that branch:

- **Cross-reference:** If an input sense cross-refers to another sense that has already been classified, this may indicate how the input sense should be classified. See section 4.
- **Taxonomic binomial/genus term:** An animal- or plant-name definition often includes a binomial name, or at least a genus name; this can be used to find an exact classification. See section 5.
- **Synonyms:** If an input sense definition includes one or more synonyms, the classification of the synonym words may indicate how the input sense should be classified. See section 6.

- **Morphology:** A derived form can usually be assumed to be semantically close to its root, or to sibling terms derived from the same root. See section 7.
- **Compound form:** The elements of a compound lemma may be indicative of sense. See section 8.
- **Superordinate:** If the superordinate term can be extracted from the definition of an input sense, this can be compared with other senses with the same or similar superordinate term. See section 9.

For each input sense, all methods are attempted, effectively in parallel.⁴ If a number of different classifications are returned, the following procedure is applied:

1. The `Central_classifier` polls the results to look for cases where two or more methods have returned the same (or very similar) classifications;
2. If multiple classifications still remain, the classification chosen by the most reliable method is preferred; the remainder are treated as runners-up;
3. If no classifications remain (or if no classifications were returned in the first place), the `Central_classifier` defaults to the `Topic_classifier`'s best-guess branch.⁵

As step 2 indicates, the system is often dependent on *a priori* rankings of different methods (e.g. for a typical sense, classification by cross-reference is ranked as more reliable than classification by synonyms). This system, therefore, does not always take individual circumstances into account (e.g. there may be occasional senses where synonyms are a better bet than a cross-reference).

2.3 ‘Runner-up’ classifications

If the `Central_classifier` retrieves multiple candidate classifications, one of these will be selected as the ‘winner’ and treated as the primary classification. If any others remain, the top one or two are selected as ‘runners-up’. Runners-up usually indicate different lines of attack that were considered by the `Central_classifier`.

In some cases, a supposed runner-up may turn out to be a better classification than the winning classification. The editorial interface provides a means to promote a runner-up ahead of the primary classifications; see section 11.

⁴ Not all methods succeed in all cases, of course; for example, the cross-reference method will fail if the input sense has no cross-references. In such cases, a null result is returned, and is discarded.

⁵ This will almost always be too high up in the taxonomy to be correct as it stands, but often provides a good starting point for human checking to identify the correct node further down.

3. Topic_classifier module

The Topic_classifier module is responsible for generating a ranked list of the three or four most likely topics (high-level branches of the HTOED taxonomy) for each input sense. This is used to restrict the search-space available to the more targeted classification methods employed by the Central_classifier. It may also be used to assist some of those methods more directly, e.g. to help pick likely senses of a synonym.

3.1 Flattened categories

The set of labels (i.e. the categories to which the Topic_classifier can assign a sense) is the set of thesaurus branches which contain 2000+ senses. This adds up to about 200 branches in total, some of which are sub-branches of others. The Topic_classifier treats these 200 branches as a flat list of disjoint labels.

This ‘flattened’ method may seem counter-intuitive. I spent some time experimenting with ‘taxonomy-aware’ classifiers, e.g. using decision trees (classifying first at level 1, then at level 2, level 3, etc.), but these approaches proved less successful. In practice, so long as each branch is reasonably well-populated, the Topic_classifier does not really need to know about the taxonomy. For a given sense, probabilities are calculated for each label in turn, and the label with the highest score wins. This may turn out to be a branch at any of the upper levels of the taxonomy.

3.2 Feature set

The feature set used includes the following:

- lemma (or lemma elements, in the case of MWEs);
- subject labels;
- register and usage labels;
- tokens from definition text;
- tokens from modern quotation text;
- tokens from quotation titles;
- author names;
- presence/absence of taxonomic binomials;
- first date (binned by 50-year periods);

- part of speech.

Tokens are all case-stripped, Porter-stemmed, and truncated to a maximum of eight characters. For example, *historically* and *historicism* are both normalized to *historic*.

3.3 Confidence score

A confidence score between 0 and 10 is associated with the ranked list that the `Topic_classifier` computes for each input sense. (A zero score indicates that the `Topic_classifier` has failed altogether, usually because the input sense provides insufficient features.)

The confidence score is a measure of the number of features provided by the input sense, and the margin by which the top two or three labels outscored the rest. If the confidence score is low, the `Topic_classifier` may be partly or wholly disregarded by other classification methods (i.e. the search-space is not restricted), and the `Topic_classifier` will not be used as a fallback if the other classification methods fail.

3.4 Sanity check

The `Topic_classifier` module acts as a kind of sanity check on some of the more deterministic methods described below. It tends to preclude or at least deprecate some of the more egregious errors that can arise from mistakes in a particular classification method: misinterpreting a word in a definition, misidentifying a superordinate, failing to correctly separate metalanguage from gloss, etc.

At the same time, the use of confidence measures prevents the `Topic_classifier` from being too aggressive in pruning away candidates.

4. Cross-references

Cross-references are a valuable way to contextualize a given input sense. Uniquely among the classification techniques discussed here, cross-references can be used deterministically rather than probabilistically, meaning that classifications made in this way tend to be both more accurate and more reliable.

4.1 ‘Equals’-type cross-references

An equals-type cross-reference provides the easiest win for the `Central_classifier`: if the target sense is classified, the input sense can simply adopt the classification of the target sense.

For example, *emulsin* is defined as:

A neutral substance contained in almonds; = SYNAPTASE n.

Here the Central_classifier can safely ignore the definition and any other features of the input sense, and just copy the existing classification of the target sense of *synaptase*.

There are various formulae that can be treated in this way: not only a leading equals sign as in the *emulsin* example, but also formulae like ‘another name for...’, ‘short for...’, ‘variant of...’, etc.

About 15,000 senses are classified this way (7% of classified senses).

4.2 ‘Cf.’-type cross-references

‘Cf.’-type cross-references do not provide such a direct and positive means of classification; but they nevertheless provide a good indicator of the right branch, at a fairly granular level.

For example, *generically 2* is defined as:

Biol. In a generic manner; with reference to genus. Cf. GENUS n. 2a.

So we can be fairly confident that *generically 2* belongs in the adverb branch parallel to the noun branch in which *genus n. 2a* is found.

About 8,000 senses are classified this way (4% of classified senses).

4.3 Other cross-references

Other cross-references are useful not as classification methods in their own right, but as ways to improve the performance of other methods.

In particular, parenthetical cross-references within a definition often serve to disambiguate keywords, especially to make clear that a word is not being used in its primary modern sense.

4.4 Problems with cross-references

Cross-references can be susceptible to the kind of problem described in section 6.3 in relation to synonyms; namely that focussing on a cross-reference to the exclusion of the rest of the definition risks ending up with a classification that only captures one aspect of the sense, not its primary meaning.

For example, *general servant* is defined as:

A servant whose duties are general rather than limited to a particular sphere; *spec. = maid-of-all-work*.

Because the `Central_classifier` focusses on the cross-reference, it ends up with the specific classification of ‘housemaid’ rather than the more general classification suggested by the main gloss.

5. Taxonomic binomial and genus names

Definitions for animal and plant names often include an explicitly tagged taxonomic binomial or genus names. By indexing all such names in the training data, we build a model mapping binomials to HTOED classifications. This can then be used to classify any input sense containing a taxonomic term in its definition.

For example, *Java lemon* is defined as:

A small lime tree, *Citrus aurantifolia* (formerly *C. javanica*), originating in South-East Asia...

This sense can therefore be classified by checking the classification of training senses which also include *Citrus aurantifolia*. Failing that, the right branch can be found by checking the classification of training senses which include some other *Citrus* ... binomial.

About 4,500 senses are classified this way (2.3% of classified senses).

6. Synonyms

Although OED senses do not identify synonyms explicitly, OED definitions are very rich in synonym-like terms. These provide a useful aid to classification. If an input sense includes a synonym that can be reliably identified and disambiguated, then the classification of that synonym will be a good indicator of how the input sense should be classified.

It’s unusual for an OED definition to depend wholly on synonyms, but it’s quite common for definitions to include synonyms in some form as an adjunct or support to the main definitional gloss. This can be particularly valuable when dealing with adjectives; somewhat less valuable when dealing with verbs and adverbs; and least useful when dealing with nouns.

About 12,000 senses are classified using synonyms (6% of all classified senses).

6.1 Patterns

The prototypical pattern for a synonym-rich definition is something like this:

Main gloss here; foo, bar, or baz.

where *foo*, *bar* and *baz* are the synonyms.

For example, *abhorred* is defined as:

Regarded with disgust or hatred; detested, loathed, abominated.

where *detested*, *loathed*, and *abominated* serve as synonyms.

Beyond this prototypical pattern, there are nine or 10 other patterns which can also be used to identify synonyms within a definition. Slightly different patterns apply to different wordclasses.

6.2 Disambiguating synonyms

Having identified a synonym or synonyms for a given definition, the system then looks up the synonym's own OED entry, finds the appropriate sense, and examines how that sense has been classified.

'Finding the appropriate sense' is the difficult bit. It is tempting to assume that synonyms will usually be used in their main modern sense; but in practice this turns out not to be the case. Since a definition usually consists of a gloss followed by one or more synonyms (as with *abhorred* above), the gloss serves to prime a particular sense of the synonym word – which may or may not be the main sense.

For example, *generous* 4b is defined as:

Of an action, a gift, etc.: readily done or given; more than is strictly necessary or expected; large, ample, bounteous.

where *large*, *ample*, *bounteous* can be identified as synonyms. *Large* here does not have its usual modern sense, but rather has the (now somewhat unusual) sense of 'liberal', primed by the preceding gloss.

Similarly, in a list of two or more synonyms, the meaning of each synonym may be primed by the others in the list. For example, *glee* 1b is defined as:

Of the eye: quick, sharp.

where *quick* and *sharp* are primed by each other so that we understand them in their 'shrewd' sense rather than in their more prototypical 'speedy' or 'keen-edged' senses.

Hence there are two main ways to disambiguate a synonym:

1. Use the `Topic_classifier` to determine the broad subject area of the sense, then

look for a sense of the synonym that falls within this subject area;

2. If the synonym is one of a list of synonyms, look for senses of the synonyms which cluster on a particular branch of the HTOED taxonomy (e.g. the ‘shrewd’ senses of *quick* and *sharp* are clustered on the ‘sharpness, shrewdness, insight’ branch of the HTOED taxonomy).

In practice there can be problems with both methods:

- Method #1 can fail because apparent synonyms are not always direct semantic equivalents, for the reasons discussed in section 6.3 below;
- Method #2 can fail because a list of synonyms may not be synonyms of each other, and so may not lie on the same taxonomic branch: the purpose of a list of synonyms is often to stake out the wider semantic territory, rather than to indicate a specific single meaning.

Because disambiguation can be problematic, it is often easier to focus on unambiguous synonym words. For example, *grait* 2c is defined as:

Of a stroke: clean, unimpeded.

where *clean* and *unimpeded* can be identified as synonyms. But because *clean* is polysemous, it is easier to focus instead on the less ambiguous *unimpeded*. However, this can exacerbate the problems discussed in section 6.3 below: the more unambiguous synonyms are often the more partial.

6.3 Are these really synonyms?

The patterns mentioned in section 6.1 above identify words that occupy a synonym-like slot in the definition; but this does not guarantee that they are actually synonyms in the strict sense. In fact, in the prototypical ‘gloss + synonyms’ pattern, the supposed synonyms may really be extensions, generalizations, or weakenings of the main gloss, rather than restatements of it.

A consequence of this is that a classification based on a synonym may capture certain aspects of the sense, but miss the core meaning.

For example, *mus* 2 is defined as:

Given to or characterized by meditation; contemplative, thoughtful, dreamy.

where *contemplative*, *thoughtful*, *dreamy* are identified as synonyms. But *dreamy* here is rather different from the main gloss *given to or characterized by meditation*. If the Central_classifier focusses on *dreamy*, the sense will end up with a classification that reflects a minor extension of the sense rather than its core meaning.

So although synonyms are in principle a very direct and widely-available aid to classification, in practice the issues of disambiguation mean that these are not always usable. Moreover, some apparent synonyms may really be distractions from the core sense. It is often better to treat synonyms as a supplement to other methods, rather than as a classification method in their own right.

7. Morphology

A derivative form can usually be assumed to be on the same branch of the HTOED taxonomy as its parent or root word. If the derivative is in a different wordclass from its root (e.g. an *-ize* verb derived from an adjective), it can be assumed to be in a branch of the HTOED taxonomy parallel to that of its root.

If the root word has more than one sense, a run-on derivative lemma can usually be assumed to be related to the main sense of the root. However, this becomes more problematic if the root word has many possible senses; classification by morphology is not usually attempted in such cases.

This approach can be adapted for ‘sibling’ derivatives, i.e. two derivative subentries derived from the same root word. For example, the likely classification of *causationism* can be inferred from the existing classification of its sibling *causationist*.

About 16,000 senses are classified this way (8% of classified senses).

8. Compounds

About 34% of all input senses are compound subentries. There are also many main-entry senses which have a compound form. A special module (the `Compound_classifier`) is dedicated to determining candidate HTOED classifications based on the compound form itself.⁶

8.1 Initial assumptions

Our initial approach to handling compounds was to assume by default that the meaning of a compound lemma (and therefore its HTOED classification) is encoded in the lemma form, i.e. that the compound is endocentric. This assumption was strongest in the case of undefined subentries (21% of all input senses).

Most compounds (especially nominals) were taken to be head-final, i.e. the last element is a hypernym, and the first element is a qualifier.

⁶ This draws on an extensive body of research into compounding and semantics in English; see Bauer (2009), Booij, (2007) and Lieber (2004).

The appropriate HTOED branch (if not the specific HTOED class) was therefore assumed to be related to one of the senses of the last element (and usually one of its main senses). Thus *furniture-van* is a hyponym of one of the main senses of *van*; *wheat-maggot* is a hyponym of one of the main senses of *maggot*.

But early testing found that these assumptions produced poor results. In particular, the assumption that the meaning of a compound can be deduced from the main sense of its last element turned out to be flawed in many cases. For example:

- *ship-jumper* is not a hyponym of any listed sense of *jumper*;
- *character assassin* is not a hyponym of any listed sense of *assassin*.

8.2 Probabilistic model of compounding

This led to a different strategy: rather than assuming compounds to be endocentric and head-final, we built a Bayesian model of compound semantics within OED, using both the first and last elements of the lemma. For each training-data sense with a compound lemma, the first and last elements are indexed against the HTOED classification of the sense. For a given input sense with a compound lemma, the most likely branch(es) of the HTOED taxonomy can then be predicted from these models.

Having identified a branch, the Compound_classifier can then revert to the more naïve assumption: the specific class within this branch is identified by focusing on the last element; either by looking for other compounds within the selected branch that have the same last element, or by looking for a sense of the last element that falls within the branch.

For example, the undefined compound *matrimonial broker* is classified as follows:

1. The Compound_classifier evaluates the two elements *matrimonial* and *broker* against the Bayesian model. This finds that initial *matrimonial* is strongly correlated with the *community » kinship or relationship* branch, whereas final *broker* is most strongly correlated with *occupation » trade and commerce*, and more weakly correlated with *community » kinship or relationship*.
2. The net result is that *community » kinship or relationship* is selected as the most likely branch.
3. The Compound_classifier then tries to find the specific class within the *community » kinship or relationship* branch. It tries two approaches in parallel: (a) it checks for senses of *broker* that fall within this branch; (b) it checks for other compounds with *broker* as the last element which falls within this branch. Approach (a) fails in this instance, but approach (b) finds a cluster of *-broker* compounds in the *community » kinship or relationship » marriage or wedlock »*

match-making » *match-maker class (match-broker, flesh-broker, wife-broker, etc.)*. This is therefore selected as the class to which *matrimonial broker* will be assigned.

The process works very neatly with the example of *matrimonial broker*, but many examples are not so clear-cut. Often the `Compound_classifier` will draw on the `Topic_classifier` to help arbitrate between competing possibilities.

8.3 Successes

The following are examples of compounds which were incorrectly classified by methods based on the earlier endocentric, head-final assumptions, but which are correctly classified by the more probabilistic approach of the `Compound_classifier`:

- *truth-speaking*: classified as *mental capacity* » *faculty of knowing* » *conformity with what is known, truth* » *sincerity, freedom from deceit* » *sincere*
- *mimosa scrub*: classified as *the earth* » *land* » *landscape* » *fertile land or place* » *land with vegetation* » *wooded land*
- *vision-monger*: classified as *mental capacity* » *expectation, looking forward* » *foresight, foreknowledge* » *prediction, foretelling*
- *quiet-footed*: classified as *sensation* » *hearing* » *inaudibility* » *inaudible* » *silent* » *of footsteps*
- *vine-clad*: classified as *the earth* » *land* » *landscape* » *fertile land or place* » *land with vegetation* » *covered with vegetation* » *wooded*

8.4 Casualties

Not all compound-handling is improved by the `Compound_classifier`; some compounds were better served by the earlier approach.

For example, *junction piece* (which seems to be something to do with plumbing) gets classified as *travel* » *travel by railway* » *railway system or organization*, due to the fact that *junction* is strongly correlated with railways.

Still, when the `Compound_classifier` gets things wrong, it at least tends to do so with a certain wit, as when it misclassifies *butt mark* (an archery term) as *...animal husbandry* » *animal keeping practices general* » *branding or marking*.

8.5 Compounds with unusual elements

There are cases where the `Compound_classifier` draws a blank for a given input sense, because either the first or last element of the lemma is unusual and so does not appear in the predictive model.

In such cases (for undefined compounds, at least), the `Central_classifier` will disregard the `Compound_classifier` and fall back to a more naïve approach, usually reverting to the assumption that the lemma is a hyponym of the main sense of its last element. Failing that, it may just leave the sense unclassified.

8.6 Figurative, poetic, and metaphoric compounds

Many of the OED's undefined compounds are figurative or metaphoric to some extent. The intended meaning is often vague or unclear (often there is only a single quotation).

For example, *strife-race*, which has the single quotation:

The strife-race, for we must run, and fight as we run, strive also to outstrip our fellow-racers,

gets classified as *leisure » sport and outdoor games » types of sport or game » racing or race » racing on foot* – which is not bad, except that it completely misses the fact that *-race* here is used metaphorically.

Some of the more poetic compounds involve deliberate repurposing of the first or second elements. For example, *panther-peopled* (‘Amid the panther-peopled forests...’) means not ‘peopled’ at all, but rather ‘occupied by panthers’. The `Compound_classifier` does not really get to grips with such compounds at all.

It is debatable whether it is even worth attempting to include such compounds in the classification exercise. But that is a moot point, given that currently there is no sure way to distinguish between literal and figurative compounds.⁷

9. Superordinates

For noun senses in particular, identifying the superordinate within the definition is often a critical part of the classification process.

The classification process is based primarily on the training data: having identified the superordinate of a given input sense, the `Central_classifier` checks for training senses

⁷ On the relationship between HTOED and metaphor, see Alexander & Bramwell, 2012.

that have an identical or similar superordinate, and examines how these are classified. Contextual information, notably the `Topic_classifier`'s evaluation, may be used to arbitrate in the case of several competing possibilities.

About 33,000 senses are classified using superordinates (17% of all classified senses).

9.1 Process

This process can be broken down into a series of subtasks:

1. Separate the definition proper (the core gloss) from any metalanguage or secondary clauses;
2. Tokenize and p.o.s.-tag the gloss;
3. Chunk into noun phrases; the first noun phrase is presumed to be the superordinate in raw form;
4. Normalize the superordinate to allow fuzzy matching;
5. Retrieve training senses with (fuzzily) matching superordinates;
6. Cluster matching training senses into candidate HTOED branches;
7. Select the best HTOED branch, if there is more than one candidate (using the `Topic_classifier` or other secondary indicators).

9.2 Difficulties

There are potential difficulties with each of these steps, but the critical problems lie in steps 1 and 4. The general problem with step 1 (extracting the core gloss from metalanguage) is discussed in section 12.3. With respect to superordinates, this issue means that a metalanguage phrase may be erroneously identified as the superordinate.

Step 4 (normalization of the superordinate noun phrase) is required because, taken literally, many superordinates are unique or near-unique noun phrases. For example, *lagre* is defined as:

In sheet-glass making: A sheet of perfectly smooth glass, placed between the flattening stone and the cylinder to be flattened.

The noun phrase containing the superordinate here is identified as *a sheet of perfectly smooth glass*. Since no other sense is defined in exactly the same way, this would draw a blank with the training data. However, if this is normalized to *glass sheet* (rearranging the syntax, and omitting possibly extraneous words), this now has more

chance of matching training-data senses (given that the training data is also normalized in the same way).

Normalization of this kind is difficult: it is difficult to figure out what can be omitted, and sometimes difficult to reorganize into an optimal form. It is also tricky to figure out how far to normalize. For example, in some cases it may be beneficial to normalize synonyms towards their prototypes (so that e.g. *tracts of arable land* and *tilled field* would both be normalized to *field*); but in other cases this would over-generalize.

9.3 Uninformative superordinates

The most common superordinate is *person* (and its variant *one*, as in ‘One who...’), closely followed by *man*. These provide no real help with classification, since *person/man* senses are distributed pretty evenly across the HTOED taxonomy.

A *person/man* superordinate can be made more specific by extending the ‘scope’ of the superordinate to include the following clause (normalized as outlined above). Some of this has already been attempted, but more work is needed.

9.4 Ontological bias

The weight given to the superordinate within a definition tends to give the classifier an ontological rather than functional bias. That is to say, it tends to classify according to what a thing actually *is*, rather than what a thing does or is used for.

For example, *alum curd* is defined as:

Milk or egg white curdled with alum, used chiefly as a poultice.

This ends up being classified as *the external world » the living world » food and drink » food » dairy produce » milk » curds*. From an ontological point of view, this is perfect (that is exactly what alum curd is). But it overlooks the medical function, which is arguably the more salient aspect here. The HTOED taxonomy tends to be organized from a functional and human-oriented point of view, rather than from a strictly ontological point of view.

9.5 Adjectives

Strictly, the superordinate-based method described above only really applies to noun senses. However, the principle can be extended to certain kinds of adjective sense. In particular, adjectives defined in terms of a noun phrase (introduced with phrases like ‘of or relating to’, ‘designating’, etc.) are susceptible to superordinate-like classification.

For example, *all-in* adj. 2 is defined as:

Designating a form of wrestling with few or no restrictions on the tactics that may be employed; of, relating to, or involved in this kind of wrestling.

Here we can say that *a form of wrestling* is a kind of superordinate, not of the adjective sense itself, but of its nominal equivalent. So we can ‘pretend’ that this is a noun sense with the superordinate *a form of wrestling*, classify it accordingly, and then convert that classification to an equivalent adjective branch.

About 2,500 adjective senses are classified in this way (1.3% of all classified senses).

10. Results and evaluation

Of the 821,000 senses in the OED data set:

- 557,000 (68%) are training senses, i.e. senses that already have at least one HTOED classification;
- 264,000 (32%) are input senses, i.e. senses for which a new HTOED classification is to be computed.

10.1 Output summary

Of the 264,000 input senses processed by the classifier:

- 227,000 (86%) were assigned a classification;
- 25,000 (9.5%) were left unclassified (i.e. the classifier failed to find any classification);
- 12,000 (4.5%) were rejected as intractable.⁸

10.2 Evaluation

The accuracy of the classifier was evaluated by taking a random sample of 1000 senses from the 227,000 senses assigned a classification. For each sense, an evaluator was asked to judge whether the assigned classification was accurate, i.e. represented a valid categorization of the definition.

⁸ These include senses in wordclasses not covered by HTOED (chiefly prepositions, conjunctions, and pronouns); and senses whose definition indicates that they are semantically too vague to be meaningfully classified (e.g. proverb senses, senses with a long list of lemmas, senses defined as ‘miscellaneous’).

Note that this rubric is designed to check for *a* valid categorization, not all possible valid categorizations: see the discussion of multi-part definitions at section 12.2.

Overall, we found that:

- 25% of classifications were accurate, i.e. the correct node of the HTOED taxonomy had been identified;
- 22% of classifications were immediate neighbours of the correct node, i.e. a parent, child, or direct sibling node;
- 18% of classifications were second- or third-generation ancestors of the correct node, i.e. on the correct branch but not specific enough;
- 33% were either straightforwardly incorrect (i.e. on the wrong branch) or were not specific enough to be of any use (i.e. on the right branch, but too high up from the correct node).
- A small residue (<2%) were cases where the evaluator was uncertain of the correct classification (chiefly technical definitions, and obscure undefined compounds).

Only primary classifications were considered; runner-up classifications (see section 2.3) were disregarded.

11. Editing interface

A web-based interface allows results to be reviewed and analysed by a number of different features, including wordclass, sense type (main sense or subentry, defined or undefined), HTOED branch, and principal method of classification:

154	Maastrichtian (noun) s.v. <i>Maastrichtian</i>	→ "The Maastrichtian stage or age. Usu. with the." the earth » structure of the earth » age or period » stratigraphic units » [noun] » secondary or Mesozoic →	nbor
155	maat (noun) s.v. <i>maat</i>	→ "A friend; a companion or mate. (Freq. as a familiar form of address)." the community » social relations » association, fellowship, or companionship » a companion or associate » [noun] →	syns
156	maat (noun) s.v. <i>maat</i>	→ "A formally selected or appointed partner in business or other matters; an agent. Obs. rare." authority » delegated authority » one having delegated or derived authority » [noun] » one who acts for another →	syns
157	maat (noun) s.v. <i>maat</i>	→ "Naut. An assistant officer on a ship; a mate. Now rare." travel » travel by water » one who travels by water or sea » sailor » [noun] » mate →	syns
158	Maatschappij (noun) s.v. <i>Maatschappij</i>	→ "A name given to: an early form of local government among the Boers in the regions that were to become the historical republics of Orangia, the Transvaal, and Natalia." authority » rule or government » a or the system of government » centralized or regionalized systems » [noun] » local government →	supe
159	mab (noun) s.v. <i>mab</i>	→ "= mab-cap." <i>"The ordinary morning headdress of ladies continued to be distinguished by the name of a mab, to almost the end of the reign of George the second."</i> the living world » clothing » types or styles of clothing » headgear » [noun] » cap » types of » worn for specific purpose » indoor cap for women →	eqxr

Figure 2: Editorial interface in review mode

The interface also has an ‘edit mode’ which provides controls for a user to approve, reject, or adjust a classification:

154	Maastrichtian (noun) s.v. <i>Maastrichtian</i>	<p>"The Maastrichtian stage or age. Usu. with the."</p> <p>the earth » structure of the earth » age or period » stratigraphic units » [noun] » secondary or Mesozoic → + (examples: [None] ⚠)</p> <p>✓ ⚙ ✕</p>	nbor
155	maat (noun) s.v. <i>maat</i>	<p>"A friend; a companion or mate. (Freq. as a familiar form of address)."</p> <p>the community » social relations » association, fellowship, or companionship » a companion or associate » [noun] → + (examples: <i>friend, copartner, partner, neighbour</i>)</p> <p>✓ ⚙ ✕</p> <p>emotion or feeling » love » friendliness » [noun] » friend → + (examples: <i>friend, ally, gossip, fellow</i>)</p> <p>✓ ⚙ ✕</p>	syns
156	maat (noun) s.v. <i>maat</i>	<p>"A formally selected or appointed partner in business or other matters; an agent. Obs.rare."</p> <p>authority » delegated authority » one having delegated or derived authority » [noun] » one who acts for another → + (examples: <i>agent, procureur, procurer, proxy</i>)</p> <p>✓ ⚙ ✕</p>	syns

Figure 3: Editorial interface in edit mode

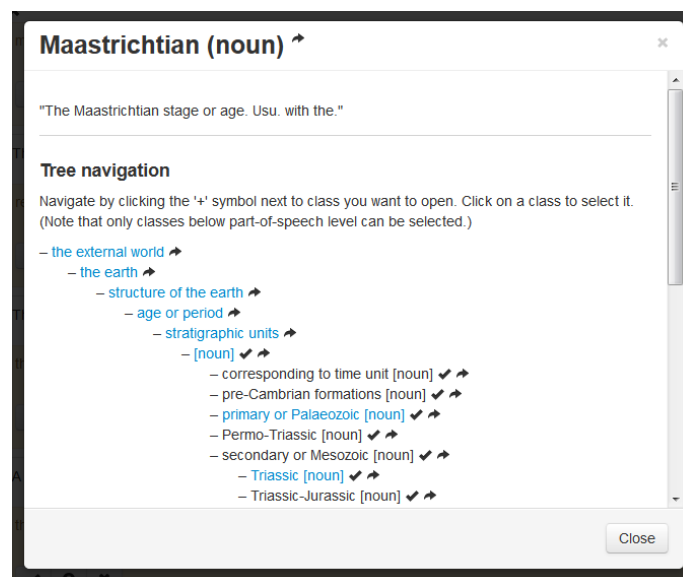


Figure 4: Modal dialogue for adjusting an incorrect classification

We currently have a programme under way to systematically check and approve classifications. Approved classifications are fed back to the source database, becoming part of the training data next time round. This allows for an ongoing iterative process.

12. Limitations and further development

12.1 Taxonomy

A key limitation of the project is that it only attempts to classify input senses in terms of the existing HTOED taxonomy; it does not suggest or create new categories. This means that often there is no correct node to which a given input sense could be assigned: the ‘bottom-up’ construction of the taxonomy means that it is shaped by the existing OED content, with no provision for new senses representing new concepts.

12.2 Multi-part definitions

In general, the classifier treats each input sense as atomic: that is to say, it assumes that a single sense represents a single coherent meaning or usage.

In reality this assumption is flawed, because many individual senses can be decomposed into two or three distinct meanings. Indeed, the original editors of HTOED routinely interpreted OED senses in this way, and so many training-data senses have multiple HTOED classifications.

But the multiple meanings within a single sense can be signalled in more or less explicit ways, and can be hard to distinguish from single-meaning senses. For example, the definition of *scene queen* has two quite different meanings presented as semicolon-separated clauses:

A woman who is prominent in a particular scene, esp. a particular music scene; (esp. in gay usage) a homosexual man who goes to gay bars, clubs, etc...

The definition of *overpower* v. 3 also has semicolon-separated clauses; but here these are really just restatements or nuances of the same core meaning:

Of an emotion, fatigue, etc.: to overcome (a person, etc.) by intensity; to be too much or too intense for; to overwhelm.

It is very hard to define formally what differentiates the *scene queen*-type multi-part definition from the *overpower*-type single-sense definition.

We allow the classifier to treat certain input senses as having multiple meanings (and therefore to assign multiple HTOED classifications), where this is unambiguous; but the default approach of treating each input sense as atomic means that the assigned classification often fails to reflect the semantic range indicated by the definition.

12.3 Gloss and metalanguage

OED definitions consist broadly of two kinds of material:

- semantic gloss;
- metalanguage: various forms of grammatical, contextual, and usage information.

For the purposes of HTOED classification, the metalanguage is usually redundant, and is best jettisoned so that the classifier can focus on the gloss. This is a necessary first step for many of the analytic strategies described above. If metalanguage is confused for gloss, or vice versa, this can cause some significant problems.

In practice, separating gloss from metalanguage can be difficult, since OED definitions do not explicitly demarcate them.

Certain known patterns can be tested, for example, metalanguage often precedes and/or follows the gloss as separate sentences (sometimes bracketed). For example, in *mash* n. 3b:

(Without article.) The state of being mashed or reduced to a soft pulp. Chiefly in *to beat* (*also boil, etc.*) *to mash*. Also in extended use.

The gloss is *the state of being mashed or reduced to a soft pulp*; the preceding and following sentences are metalanguage which can be discarded.

But gloss and metalanguage are often more fluidly integrated, making automatic separation more difficult. For example, *club-ball* is defined as:

A term applied by Strutt and subsequent writers to games in which a ball is struck by a club or bat, esp. to the earlier types of these.

where the definition proper is *game[s] in which a ball is struck by a club or bat*, and the rest is metalanguage. But the classifier currently misconstrues *a term applied by...* as the start of the definition proper; this leads to the sense being misclassified.

There is no magic-bullet solution to the general problem of separating gloss from metalanguage. Really, it is just a matter of trying to account for more and more patterns as they are observed; this gradually improves performance, but is unlikely ever to be exhaustive.

12.4 Identifying the main sense of a word

When analysing a sense, a typical task that the classifier needs to perform is to find

the meaning of certain keywords within the definition, e.g. a superordinate or synonym term. For example, in the definition *Stocks or shares in a mining company*, we need to be able to determine the sense in which *stocks* and *shares* are being used.

When a word appears in a definition, particularly as a synonym, the default assumption is that the word is being used in its primary modern sense. Although not impossible, it is unusual for an OED definition to use a word in an obscure, historical, figurative, dialect, or slang sense – at least not without some explicit indication.

Hence, to analyse a definition effectively, any system needs to be able to:

1. identify the primary modern sense of a word, as given in OED;
2. determine when the default assumption does not apply, i.e. when there is some indication that the word is being used in a different sense.

The first is an interesting problem in its own right, given that OED lists senses in chronological order, rather than by frequency or prototypicality. There are several promising approaches to this, both internal (evaluating the structure and relative significance of senses within the entry) and external (comparing senses in the OED entry with the corresponding entry in dictionaries which *do* rank senses by prototypicality). But these are not altogether reliable.

The second task – determining when the main-sense assumption does not apply – is handled by looking for explicit markers (e.g. the word in question is followed by a cross-reference pointing to a particular sense of that word); or by testing if the topic of the sense as a whole suggests a more technical sense of a given word within the definition. For example, *prosiphon* is defined as:

The primitive siphon in an embryonic ammonoid, consisting of a kind of ligament attached to the protoconch.

Here the *Topic_classifier* establishes that the sense as a whole is zoological; so in analyzing the superordinate *siphon*, the classifier is able to prefer the specifically zoological sense of *siphon* over the more general main sense.

As these examples suggest, there is no attempt to perform full word-sense disambiguation of terms in definitions. Instead, a more primitive default/exception model is employed: by default, a term is assumed to be used in its main sense, unless the contextual evidence suggests that something else may be preferred.

12.5 External methods

All the methods discussed so far are internal methods, to the extent that they only draw on data from within OED and HTOED.

It is also worth considering what other resources could be brought to bear on the problem, especially resources that deal in hypernymy (e.g. Wordnet) or synonymy (e.g. Wiktionary). In general, external resources are of limited value because of the rarefied nature of OED content: most OED lexemes and senses do not appear in other lexical resources, and this is even more true of the input senses considered here. Still, for those OED terms which *do* appear in a resource like Wordnet or Wiktionary, these may provide more direct evidence for a classifier.

13. Acknowledgements

Thanks to Kate Wild, Andrew Ball, Liz Ashdowne and Michael Proffitt (all at OED) for reviewing this project at each stage of development, and for numerous valuable suggestions.

14. References

- Alexander, M. & Bramwell, E. (2012). Mapping Metaphors of Wealth and Want: A Digital Approach. In Mills, C., Pidd, M. & Ward, E. (eds.) *Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities*. Sheffield: HRI Online Publications, 2014. Available online at: <http://www.hrionline.ac.uk/openbook/chapter/dhc2012-alexander>
- Bauer, L. (2009). Typology of compounding. In Lieber, R. & Štekauer, P. (eds.) *The Oxford Handbook of Compounding*. Oxford: Oxford University Press, pp. 343-356.
- Booij, G. (2007). *The Grammar of Words: An Introduction to Linguistic Morphology*. 2nd edition. Oxford: Oxford University Press.
- Crystal, D. (2014). *Words in Time and Place: Exploring Language Through the Historical Thesaurus of the Oxford English Dictionary*. Oxford: Oxford University Press.
- Kay, C. & Wotherspoon, I. (2002). Turning the dictionary inside out: some issues in the compilation of a historical thesaurus. In J. E. Diaz Vera (ed.) *A Changing World of Words: Studies in English Historical Semantics and Lexis*. Amsterdam: Rodopi, pp. 109-135.
- Kay, C., Roberts, J., Samuels, M. & Wotherspoon, W. (2009). *Historical Thesaurus of the Oxford English Dictionary: With additional material from A Thesaurus of Old English*. Oxford: Oxford University Press.
- Levin, B. & Hovav, M. R. (1998). Morphology and lexical semantics. In Spencer, A. & Zwicky, A. (eds.) *Handbook of Morphology*. Oxford: Blackwell, pp. 248-271.
- Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge: Cambridge University Press.
- Mooney, R. J. (2005). Machine Learning. In Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, pp. 376-394.
- Murphy, M. L. (2003). *Semantic Relations and the Lexicon: Antonymy, Synonymy, and other Paradigms*. Cambridge: Cambridge University Press.

Roberts, J. & Kay, C. (1995). *A Thesaurus of Old English*. London: King's College London Medieval Studies XI.

Taylor, J. (2003). *Linguistic Categorization*. 3rd edition. Oxford: Oxford University Press.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



What is a Target Language in an Electronic Dictionary?

Anna Helga Hannesdóttir

University of Gothenburg, Department of Swedish, Box 405, SE-405 30 Gothenburg

E-mail: anna.hannesdottir@svenska.gu.se

Abstract

In a printed bilingual dictionary, one of the languages acts as the source language and the other the target language. In an electronic dictionary, where both languages can be made equally accessible, the relationship between the two languages is much more complicated. This paper will discuss the consequences of this multiple access in bilingual lexicography. The focus will also be on the target language vocabulary, when it is made as accessible as the source language. The point of departure is the Swedish vocabulary presented in the multilingual online-only resource *ISLEX*, where Icelandic is the source language and Swedish one of the target languages. While the Icelandic vocabulary in *ISLEX* is carefully selected and representative of the Icelandic lexicon, the Swedish vocabulary consists of a rather arbitrary selection of the Swedish lexicon, revealing unfortunate equivalent lacunae, i.e. the absence of words of frequent occurrence and central to colloquial Swedish. Some implications of multiple access for the typology of bilingual dictionaries will be discussed.

Keywords: bilingual e-lexicography; multiple access; source/target language; equivalent lacunae; dictionary typology

1. Introduction

In a printed bilingual dictionary, the function of the two languages is clear: one acts as the source language (SL) and the other the target language (TL). The TL is in all aspects subordinate to the SL. This is the case for the TL vocabulary provided in the dictionary, the examples given to illustrate the usage of the headword, collocations, idioms etc. There are no TL units in the dictionary that are not motivated by specific qualities of the SL and all information about the TL is accessed only through the SL. While the lexicographic description necessarily takes either of the two languages in question as a point of departure for the information provided, an electronic dictionary can offer the user equal access to units of both languages. For the user, the function of the two languages is not as clear-cut as in the printed dictionary since the distinction between the SL and the TL is partly neutralized. The TL occurs as a lexical component in its own right. This has changed the very basis of the bilingual lexicography.

This paper first discusses some of the differences between printed and online bilingual dictionaries, focusing on the concepts of source language and target language. Then the multilingual *ISLEX* online-only resource is presented and the Icelandic and Swedish vocabularies, respectively, are described. One consequence of the accessibility of the target language for bilingual lexicography is the equivalent lacunae occurring in

the Swedish vocabulary in *ISLEX*. The typology of bilingual dictionaries is also discussed and modified.

2. Bilingual dictionaries on the Internet

In a printed, bilingual dictionary, the lemma selection and the description of the lemmas and equivalents are adjusted to a well-defined user group. The users are taken to be *either* mother tongue (L1) speakers of the SL, using the dictionary for encoding tasks, *or* mother tongue speakers of the TL, using the dictionary for decoding texts in the foreign SL (Figure 1). The L1 users are expected to have good knowledge of their mother tongue, while their skills in the foreign language (L2) are taken to be insufficient. The description of the source language is adapted to the users' skills and needs, and so is the description of the equivalents. It is, of course, the L2 that is provided with an elaborated description, adjusted to the role as the source or target language, respectively.

Source Language in relation to user's mother tongue		Target Language in relation to user's mother tongue	User's activity
L1	>	L2	encoding
L2	>	L1	decoding

Figure 1: The functions of the languages in the dictionary, related to the user's mother tongue and activity

Many of the bilingual dictionaries now available on the internet are simply digitalized versions of existing printed dictionaries, i.e. *p-dictionaries* rather than *e-dictionaries* (Fuertes-Olivera & Bergenholtz, 2011), and are thus subjected to the same restrictions in accessibility as their printed predecessors. In dictionaries conceived and edited as an online-only resource, the material in the dictionary database can be accessed in far more elaborated ways, which makes the relationship between the two languages much more complex than it is in a printed dictionary. Both of the languages can be made mutually accessible, and both can serve L1 and L2 users alike. Users consulting the dictionary for decoding a text in L2 need a comprehensive set of words and fixed phrases in that language, while for encoding tasks they also need elaborated information regarding the morphological, syntactic and pragmatic features of the L2 units.

In order to fulfil the needs of L1 and L2 users alike, both languages in a bilingual e-dictionary should provide a comprehensive stock of lexical units, as well as a detailed description of these units. This entails a theoretical as well as methodological challenge for the bilingual e-lexicography regarding the coverage and description of both languages.

3. Source Language and Target Language

One aspect of the multiple accessibility of the target language in an e-dictionary is the target language itself. While the subset of the source language lexicon presented in a bilingual dictionary is carefully selected, the target language representation is subordinate and reactive to the source language. In the printed dictionary, the target language only appears as an answer to a query concerning a source language unit, and the target language features are focused upon only in relation to that specific source language unit. The inevitable lemma lacunae, i.e. SL units absent in the stock of lemmas, are due either to the lexicographer's rational consideration, estimating these lemmas as too peripheral or special to be included in that particular dictionary, or unintentionally caused by random lapses of the lexicographer. The lemma lacunae rarely affect a complete structurally defined, coherent subgroup of the lexicon.

When the target language is also accessible, a new lexicographic phenomenon emerges, i.e. the equivalent lacunae. Unlike the lemma lacunae, the equivalent lacunae can be extensive and they can affect a clearly definable subset of the lexicon. When all the lexical information presented in both of the languages can be accessed, the dichotomy between the source language and the target language is technically neutralized. This raises the question asked in the title: what is a target language in an electronic dictionary? As will be illustrated below, multiple and equal access to the two languages featuring in a bilingual electronic dictionary results in great demands on new theories and new methodology in bilingual lexicography.

4. The *ISLEX* Dictionaries

The multilingual *ISLEX* e-dictionaries were launched on the internet in November 2011. The source language is Icelandic and the mainland Scandinavian languages Danish, Norwegian Bokmål, Norwegian Nynorsk and Swedish are the target languages. Recently, Faroese was added as a target language, and the compilation of an Icelandic–Finnish version is now in progress. All the languages treated in the *ISLEX* dictionaries can be considered as “small” languages, varying from 50,000 speakers of Faroese and 320,000 Icelandic speakers to 8,500,000 speakers of Swedish. Hence, as is often the case with bilingual dictionaries of “small” languages, the main objective of the *ISLEX* dictionaries is to serve as many users in as many linguistic activities as possible. All the Icelandic material in *ISLEX*, i.e. lemmas, examples, fixed phrases and idioms, is provided with equivalents, paraphrastic explanations or translations into the Scandinavian languages. The *ISLEX* project, including its technical aspects, has been presented at several international conferences, e.g. EURALEX 2008 (Sigurðardóttir et al., 2008) and LREC 2014 (Úlfarsdóttir, 2014).

The Icelandic editors at the University of Iceland were in charge of the overall planning and management of the project. The Scandinavian partners were The Society for Danish Language and Literature in Copenhagen, The University of Bergen, Norway and The University of Gothenburg, Sweden. From the outset, *ISLEX*

was planned as an online-only resource, and the opportunities offered by the electronic technique were well utilized in the planning, editing and development of the dictionary. The ISLEX content is set in an object-relational database, which was designed, developed and is now being maintained, and also elaborated further, in Iceland. The editorial environment of the dictionary and the user interface were also designed in Iceland.

From this database alone, different dictionaries are now generated. They are published online and can all be accessed free of charge. The website addresses, www.islex.is, www.islex.dk, www.islex.fo, www.islex.no, www.islex.se, respectively, lead to the homepages of the individual dictionary. The meta-language shown in the entries is determined by the country suffix, which means that [islex.dk](http://www.islex.dk) generates Danish, [islex.se](http://www.islex.se) Swedish, etc. Also, the language constellation offered initially in the search process is generated by the suffix, [.dk](http://www.islex.dk) leads to the Icelandic–Danish dictionary. The users can, however, easily change both the meta-language and the language combination and they can also view all the target languages simultaneously (Figure 2). The dictionaries have been very well received by the target user groups (Úlfarsdóttir, 2014) as well as by reviewers (Sanders, 2013).

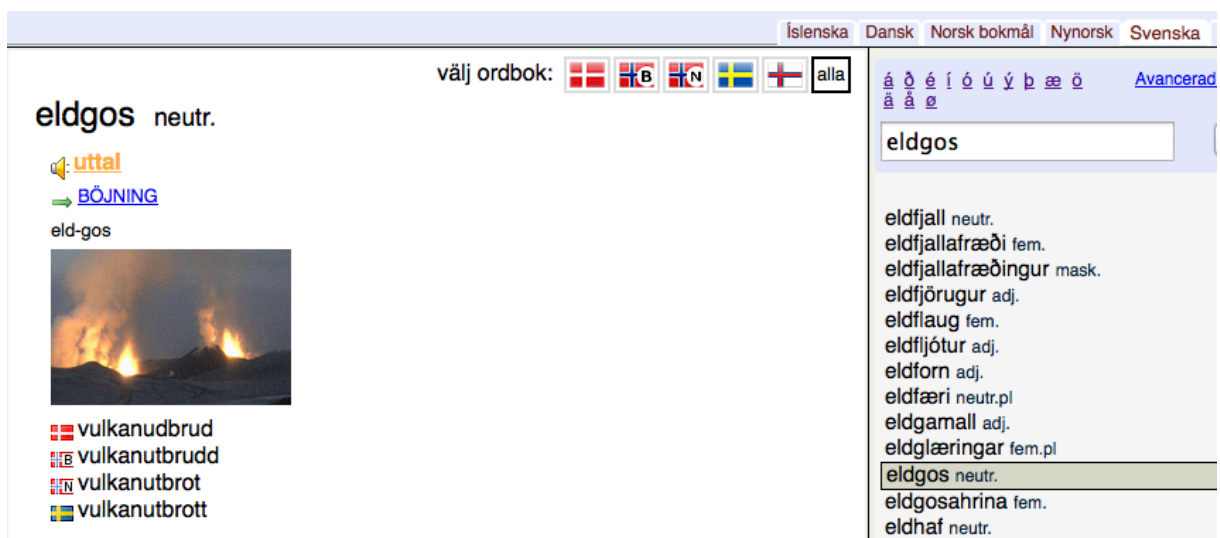


Figure 2: The result of the query for the lemma *eldgos* ('volcanic eruption') with equivalents in Danish, Swedish and the two Norwegian varieties

Icelandic is, however, always one of the languages offered to the users, more precisely in the capacity of source language.

In the *ISLEX* dictionaries, the multiple search options offered by the electronic technology are well employed. The user can search not only for the Icelandic lemmas but also, by using the free text search, for all other Icelandic lexical units and strings of text occurring in the dictionary. Also, the equivalents can be searched out, as well as every word or string of text, occurring in the translations of the Icelandic material.

Technically, the *ISLEX* dictionaries are thus not only bi- or multilingual but also biscopal or bidirectional, since both languages are equally accessible.

Another objective of the *ISLEX* project is that the dictionaries should be multifunctional, i.e. they are supposed to serve Icelandic users as well as the Scandinavian ones, in decoding and encoding activities alike. In terms of traditional bilingual lexicography, and in the ways the dictionaries were edited, Icelandic is the source language and the point of departure for the lexical description of the Scandinavian languages. The lexicographic representation of each of the Scandinavian languages is therefore subordinate to the Icelandic material, since it is the Icelandic headword that is provided with equivalents or paraphrased. The same goes for the fixed phrases and idioms. Although all the examples of usage and the fixed phrases are presented in all the languages, the Scandinavian versions are translations of the Icelandic ones, which in turn are intended to illustrate language specific features of the Icelandic lemma rather than illustrating contrastive aspects of the languages in question.

The established notions of *source* language and *target* language should be reconsidered and distinguished with respect to the lexicographic perspective on the one hand and the user perspective on the other. In the case of *ISLEX*, the lexicographic status of Icelandic is that of a *main* language, since it makes the basis of the lexicographic description also of the Danish, Norwegian and Swedish languages. The lexicographic status of these languages is therefore *subordinate* to Icelandic – the user’s activities left aside. From the user’s point of view, the Scandinavian material is just as accessible as the Icelandic material. The *search* (rather than *source*) language can thus be one of the Scandinavian languages as well as Icelandic. Depending on the user’s lexicographic activities, decoding or encoding text, and depending on which of the languages is his or her mother tongue, the *search* language can be L1 or L2. To emphasize the distinction between the lexicographic perspective and the user perspective, I will here use *main language* (ML) referring to Icelandic and *subordinate language* (SuL) referring to a Scandinavian language in a lexicographic perspective. When the user perspective is in focus I will use *search language* and *target language* respectively. The abbreviations SL and TL will henceforth relate to the user perspective only, standing for *search language* vs. *target language*.

ISLEX is primarily intended to support the Icelandic users in (1) expressing themselves in a Scandinavian language, i.e. for encoding purposes. Icelandic is then the SL and the users L1 while the TL is their L2 (ML/SL/L1>TL/L2). The Icelandic users are also supported in (2) decoding texts in any of the Scandinavian languages presented in the dictionary, by looking up an SL unit in L2 in order to find an Icelandic TL unit (ML/TL/L1<SL/L2). Furthermore, the dictionary is intended to serve Scandinavian users in (3) decoding Icelandic texts (ML/SL/L2>TL/L1) and – with certain reservations – in (4) producing texts in Icelandic (ML/TL/L2<SL/L1) as illustrated in Figure 3. The angle bracket illustrates the direction of the search in

relation to the user’s mother tongue.

User’s L1	User’s activity	ML in relation to user	Search direction related to user’s mother tongue	SuL in relation to user
1 Icelandic	Encoding	L1	>	L2
2 Icelandic	Decoding	L1	<	L2
3 Dan/Nor/Sw	Decoding	L2	>	L1
4 Dan/Nor/Sw	Encoding	L2	<	L1

Figure 3: In the electronic dictionary, the main language and the subordinate language are equally accessible, ML as well as SuL is L1 to some users and L2 to others and ML and SuL alike are consulted in encoding as well as decoding activities.

Henceforth, I will focus on the Icelandic–Swedish dictionary in *ISLEX*, i.e. *islex.se*. The Icelandic users consulting *islex.se* for decoding a text in Swedish should need a comprehensive Swedish vocabulary; single word units as well as fixed phrases. When consulting the dictionary for encoding tasks, users will also need elaborated information regarding the morphological, syntactic and pragmatic features as well as the selectional restrictions and constructional preferences of the Swedish units. The Swedish user has the same needs but the other way around, i.e. an extensive Icelandic lemma list for decoding Icelandic texts and generous information regarding the formal features of the Icelandic units for encoding tasks. Adjusting the lexical description of each of the languages to the needs of an L2 user, the description of both languages runs the risk of suffering from a rather heavy overload of information, at least from the L1 user’s point of view. That problem is, indeed, a technical as well as a lexicographic one.

5. The Icelandic Vocabulary in *ISLEX*

Icelandic is the point of departure for the lexical description of Swedish as well as for all the other languages in *ISLEX*. The entries are based on an Icelandic lemma, which in turn can be a single- or multi-word unit. The lemma is completed with adequate information regarding its grammatical, syntactic, phraseological etc. features. Recorded pronunciation of the headwords, single word units as well as multi-words units, is also added.

The Icelandic material in *ISLEX* consists of ca. 50,000 lemmas, 30,000 exemplifying sentences and 14,000 collocations, idioms and fixed phrases of different kinds (Úlfarsdóttir, 2013). All this material is carefully selected with respect to adequacy and representativeness in relation to the Icelandic lexicon and to the manifold

objectives of the dictionary. The emphasis lies on the modern Icelandic lexicon and a great many of the lemmas make their very first dictionary appearance in *ISLEX*. Words denoting culture-specific phenomena in Iceland of today as well as some words central to the medieval Icelandic saga literature are also included. Thus, words such as *æðarvarp* ‘area where eider ducks nest’, *þorramatur* ‘traditional Icelandic late winter food’ and *landnámsöld* ‘Age of Settlement’ are lexical entries in *ISLEX* (the English translations are given in Hólmarsson, Sanders & Tucker (1989), s.v. *æðarvarp*, *þorramatur* and *landnámsöld*). The same applies to a number of words denoting parts of the Icelandic traditional women’s costume, traditional Icelandic food and other folkloristic phenomena. There is, similarly, a number of words denoting the traditional or typical Icelandic professions farming and fishing. Also, the vocabulary related to the Icelandic landscape with volcanoes, lava fields and glaciers is included, and neologisms, words and phrases related to the Icelandic banking collapse in 2008 are also added. Albeit far from a complete coverage of the Icelandic vocabulary, systematic, unintentional lemma lacunae are not to be expected in *ISLEX*.

6. The Swedish Vocabulary in *islex.se*

The Swedish vocabulary is, unlike the Icelandic one, not the result of a carefully conducted and well-conceived selection process. While there are ca. 50,000 Icelandic lemmas in *ISLEX*, the number of unique Swedish equivalents in *islex.se* amounts to ca. 41,000 (Úlfarsdóttir, 2013). As can be expected, these 41,000 equivalents constitute a somewhat arbitrary selection of the Swedish lexicon. Not only is the coverage of the Swedish lexicon inferior to the coverage of the Icelandic one in numbers of lexical items, but the degree of representativeness in terms of basic words among these 41,000 is also rather insufficient compared to the number of Icelandic lemmas. In a printed dictionary, neither the number nor the representativeness of the target language is a problem – the number of unique equivalents has not yet become a sales argument like that of the input lemmas.

One reason for the quantitative discrepancy regarding Icelandic lemmas and the Swedish equivalents lies in the structural differences in the lexical systems of the two languages. These differences are reinforced by the lexicographic status of the languages, Icelandic being the point of departure for the description of the Swedish language, rather than because of accessibility, whereby Icelandic is the source language and Swedish the target language. That distinction is indeed neutralized in the e-dictionary with multiple search options. However, Icelandic is the language that conducts the lexical description of the Swedish language. There is no incentive for the Swedish lexicographer to insert Swedish words or phrases unless they are triggered by the Icelandic units or by phrases illustrating the use of these units. This imbalance results in a considerable amount of what can be labelled as equivalent lacunae, i.e. TL words – in this case Swedish words, which – unlike the case in the printed dictionary – actually were directly accessible if they only were included in the *ISLEX*

dictionary.

Two types of systematic equivalent lacunae will be discussed below. One of these is due to discrepancies in word formation strategies in the two languages, the other type is due to the very subject of *ISLEX*, namely the Icelandic language, nature, culture and society – not the Swedish language, nature, culture and society.

6.1 The Swedish *-era/-iera* Verbs as Equivalents in *islex.se*

One systematic difference between Icelandic and Swedish concerns the policy towards loanwords. In Swedish there is a generous attitude towards loanwords, and a significant part of the lexicon consists of words and word formation elements of West-Germanic or Greco-Romance loans. In Icelandic, on the other hand, the modern international vocabulary, based on Greco-Romance elements, is scarce and there is a reluctance to include such words in Icelandic (Vikør, 1993: 211). Also Greco-Romance prefixes like *in-*, *multi-*, *re-*, *un-* and the like are seldom used in Icelandic word formation, while they are incorporated in the productive material in Swedish. The same goes for the suffixes, *-tion*, *-era* etc., originating in the classic languages and productive in the Swedish word formation system. The Swedish reverse dictionary (Allén & Sjögreen, 2007) contains 2038 Swedish verbs derived from Greco-Romance stems through any of the suffix variants *-era*, *-iera*, *-fiera*, *-ficera* etc. (Hannesdóttir, 2014). Of these 2038 verbs, 1071 are included in the largest printed Swedish–Icelandic dictionary (*Svensk-isländsk ordbok*, 1983). In this dictionary of 60,000 lemmas, where Swedish is the source language, the stock of lemmas is composed with the same users in mind as *islex.se*, i.e. Swedes and Icelanders. It is also intended to be multifunctional and serve Icelanders as a decoding dictionary and Swedes as an encoding dictionary. Of the 1071 verbs included in this Swedish–Icelandic dictionary, 360 occur as equivalents in *islex.se*. Quite a great number of the verbs in the reverse dictionary, as well as those in the Swedish–Icelandic dictionary, are rather peripheral in the Swedish lexicon as such. Many of the verbs are, however, of frequent occurrence and central to the colloquial Swedish of today.

A more relevant object of comparison regarding the Swedish lexicon of today is the Swedish lemma stock of the bilingual learning dictionaries in the *Lexin* project, a series of dictionaries between Swedish and the languages of some of the largest immigrant groups in Sweden. The bilingual dictionaries are based on the printed monolingual Swedish dictionary *Svenska ord* (1984; 1992; 1995). In 2011, the fourth edition of the Swedish dictionary was launched online. The material in *Svenska ord* is the point of departure for selecting the Swedish lemmas and their lexicographic description for all the bilingual dictionaries. The database contains ca. 28,000 lemmas (Hult et al., 2010). Today there are 15 different *Lexin* dictionaries available online while another five dictionaries are available only in printed form. As presented on the homepage of *Lexin*, the dictionaries are specially adapted for use in the teaching of Swedish as a second language. They therefore contain only the most common Swedish

words. Swedish is the source language in the early printed dictionaries and it is still the basis for the target language description as new dictionaries between Swedish and the languages of new immigrant groups are edited and appearing as online-only resources. In these dictionaries, as well as the ones that have been digitalized and published online, both languages, i.e. the lemmas and the equivalents, are equally accessible.

It appears that a number of the 700 *-era/-iera* verbs that do not occur as Swedish equivalents in *islex.se* are included as Swedish lemmas in *Lexin*. Among those we find *associëra* ‘associate’, *devalvera* ‘devalue’, *figurera* ‘appear, figure’, *fingera* ‘simulate’, *fixera* ‘fix, determine’, *imponera* ‘impress’, *initiera* ‘initiate’, *koncentrera* ‘concentrate’, *konversera* ‘converse’, *moralisera* ‘moralize’, *precisera* ‘specify, clarify’, *ruinera* ‘ruin, destroy’ and *socialisera* ‘socialize’, and a fair number of other verbs. *Lexin* is considerably smaller than *ISLEX* but explicitly concentrates on the most common and basic words in Swedish.

Equivalent lacunae as those in *islex.se* are significant when a target language has been made just as accessible as the source language. All the verbs mentioned here are included as lemmas in the somewhat larger Swedish–Icelandic bilingual dictionary, aimed at the same user groups as *islex.se*. They should definitely, one way or another, be included in the Swedish vocabulary presented in *ISLEX*.

6.2 The Swedish *-era/-iera* Verbs Occurring at Free Text Search in *islex.se*

A free text search through the Swedish material in *islex.se* for *-era/-iera* verbs occurring in the translations of examples and other illustrative material but not as equivalents, gives another 132 verbs in addition to the 360 (Hannesdóttir, 2014). Even if lexical items occurring only in the Swedish translations lack information regarding the morphological features added to the Swedish equivalents, the presence of them in the translations is far better than no occurrence at all.

The total of almost 500 *-era/-iera* verbs in *islex.se* is still less than half the number of such verbs listed in the printed, larger, Swedish–Icelandic dictionary. When the lexical systems of two languages are confronted in the way they are in the bilingual dictionary, the discrepancies with respect to the way various concepts become crystallized, established, denoted and lexicalized in the two languages in question become clear. The verbs discussed here all share the semantic feature of denoting highly abstract actions. They all represent concepts so well established in the Swedish speech community that they have become lexicalized in form of a single word. The absence of a lexical representation of these concepts in the Icelandic lemma list in *ISLEX*, might partly be due to the word formation strategies of Icelandic, blocking loanwords of this kind and preferring domestic derivational suffixes to Greek and

Latin ones. The denotations of concepts, if established at all, might therefore be lexicalized in form of multiword units and phrases rather than single words (Hannedóttir, 2014). Of the 13 above-mentioned *-era/-iera* verbs, present in *Lexin* but absent as equivalents in *islex.se*, only three occur in free text search through the translations of Icelandic phrases or examples: *imponera*, *koncentrera* and *konversera*. Six of the word stems can be recognized in participles or nouns, as e.g. *fixerad* ‘fixed’, *moraliserande* ‘moralizing’ and *precision*.

6.3 Culture Specific Words in *islex.se*

As aforementioned, the Icelandic society and culture is the subject of description in the *ISLEX* dictionaries. While the coverage of the culture specific, Icelandic vocabulary is quite sufficient for the decoding Swedish users, the number of Swedish culture specific words occurring as equivalents is rather poor. These words denote concepts that are not established and therefore not lexicalized in Icelandic.

A significant number of words denoting Swedish food and feasts lacking in *islex.se* are treated in *Lexin*, such as e.g. *kräftskiva* ‘crayfish party’, *nypon* and *nyponsoppa* ‘roship’ and ‘roship soup’, *surströmming* ‘fermented Baltic herring’ and *kavring* ‘dark, sweetened rye bread’. The printed Swedish–Icelandic dictionary includes four of these five words, i.e. all those mentioned except *kavring*. In *Lexin* we also find words related to the Samic culture: *sametinget* ‘the Sami Parliament’, *samekultur* ‘Sami culture’ and *renhjörd* ‘reindeer herd’. This field is poorly represented not only in *islex.se* but also in the printed Swedish–Icelandic dictionary. The few words that actually are included as lemmas or sublemmas in the printed dictionary are compounds with the *Lapp* element rather than *Same*: *lappdräkt* ‘Samic costume’ etc.

Words for common Swedish phenomena absent in *islex.se* but included in *Lexin* as well as the Swedish–Icelandic dictionary are e.g. *semestra* ‘spend one’s holiday’, *sommargäst* ‘holiday visitor’, *vinterbona* ‘prepare for winter conditions’, *hötorgskonst* ‘kitsch art’, *kullersten* ‘cobblestones’ and *bostadskö* ‘housing queue’ (the English equivalents and paraphrastic explanations from www.ne.se/ordböcker).

Words such as these are common and frequent Swedish words, likely to turn up in Swedish texts and they should definitely be among the Swedish words presented in a bilingual, bidirectional and multifunctional dictionary such as *islex.se*.

7. Consequences of Multiple Accessibility for the Bilingual Lexicography

The entire process of dictionary making – bilingual as well as monolingual – has been revolutionized by the computerization of the process and the alternative digital publication forms. The discussion concerning how lexicography has benefitted from technological developments is dominated by the monolingual perspective, and not much has been said regarding bilingual e-dictionaries. However, many of the points at

issue concern general features in mono- and bilingual lexicography alike. Thus, the advantages brought about by the technical improvement have facilitated the lemma selection process and the selection of good examples; these moments are now based on large corpora and powerful search tools (Kilgarriff et al., 2008; Trap-Jensen, 2013). The scantiness in the description of the semantic, pragmatic, morphological etc. features of the lexical units are no longer called upon since the space is not the same issue in the electronic format as it is in the printed dictionary. And the lexicographer's work does not necessarily concern one specific lexicographic product but rather a database from which a number of dictionaries can be produced with a number of alternatives regarding presentation and visualization of data. The opportunities offered by the rapid technological developments are far from being utilized optimally. One main problem is that the lexicographers have not kept pace with the opportunities offered by the technological progress.

The reversal of bilingual dictionaries has been at stake for quite some time. The reversal projects hitherto reported in the lexicographic literature, first and foremost aim at printed dictionaries (i.e. the OMBI project: Maks, 2007; Martin, 1996; 2007). In bilingual e-dictionaries, where all the material in both languages is made equally accessible, some new criteria should be taken into consideration already in the planning phase of the project. In order to avoid massive equivalent lacunae of the kind discussed above, the point of departure must be a representative selection of not only the main language but of the units representing the subordinate language too. Also the selection of examples should be "chosen entirely on the basis of their translations" (Atkins & Rundell, 2008: 507). The examples must be contrastively sound, not only in order to avoid causing problems of ambiguity in one of the languages but also, as far as possible, focusing the deviations in usage in the two languages.

The ISLEX database maintains high technical standards. It was, from the outset, designed as an online-only resource. The software solutions chosen at the beginning of the project are flexible and, from the editorial point of view, well adapted to its purpose. The different fields, defined for the different types of data categories designed for the Scandinavian languages, can be expanded, added or omitted at the discretion of each one of the Scandinavian lexicographers. As is often the case at the planning stage of a dictionary project, there were more questions than answers, and there are certainly some shortcomings of the dictionaries. Some, e.g. the equivalent lacunae discussed above, can be attributed to specific linguistic features. Others should rather be ascribed to the theoretical aspects of bilingual lexicography as it developed from the late 20th century, based on lexicographic practice established during centuries of bilingual dictionaries being published in printed form. The roles of the languages involved were then given once and for all as illustrated in Figure 1.

First and foremost we were not aware of what impact multiple accessibility would have on the dictionaries. Actually, the question of access to the data and presentation

alternatives was not at stake until quite late in the editing process. The lexical description of the material in *islex.se* is strongly based on the theory of the different functions of the two languages included in a bilingual dictionary; one being the source language and the other the target language. This distinction is consequently based on directionality and accessibility being restricted to one of the languages and it is, as well as the terms themselves, outdated in the bilingual e-dictionary. Here, I have used the terms *main* language vs. *subordinate* language, focusing on the criteria for lexicographic description rather than the access criteria. However accessible, the representation of the subordinate language will in many respects depend on the main language. This calls for methodological development of the bilingual lexicography.

What is now provided by *ISLEX* is an efficient and well structured database and an adequate lexicographic description of the Icelandic lemmas. The selection of the Icelandic material is strictly language specific, i.e. neither the lemmas nor the examples are selected considering the contrastive aspects actualized in the bilingual dictionary. It should be borne in mind, however, that *ISLEX* is conceived not as a bilingual but as a multilingual dictionary. One and the same Icelandic material in the *ISLEX* database is intended to provide a representative basis for bilingual dictionaries between Icelandic and a number of other languages. Technically, the *ISLEX* e-dictionaries make good use of many of the technical possibilities offered by computer science and language technology. From a lexicographic point of view, it is indeed made by the book on bilingual lexicography. The problem is that the traditional view on bilingual lexicography is long since outdated.

8. Conclusions

What then is the target language of an electronic dictionary? In terms of accessibility, the distinction between source language and target language should not be relevant at all. As discussed in this paper, both languages in the bilingual e-dictionary can be equally accessible. In terms of lexicographic status on the other hand, it still seems suitable that one of the languages is made the point of departure for the lexicographic description. As the lexical description does not have to do with accessibility, I have chosen to use the term *main* language rather than *source* language. The real challenge for bilingual e-lexicography is to develop methods for an adequate description of the language subordinated to the main language, a description where a suitable stock of lemmas is presented and the grammatical, semantic, combinatorial and pragmatic features of these lemmas are accounted for. The description of the Swedish language in *islex.se* is not yet there.

What has become obvious reviewing the process of editing *ISLEX* as well as the resulting product itself is that the theories and methods of bilingual lexicography do not keep up with the development in computer science. The lexicographers must loosen their grip on several traditional notions established long ago. In particular, the lexical description of the languages should be based on the multiple accessibility at

hand in e-dictionaries rather than on the restricted accessibility of printed dictionaries. Much more information is available in e-dictionaries, and the creative user looks up whatever we generously make accessible. We must take the consequences of our generosity by furnishing the lexicographic material offered with as much relevant information as possible, whether the user is a speaker of the main language or of the subordinate language.

9. References

- Allén, S. & Sjögren, Ch. (2007). *Norstedts svenska baklängesordbok*. Stockholm: Norstedts Akademiska Förlag.
- Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Fuertes-Olivera, F. & Bergenholtz, B. (2011). Introduction: The Construction of Internet Dictionaries. In P.A. Fuertes-Olivera & H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, pp. 1–16.
- Hannesdóttir, A. H. (2014). Lemman och ekvivalenter i nya roller. In R. Vatvedt Fjeld & M. Hovdenak (eds.) *Nordiske studier i leksikografi 12. Rapport fra konferanse om leksikografi i Norden, Oslo 13.–16. august 2013*, pp. 193–211.
- Hólmarsson, S., Sanders, Ch. & Tucker, J. (2008). *Íslensk-ensk orðabók*. Reykjavík: Iðunn. Accessed at: <http://www.snara.is>. (20 May 2015)
- Hult, A.-K., Malmgren, S.-G., Sköldberg, E. (2010). Lexin – a report from a recycling lexicographic project in the North. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, EURALEX 2010. Ljouwert: Fryske Akademy*, pp. 800–809.
- ISLEX*: Accessed at: <http://www.islex.se>. (22 May 2015)
- Kilgarriff, A., Husak M., McAdam K., Rundell M. & Rychlý P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress, EURALEX 2008. Barcelona: Institut Universitari de Lingüística Aplicada, Univeritat Pompeu Fabra*, pp. 425–432.
- Lexin*. Accessed at: <http://lexin.nada.kth.se>. (25 May 2015)
- NE*. Accessed at: <http://www.ne.se.ezproxy.ub.gu.se/ordböcker/>. (23 May 2015)
- Maks, I. (2007). OMBI: The Practice of Reversing Dictionaries. *International Journal of Lexicography* 20(3), pp. 259–274.
- Martin, W. & Tamm, A. (1996). OMBI: An Editor for Constructing Reversible Lexical Databases. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström & C. Røjder Pappmehl (eds.) *EURALEX '96 Proceedings I–II. Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden*, pp. 675–687.
- Martin, W. (2007). Epilogue: Back to the Future. *International Journal of*

- Lexicography* 20(3), pp. 329–334.
- Sanders, Ch. (2013). ISLEX foråret 2013. *LexicoNordica* 20, pp. 259–277.
- Sigurðardóttir, A., Hannesdóttir, A., Jónsdóttir, H., Jansson, H., Trap-Jensen, L. & Úlfarsdóttir, Þ. (2008). ISLEX – an Icelandic-Scandinavian Multilingual Online Dictionary. In E. Bernal & J. DeCesaris (eds.) *Proceedings of the XIII Euralex International Congress, EURALEX 2008. Barcelona: Institut Universitari de Lingüística Aplicada, Univeritat Pompeu Fabra*, pp. 779–798.
- Svensk-isländsk ordbok: *Norstedts Svensk-isländska ordbok*. ([1983] 2005). 4th edition. Stockholm: Norstedts Akademiska Förlag.
- Trap-Jensen, L. (2013). Researching Lexicographical Practice. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London, New Dehli, New York, Sydney: Bloomsbury, pp. 35–47.
- Úlfarsdóttir, Þ. (2013). ISLEX – norræn margmála orðabók. *Orð og tunga*, 15, pp 41–71.
- Úlfarsdóttir, Th. (2014). ISLEX – a Multilingual Web Dictionary. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.) *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Accessed at: <http://www.lrec-conf.org/proceedings/lrec2014/index.html>. (25 May 2015)
- Vikør, L. (1993). *The Nordic Languages. Their Status and Interrelations*. Oslo: Novus Press.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography

Ana Zwitter Vitez^{1,2}, Darja Fišer²

¹ Department of Applied linguistics, Faculty of Humanities,
University of Primorska, Titov trg 5, 6000 Koper

² Department of Translation. Faculty of Arts, University of Ljubljana,
Aškerčeva 2, 1000 Ljubljana

E-mail: ana.zwitter@guest.arnes.si, darja.fiser@ff.uni-lj.si

Abstract

As user-generated content is on the rise both in terms of volume and importance, the long established relation between spoken and written communication needs to be re-examined in lexicography. This is the aim of this paper, in which we perform a corpus-based analysis of typical non-canonical words in spoken and computer-mediated communication in Slovene. The results show that the spoken and the Twitter corpus contain a similar proportion of non-standard pronunciation/spelling variants, interaction words and informal lexemes. On the opposite end of the spectrum are news comments which contain a higher proportion of nouns and a smaller proportion of non-canonical words. The presented study brings a language-independent methodology of identifying typical elements of spoken and written informal texts.

Keywords: lexicography; non-canonical language; computer-mediated communication; spoken language

1. Introduction

Contemporary corpora-based dictionaries are increasingly tackling language material from informal genres, such as tweets, forums, blogs, and comments on news portals. The stereotype of user-generated communication is that it is a hybrid between spoken and written language. Nevertheless, research shows that “netspeak is better seen as a written language which has been pulled some way in the direction of speech rather than as spoken language which has been written down” (Crystal, 2007: 47). To what extent is this true? What are the main similarities and differences between typical spoken and user-generated structures? And how should these typical structures of informal spoken and written genres be included in dictionaries? In order to attempt to answer these questions it seems reasonable to establish a methodology which enables a systematic comparison of spoken and user-generated informal communication.

This paper presents the results of a corpus-based analysis of non-canonical words in user-generated and spoken communication in Slovene. The rest of the paper is structured as follows: in Section 2 we introduce related work analysing spoken and user-generated structures in lexicography; in Section 3, we bring out the analysed

datasets; the methodological Section 4 focuses on the procedure and the main levels of analysis (part of speech, standardization, categorization, linguistic phenomena). In Section 5, we examine the results showing on which levels the analyzed subcorpora of user-generated content display the most spoken language characteristics and in the concluding section, we discuss the value of the results for Slovene and international lexicographic practices.

2. Spoken and user-generated structures in lexicography

Numerous previous studies have confirmed that “there is a whole world” (Morel, Danon Boileau, 1998) between spoken and written texts. These differences have led to the fact that spoken discourse was included in lexicography as soon as technical constraints permitted it. The first Cobuild dictionary (Sinclair et al., 1987), based on the Collins corpus, included examples of English “that people speak and write every day”, including material from radio, TV and everyday conversations. Nevertheless, Moon (1998) argues that the extensive differences between written and spoken language should launch reconsideration in dictionary-making on the levels of phonology, phraseology, collocations, colligations, parts of speech and syntactic structure.

With an increasing quantity of user-generated content on the internet, the relation between spoken and written communication presents a new research challenge. Different disciplines have acknowledged the role of linguistics in the analysis of “netspeak”: D. Crystal (2007) exposes sociolinguistics, stylistics, teaching, and applied linguistics. M. Beißwenger (2012) adds the importance of analysing user-generated contents for lexicography, while exposing genre-specific discourse markers and ‘netspeak’ jargon (like ‘imho’ for ‘in my humble opinion’), and new vocabulary, e.g. ‘funzen’ (an abbreviated variant of the German verb ‘funktionieren’, en.: ‘to function’). Due to the accessibility of user-generated texts, updating vocabulary has become a regular practice: M. Rundell (2014) reports about four updates per year in Macmillan where new words, meanings, and phrases are added (typically at a rate of around 120 to 150 per update).

In Slovene linguistics, historical, political and discipline-specific factors have promoted a protective view of the language, keeping the process of language standardisation separated from the data on actual language use (Verovnik, 2004). Monolingual lexicography is still finding its digital form (Kosem, 2015), but the prevalent doctrine of contemporary lexicography is becoming descriptive, turning away from the position of “how people ‘ought to’ use language” (Atkins & Rundell, 2008: 2). It therefore seems to be the right time to examine the relation between the written user-generated contents and the spoken discourse and start including user-generated contents into dictionaries.

In principle, we know what to do, but in practice, different approaches reveal

potentials and traps when trying to systematically compare spoken and user-generated communication. Linguistic studies (Akinnaso, 1982; Chovanec, 2009; Sindoni, 2013) seem to be comprehensive but are usually not based on quantitative research. On the other hand, different computational approaches give very detailed results on certain linguistic phenomena (Leech et al., 2001; Baron, 2010; Bamman et al., 2014), but only offer results on specific structures. It seems that a systematic corpus study of spoken elements in user-generated discourse could provide valuable insights and could help to resolve the dilemma of including these elements into lexicographic practice.

3. Analysed datasets

For the study presented in this paper we used three corpora:

- 1) a corpus of Slovene called Kres (Logar Berginc et al., 2012) which contains 100 million tokens, sampled from the reference corpus Gigafida. It contains equal proportions of literary, non-fiction, newspaper and internet texts. The corpus has been PoS-tagged and lemmatized. In our study we used it as a baseline corpus displaying canonical, standard written language use.

Example 1)

Example	Kljub obilju, v katerem živimo, pa danes mineralov marsikomu primanjkuje, za kar je kriva nepravilna prehrana.
Translation	<i>Despite the abundance in which we live nowadays, many people lack minerals, which is consequence of poor nutrition.</i>

- 2) the corpus of spoken Slovene called Gos (Verdonik & Zwitter Vitez, 2011) which contains 1 million tokens, transcribed from 120 hours of recorded spontaneous private and public speech on TV, radio, in schools, meetings, bars and at home, sampled for sex, age, region and education level of the speakers. The transcriptions were performed in two ways: one resembles speech as closely as possible while the other one is normalized in accordance with standard spelling conventions, which simplifies corpus querying but also enables the analysis of lexical variants. The transcriptions were also PoS-tagged and lemmatized. In our study we used it to identify the phenomena that are characteristic of spoken discourse.

Example 2)

Example	pa sej itak ni nič februarja itak je eee dons je bla angleščina jutri je pa nemščina to je pa to
Translation	<i>well in any case there's nothing in February today we had English tomorrow we have German and that's it</i>

- 3) the corpus of Slovene user-generated content called Janes (Fišer et al., 2014) which contains 160 million tokens, collected from Twitter, forums, comments on news portals and blogs. As the corpus is rich in non-canonical lexical variants, they were standardized (Ljubešić et al., 2014) before they were PoS-tagged and lemmatized. Social media are used in two very distinct ways: as one of the official news channels by news media, government institutions, private companies and organizations who use the traditional communication conventions, and proper user-generated content in which non-professional users share their personal opinions and experience with their social network in more relaxed settings, often resorting to non-canonical communication conventions. Each text in the corpus was automatically annotated with a standardness measure at the technical and linguistic levels (Ljubešić et al., in press), making it possible to analyse only those parts of the corpus that contain non-standard language, for example.

Example 3)

Example	a se men sam zdi al si neki našpičena dons ? : -(
Translation	<i>is it just me or you really are a bit pissed off today ? : -(</i>

4. Methodology

The goal of the study presented in this paper was to analyse the spoken language elements in computer-mediated communication. We performed this analysis by first identifying the lexical spoken-language features with respect to standard written communication. We then compared lexical features of computer-mediated communication with traditional written communication and checked to what extent the characteristics of the user-generated contents resemble spoken language. As this is the first systematic comparison of Slovene spoken, user-generated and standard corpora, we wanted to analyse single-word units that are typical of each of the corpora. This was achieved by a three-way comparison of keyword lists (Kilgarriff et al., 2004) which were generated in the SketchEngine by comparing both the spoken-language Gos corpus and the Janes corpus of user-generated content against the Kres corpus of written Slovene. While a single keyword analysis was performed on the entire Gos corpus, three Janes subcorpora were examined separately; tweets, forum messages and news comments. We opted for an independent analysis of the three genres because we believe they display important distinctive characteristics and do not resemble spoken language in the same way and to the same degree. Since we were interested in non-canonical language phenomena, only non-standard texts (i.e. those from bands 2 and 3 of the linguistic standardness measure) were included in the analysis.

GOS		Forums		Twitter		Comments	
eee	<i>eee</i>	avto	<i>car</i>	btw	<i>btw</i>	ane	<i>isn't it</i>
mhm	<i>mhm</i>	tud	<i>also</i>	oz.	<i>or</i>	nebi	<i>wouldn't</i>
eem	<i>eem</i>	mal	<i>a little</i>	cca	<i>around</i>	nevem	<i>don't know</i>
sej	<i>any case</i>	tko	<i>like this</i>	slo	<i>Slovene</i>	ala	<i>like</i>
tud	<i>also</i>	blo	<i>was</i>	lol	<i>lol</i>	kriv	<i>guilty</i>
zdej	<i>now</i>	tut	<i>also</i>	cez	<i>in</i>	krivi	<i>guilty (pl.)</i>
tko	<i>like this</i>	gor	<i>up</i>	bos	<i>you will</i>	obsojen	<i>prosecuted</i>
aha	<i>oh</i>	jst	<i>I</i>	nic	<i>nothing</i>	fajn	<i>nice</i>
blo	<i>was</i>	mam	<i>have</i>	prevec	<i>too much</i>	cel	<i>whole</i>
tak	<i>like this</i>	gume	<i>tires</i>	mogoce	<i>maybe</i>	neprimerno	<i>inappropriate</i>

Table 1: Top 10 words from the analysed corpora¹

The top 200 word forms were manually analysed on each of the four generated keyword lists. Each analysis consisted of four steps:

(1) Part of speech: we annotated each keyword with part-of-speech information. Since many word forms are ambiguous, we used the most frequent part of speech annotation only.

word	PoS
tko	<i>like this</i> adverb
aha	<i>oh</i> interjection
blo	<i>was</i> verb

Table 2: Example of PoS annotation

(2) Standardization: First, we checked whether the keyword was canonical. If it was not, we normalized it with its standard variant. If the word form was ambiguous and could be standardized in several ways, we used the most frequent option and annotated it with a special “VARIANT” flag.

word	normalization	Translation 1	Translation 2
pol	potem_VAR	<i>then</i>	<i>half</i>

Table 3: Example of ambiguous normalization

(3) Categorization: We checked whether the keyword form was part of the standard vocabulary. If it was not, we attempted to assign them to different categories, which led us to the next 10 categories, displaying either lexical or orthographic deviations from the norm: abbreviation, omitted diacritics, discourse marker, foreign expression, informal expression, expression signalling interaction in communication, medium-specific expression, spelling resembling pronunciation, non-standard

¹ The translations into English are in italics.

tokenization and topic-specific expression. If the keyword displayed characteristics of several categories, we assigned it the most salient one.

Category	Example	Translation
pronunciation	reku	<i>said</i>
interaction	hvala	<i>thank you</i>
standard	vedno	<i>always</i>
topic	servis	<i>service</i>
informal	folk	<i>people</i>
diacritics	cist	<i>totally</i>
medium	prijavi	<i>report</i>
tokenization	nebi	<i>would not</i>
discourse	hm	<i>hm</i>
abbreviation	cca.	<i>about</i>
foreign	good	<i>good</i>

Table 4: Categorization of the analysed keywords

(4) Linguistic phenomenon: We examined the non-canonical word forms in all 10 categories and identified the linguistic phenomenon at play in each case.

Linguistic phenomenon	Example	Translation
reduction	boljš	<i>better</i>
neutralization	dej	<i>come on</i>
from English	ful	<i>totally</i>
deixis	tale	<i>this</i>
article	ta	<i>the</i>

Table 5: Linguistic phenomenon of deviation

The results of the analysis of the spoken-language corpus and the user-generated subcorpora were compared in order to determine the degree and distribution of interference of speech/written discourse in computer-mediated communication. In the end, an analysis of the extent and distribution of orthographic variation of the non-canonical keywords found in all four analysed samples was performed.

5. Analysis and results

5.1 PoS categorization

In order to get a general picture regarding the material we are dealing with, the keywords in Gos and in user-generated corpora were annotated with part-of-speech information (Figure 1).

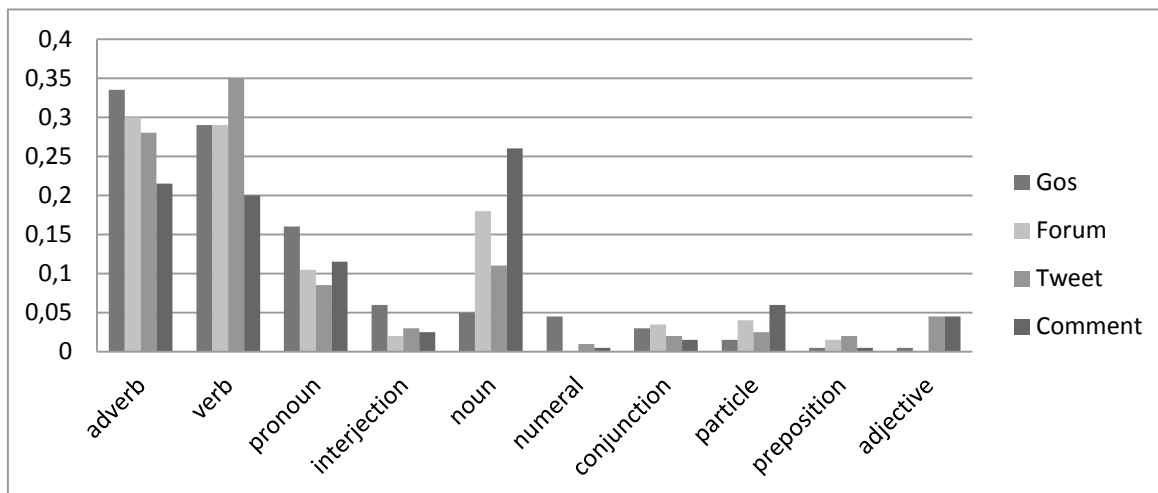


Figure 1: PoS distribution in spoken and user-generated corpora.

The results show that the most frequent PoS categories in the Gos corpus are adverb (33%), verb (29%), pronoun (16%) and interjection (6%). Within the top three typically spoken keywords we find hesitation marks *eee*, *mhm* and *eem* which are the consequence of simultaneous planning and uttering spoken discourse and are thus not present in the user-generated corpora. The high frequency of adverbs (e.g. *čist - totally*) is probably related to their original function of modifying other words, which helps to express the author's opinion. Numerous frequent verbs in the Gos corpus have a different pragmatic function from that assigned in the PoS process (Example 4):

Example 4)

Example // *zakaj kako a več mislim eee poznaš eee [ime] od prej?/*

Translation // *why how you know I mean eee do you know [name] from before?//*

Example 4 shows that the verb *mislim* (e.g. *I think*) plays an important role in keeping attention of the addressee while formulating the rest of the utterance, so it does not function within its traditional syntactic structure (e.g. *I think that...*) but rather as a discourse marker (e.g. *I mean*).

The Forum subcorpus has a similar proportion of adverbs (30%) and verbs (29%). Many verbs relate to the expression of personal opinions or evaluations (e.g. *me zanima - I am interested*, *zgleda - it seems*, *vidim - I see*). Contrary to spoken discourse, the non-standard forum discourse is marked by frequent nouns related to the topic of conversation (e.g. *gume - tires*, *cena - price*, *poraba - consumption*) and the nature of the conversation (*problem*, *odgovor - answer*) where a predictable set of formulations is used, as shown in Example 5.

Example 5)

Example *Hvala za odgovore in lep dan.*

Translation *Thank you for you answers and have a nice day.*

The Twitter subcorpus consists of a slightly lower proportion of typical adverbs (28%) and a significantly higher proportion of verbs (35%) expressing the author's point of view (e.g. *zgleđa* - *it seems*) or illocutionary verbs expressing promise, inquiry or request of interaction with other authors (e.g. *rabim* - *I need*, *poznam* - *I know*, *dobiš* - *you get*):

Example 6)

Example **Rabim prostovoljca ki bi mi prišel skuhat mlečni riž.**

Translation *I need a volunteer who would cook a rice pudding for me.*

The Comments corpus contains fewer verbs and adverbs but a significantly higher proportion of nouns (26%) among the top 200 analysed keywords, than the Gos corpus (only 5%). Nouns in the Comments corpus range from the emotionally marked (e.g. *sramota* - *shame*) to the topic-oriented (e.g. *denar* - *money*, *volitve* - *elections*, *gol* - *goal*):

Example 7)

Example **Sramota. Samo to bom reku.**

Translation *Shame. That's all I'll say.*

It is interesting to note that the process of manually annotating word class for 800 words without seeing their context is less than trivial because very often, a certain word has a traditional PoS identity but operates in a different way in the analysed corpus (this is why it would be interesting to see the score for inter-annotator agreement if many annotators were involved). This phenomenon can be shown by the example of the verb *recimo* (*say*) which mostly operates in the pragmatic function of a discourse connector in the Janes corpus.

5.2 Standardization

With the next level of analysis, we wanted to examine the proportion of non-canonical words among the analysed sample of 200 keywords per corpus. Within the Gos project, standardization was carried out manually (1 million words). For the Janes corpus, an automatic rudimentary standardization has been performed and added as an attribute, but it is currently too imprecise for detailed analysis. This is why we have performed the process of standardization manually for the purpose of this research following the guidelines of the Gos project.

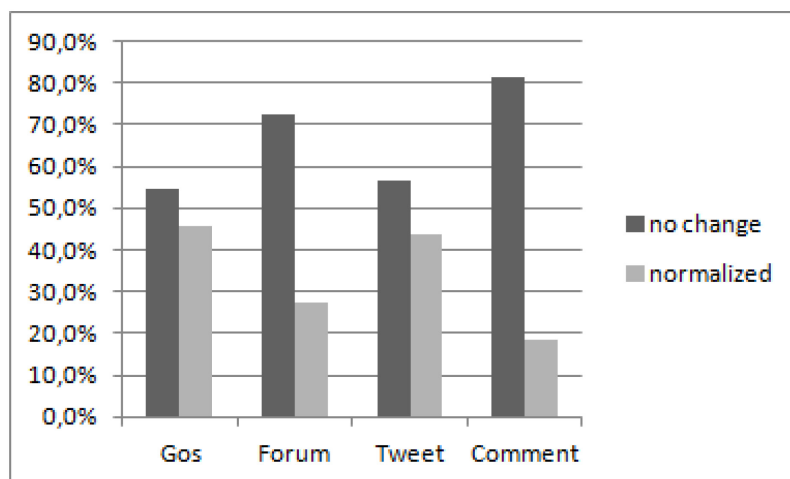


Figure 2: Degree of standardization changes needed in the Gos and Janes corpora.

The results show that in the Gos corpus, a little more than a half of the keywords (55%) were normalized. The normalization is mostly related to pronunciation variation because of reduction on most common words (adverb (44%) and verb (39%)).

Example 8)

Example	in drgač ne prideš gor k je tok strmo
Normalization	in drugače ne prideš gor ker je tako strmo
Translation	<i>and otherwise you won't get there because it's so steep</i>

As can be seen from Example 8, the most common phenomenon of pronunciation variation in the corpus of spoken Slovene is non-stressed vowel reduction. Besides this phenomenon, pronunciation variation concerns different phonetic levels (neutralization, monophthongization, diphthongization) varying from one dialect to another. Some informal words have gone through numerous phonetic changes and have a very different form compared to their standard equivalents (e.g. *pol* - *potlej*, *kva* - *kaj*, *jst* - *jaz*). At this point, it has to be mentioned that the results also depend on the transcription conventions of the Gos corpus transcription using the characters of the Slovene orthographic system following as faithfully as possible the realized acoustic forms of words, with the principal aim to show the typical deviations to the standard pronunciation, see Verdonik et al. (2013).

Regarding the Janes subcorpora, the need for standardization is mostly due to non-canonical spelling (e.g. *drgač/drugače* - *otherwise*) which is influenced by pronunciation variation in spoken discourse, but also the result of omission of diacritics not easily accessed on smartphone keyboards (*mogoce/mogoče* - *maybe*) and non-standard tokenization (e.g. *nevem/ne vem* - *I don't know*). A comparison between the Gos and the Janes corpora shows that the degree of normalization needed in Twitter subcorpus (57%) the most resembles spoken discourse.

Example 9)

Example	haha jst teb to čist resno!
Normalization	haha jaz tebi to čisto resno!
Translation	<i>haha I am totally serious!</i>

As the proportion of words that had to be normalized is higher in the Gos and the Twitter corpora than in the Comments and Forums corpora, we could conclude that spoken and Twitter communication are less standard than that used in Comments and Forums. Yet, as Example 9 shows, the degree of standardization needed is not the only indicator of informal language as communication on Twitter seems to reflect a sociolect of an urban society finding its interactive way to interpersonal communication here and now (as indicated by the frequently used interjection *haha* as an element of reaction to what has been written and the frequent second-person singular pronoun *you* as an indicator of direct interaction).

The Forum and Comments subcorpora show less resemblance with spoken discourse with respect to the degree of standardization required (28% in Forums and only 18% in Comments). It seems that non-canonic language on Forums and Comments is more topic-related: while a patient asking a doctor to explain the results of a medical report will use canonic orthography, but an adolescent discussing his height with his peers will be less devoted to standard language:

Example 10)

Example	jst sm 17 pa sm vlek 189 -.- a se da kako pomajjšati?
Normalization	jaz sem 17 pa sem velik 189 -.- a se da kako pomanjšati ?
Translation	<i>I am 17 and I am 189 cm tall -.- is there a way to get shorter?</i>

5.3 Categorization

The previous section showed that several dimensions of non-canonic language use cannot be explained by limiting the analysis to the degree of deviation from the norm in a particular corpus as they require a deeper linguistic consideration as well. This is why we performed a categorization process which shows for each of the analysed corpora whether a word belongs to standard vocabulary or to one of the 10 identified categories of non-standard forms. With this process, we wanted to examine the characteristics of user-generated language that are adopted from informal spoken discourse and those that represent innovative elements of written computer-mediated communication.

5.3.1 Canonical elements

The category of standard expressions contains words which did not display any non-canonic characteristics (e.g. *dejansko* - *actually*). The biggest proportion of them is found in the spoken corpus and in the Forum subcorpus. It must be noted, however,

that some of the words could have been classified into other groups with more context analysis (e.g. several standard forms reveal intense interaction with other participants and could have been categorized in the category ‘interaction’).

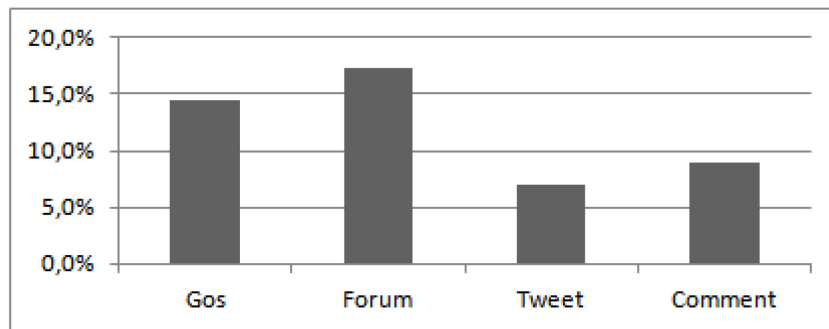


Figure 3: Standard elements in spoken and user-generated corpora.

5.3.2 Spoken language elements

We took a closer look at the non-canonic categories that can be found in spoken and user-generated corpora: non-standard pronunciation or pronunciation-like spelling, topic- or medium-related expressions, discourse markers, and informal or foreign words (Figure 4).

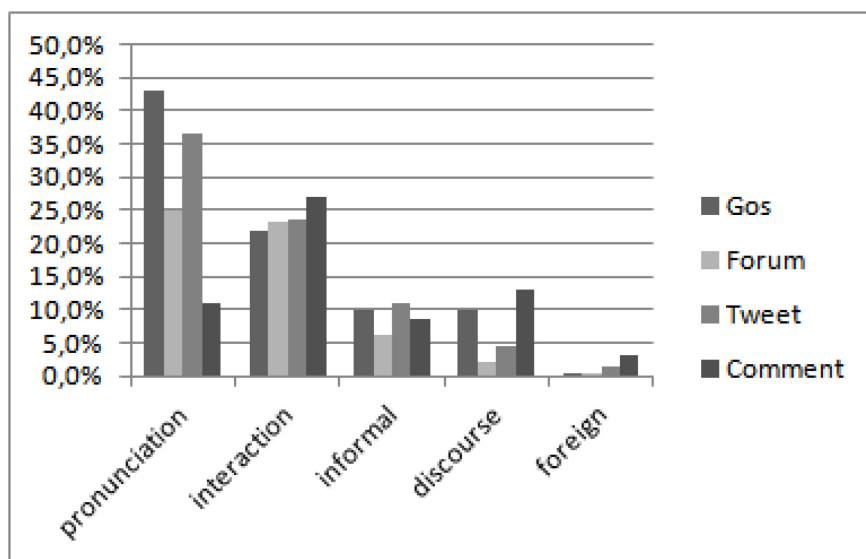


Figure 4: Non-canonic elements present in spoken and user-generated corpora.

Similar to the observations of the standardization process, the Twitter corpus seems to be the most similar to speech in terms of phoneticized spelling of words (43% in Gos vs. 36% in Twitter), interaction (26% in Gos vs. 24% in Twitter), and informal words (10% in Gos vs. 11% in Twitter). As Example 4 shows, the informal words (e.g. *razirat se* - to shave, *nažajfan* - soaped) co-occur with interaction words (e.g. *sej veš* - you

know) and discourse markers (*jah - well*), which all reflect the relaxed and interactive nature of tweeting:

Example 11)

Example	jah sej veš.. za razirat se, morš bit nažajfan:)
Normalization	jah saj veš ... ra razirat se moraš biti nažajfan :)
Translation	<i>well you know ... you have to be soaped to get shaved :)</i>

In the category of discourse markers, the Comments corpus (12%) is the closest to spoken discourse (10%). This category covers mostly adverbs (e.g. *sedaj - now, torej - so*), particles (e.g. *evo - here, pač - well*) and interjections (e.g. *aja - oh, haha*), and gives the impression of imitating the simultaneous process of planning and uttering spoken discourse:

Example 12)

Example	Haha mi je jasno kako je dobila položaj. Vsaj če držijo besede njenih sodelavcev.
Translation	<i>Haha I get it how she got the position. At least if what her colleagues say is true.</i>

Interactive words are characteristic of all analysed corpora (22–27%) and refer to other participants (e.g. *hvala - thank you*) or to the authors themselves (e.g. *gledam - I am watching*). Deictic expressions (e.g. *tole - this*) and interrogative pronouns, such as *kdo (who)* and *kje (where)* belong to this category as well because they also indicate interaction with other participants.

The biggest outlier in this analysis turns out to be the Forum subcorpus, in which we have detected significantly less pronunciation-like spelling (25%), informal lexemes (6%) and discourse markers (2%) than in the Gos corpus. The degree of use of spoken elements correlates with the degree of formality imposed by the forum topic (e.g. lower in medical discussions, higher in threads on motoring). While Twitter users display a distinctive liking for wordplay and innovative language use, the underlying communicative goal of forum users seems to be much more transactional.

5.3.2 User-generated contents-specific elements

Categories which are only present in the Janes subcorpora but not in the Gos corpus represent the most salient CMC characteristics (Figure 5).

The topic of discussion concerns mostly nouns and is most evident in Forums (e.g. *avto - car, problem*) and in Comments (e.g. *tekma - match, volitve - elections*). We were not surprised by this fact because the Janes corpus was constructed from domain-specific forums and because news comments are by definition topic-specific, unlike the topic-diverse GOS and Twitter data.

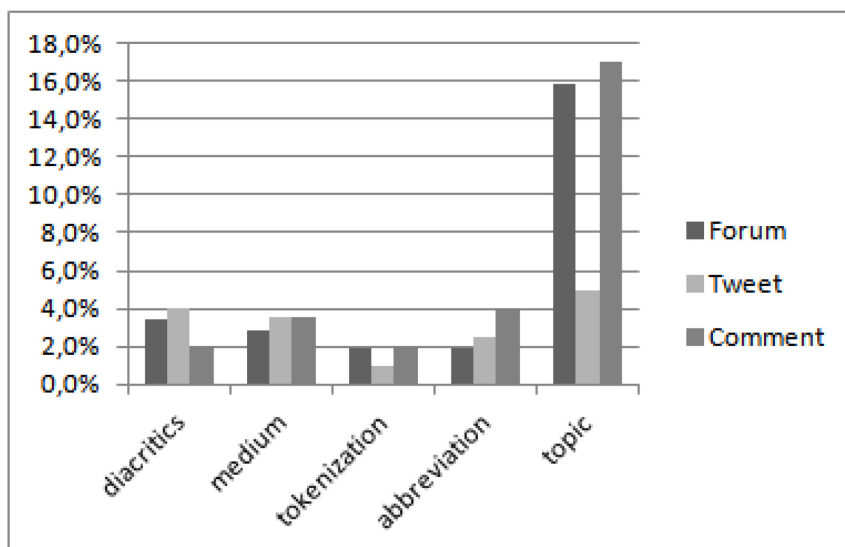


Figure 5: Non-canonic elements only present in user-generated corpora.

All three Janes subcorpora contain keywords revealing the main features of social media (e.g. *com* - *.com*, *všeč* - *like*, *videoposnetek* - *video*), the use of which is important because even though they might be limited to a particular medium at first but then become part of the general vocabulary (e.g. *všečkati* - *like*).

Omission of diacritics, shortening of words and non-standard tokenization are not substantial features in this analysis in quantitative terms because these characteristics are dispersed over different words and will not show within the top typical 200 keywords of a corpus. If a user uses a specific abbreviation, tokenization or does not use diacritic signs, we can only observe the most frequent words characterized by these phenomena. On the level of diacritic signs omission, this is the case of *boš/bos* - *you will*, while non-standard tokenization also concerns the most frequent verbs (e.g. *ne bi/nebi* - *I would not*). In our opinion, non-standard tokenization, more often present in Comments and Forums corpora than in the Twitter corpus, reflects the lack of linguistic competence rather than linguistic creativity.

5.4 Linguistic phenomena

In addition to the general non-canonical categories, we tried to identify the specific linguistic phenomenon of each non-canonical keyword. Since more than half of the analysed words did not get a linguistic label because the phenomenon was already sufficiently defined within the categorization process (discourse marker, interactive words etc.), this subcategorization only relates to some categories of the non-standard analysed words (phonetic spelling, informal and foreign words and discourse markers), which is why the results in Figure 6 are accordingly lower.

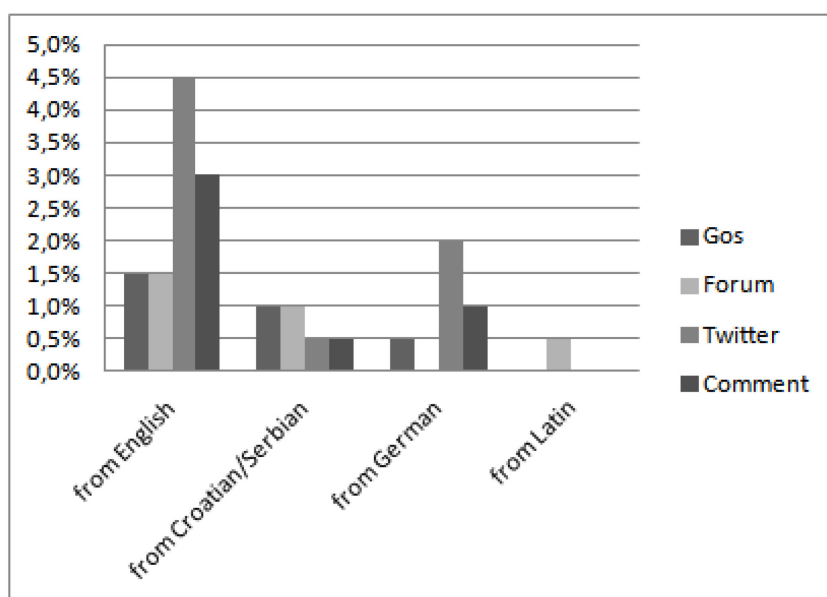


Figure 6: Pronunciation-related phenomena in the spoken and user-generated corpora.

Within the categories that were analyzed in the Gos corpus, the most frequent linguistic phenomena are phonetic reduction, posteriorization, and neutralization, which is also the case for Twitter and Forums (e.g. *drgač/drugače* - *otherwise*). In order to prevent premature speculation about the nature of pronunciation and spelling tendencies in contemporary Slovene, a larger amount of spoken and user-generated data should be studied.

Foreign words in Slovene have historically been subject to numerous stereotypes and different linguistic perspectives have shown very diverse attitudes. As Figure 7 shows, elements from four languages were identified among the top 200 analysed keywords. In the corpus of spoken Slovene, three words were derived from English (*jes* - *yes*), one from Croatian or Serbian (*kao* - *like*) and one from German (*fajn* - *fein*²). Among the user-generated corpora, the Twitter and the Comment corpus seem to contain the most foreign words, considerably more than the analyzed spoken data. On Twitter, we found seven words derived from English (e.g. *app*, *top*) and four from German (e.g. *direkt*, *ziher*), while within Comments, six words were from English and four from German. As we do not want to jump to any premature conclusions with respect to the status and trends of foreign word usage in user-generated contents, a more thorough analysis is reserved for future work.

² This expression could also have been classified as an English one, but due to the historic influence of German in Slovene, we categorized it as a German word.

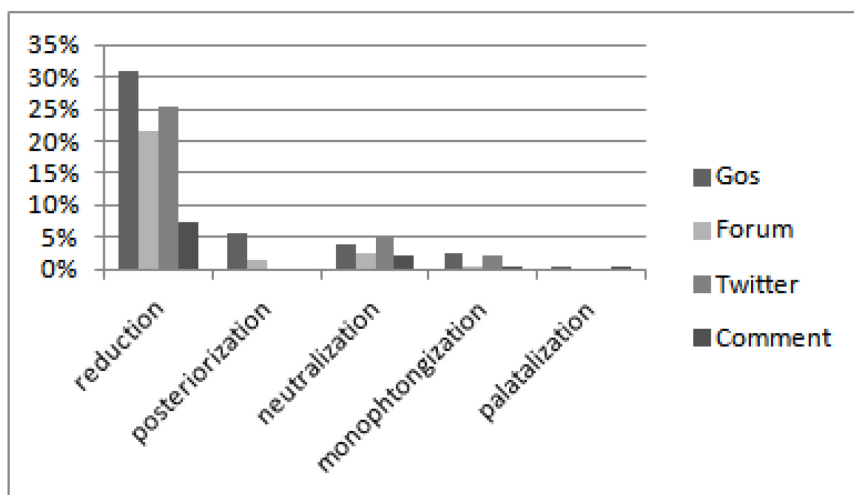


Figure 7: Foreign words in the spoken and user-generated corpora.

Other interesting linguistic phenomena that we have detected are the frequent use of deixis (*tale* - *this one*, *tam* - *over there*), typical in spoken discourse but also characteristic of user-generated corpora, and the presence of “articles” which do not exist in traditional Slovene language manuals (*una ta vesela* - *the happy one*).

6. Discussion of the results

The qualitative and quantitative analysis performed in this study expose the most salient phenomena that show common points and discrepancies between the compared corpora. The first column of Table 6 (*Spoken language*) presents the typical features of spoken discourse compared to the written standard Slovene; the second column (*Similarities*) displays the user-generated subcorpora that contain most of the detected spoken elements; and the third column (*Differences*) relates to the detected specifics of user-generated corpora that are not present in the spoken corpus.

	Spoken language	Similarities (example)	Differences (example)
normalization	high level (45%)	Twitter (<i>jst</i> - <i>I</i>)	Comments; standard words (<i>politiki</i> - <i>politicians</i>)
categorization	pronunciation (43%)	Twitter (<i>drgač</i> - <i>otherwise</i>)	Forums; topic-related vocabulary (<i>original</i> - <i>original</i>)
	interaction (21%)	all corpora (<i>strinjam</i> - <i>I agree</i>)	
	informal (10%)	Twitter (<i>ziher</i> - <i>for sure</i>)	Comments; topic-related vocabulary (<i>krivi</i> - <i>guilty</i>)
linguistic phenomenon	reduction (31%)	Twitter (<i>dobr</i> - <i>well</i>)	Comments; 1 instead of 2 words (<i>nebi</i> - <i>wouldn't</i>)
	deixis (4%)	Forums; deixis (<i>ta</i> - <i>this</i>)	
	foreign words (2%)		

Table 6: Similarities and differences between spoken and user-generated corpora.

The results show that the spoken and the Forum corpora have similar proportions of adverbs and verbs, but that the Twitter corpus shows the most similarities with spoken discourse on the levels of non-standard pronunciation and spelling variants, interaction words and informal lexemes. The most salient specific characteristics of the Comments corpus are a higher proportion of nouns than in speech and a lower level of normalization required compared to speech, while in Forums, topic-related words and non-standard tokenization are prolific.

7. Conclusions and future work

This paper presents a language-independent triangular methodology for lexical comparison of the entire spoken–written spectrum with user-generated content and its informal communication falling roughly in the middle. The results show that a considerable amount of various spoken-language characteristics permeate computer-mediated communication. This is why these characteristics are gaining in importance as they are acquiring new functions in the increasingly interactive and instantaneous online communication where the line between spoken and written discourse are blurred. For this reason, the treatment of such phenomena in contemporary lexicography needs to be re-examined and updated.

It must be noted, however, that this is only the beginning of our studies on this topic which will be extended beyond lexical level in our future work in order to comprehensively also include the context of words (i.e. phraseology, collocations, colligations, syntactic structure). We expect the greatest need for methodological changes at the syntactic level where traditional approaches via conjunction analysis cannot be used and a more important focus should be given on text comprehensibility. Regarding the detected particularities of user-generated communication, a more focused analysis should be carried out on omission of diacritics, word-shortening strategies and non-canonical tokenization.

8. Acknowledgement

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017).

9. References

- Akinnaso, F. (1982). On The Differences Between Spoken and Written Language. *Language and Speech*, 25/2, pp. 97–125.
- Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bamman, D., Eisenstein, J. & Schnoebelen T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18/2, pp. 135–160.
- Baron, N. (2010). Discourse Structures in Instant Messaging: The Case of Utterance Breaks. *Language@Internet* 7, article 4.
- Beißwenger, M. (2013). DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4): pp. 531–537.
- Chovanec, J. (2009). Simulation of spoken interaction in written online media text. *Brno Studies in English*, 35/2, pp. 109–128.
- Crystal, D. (2007). *How language works*. New York: Penguin Books.
- Fišer, D., Erjavec, T., Zwitter Vitez, A. & Ljubešić, N. (2014). Janes se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino. In T. Erjavec & J. Žganec Gros (eds). *Language technologies: proceedings of the 17th International Multiconference Information Society - IS 2014*, pp. 56–61.
- Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings EURALEX 2004*. Lorient, pp. 105–116.
- Kosem, I. (2015). Fran, pameten in intuitiven? *Slovenščina 2.0/2*, pp. 161–193.
- Leech, G., Rayson P. & Wilson A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Longman.
- Logar Berginc, N. & Krek S. (2012). New Slovene Corpora within the Communication in Slovene Project, *Prace Filologiczne*, 63, pp. 197–207.
- Ljubešić, N., Erjavec T. & Fišer, D. (2014). Standardizing tweets with character-level machine translation. In A. Gelbukh (ed.) *Computational linguistics and intelligent text processing : 15th International Conference*. Heidelberg: Springer, pp. 164–175.
- Morel, M.A. & Danon Boileau, L. (1998). *Grammaire de l'intonation*. Paris: Ophrys.
- Moon, R. (1998). On using spoken data in corpus lexicography. In T. Fontenelle, P. Hilgsmann, A. Michiel, A. Moulin, S. Thiessen (eds.) *Euralex 98 proceedings*. Liège: University of Liège, pp. 357–362.
- Pavesi C. (2014). Features of Speech in a Corpus of Learner English CMC: the case of "a lot of" In A.C. Murphy & M. Ulrych (eds.) *Perspectives on Spoken Discourse*. pp. 61–79.
- Rundell, M. (2014). Macmillan English Dictionary: The End of Print? *Slovenščina 2.0*, 2, pp. 1–14.
- Sinclair, J. (1987). *Collins Cobuild English Language Dictionary*. Collins.
- Sindoni, M.G. (2013). *Spoken and Written Discourse in Online Interactions: A Multimodal Approach*. New York/London: Routledge.

Verdonik, D. & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Verovnik T. (2004). Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX*, 46/47: pp. 241–258.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Editing an automatically-generated index with K Index Editing Tool

Kseniya Egorova

K Dictionaries, Tel-Aviv, Israel
E-mail: kseniya.a.egorova@gmail.com

Abstract

This paper presents the editing process of a new Russian–English index using dedicated software. The initial index was generated automatically from the semi-bilingual *Password* English learner’s dictionary for speakers of Russian and the editing was carried out with K Index Editing Tool (KIET). Initially, the editor was provided with the raw index produced according to a set of pre-established principles. It contained all the Russian translations from the *Password* database, converted to potential Russian headwords arranged in alphabetical order and accompanied by the part of speech of the original English equivalents. The revision process then consisted of modifying, removing or adding headwords, confirming or amending their automatically associated part of speech, and matching and re-ordering links to their English equivalents. At the final stage the index was proofread line by line for spelling and grammar mistakes, resulting in a change in index size from 31,666 to 29,039 headwords with 45,929 senses. The paper also demonstrates the main features of KIET and highlights some of the problem areas and major challenges we faced while revising the index.

Keywords: Russian–English index; automatically-generated index; editorial tool

1. Technical description

K Index Editing Tool (KIET) is a new editorial software for creating indices of *Password* semi-bilingual English dictionaries for any language. The initial bilingual list is automatically generated according to a set of pre-established editorial principles, so the Russian target language (TL) translations from the dictionary database are reversed into headwords and the original English source language (SL) headwords are converted into their potential translation equivalents. The automatic generation of the index consists of several steps including XML data parsing and building basic SQLite tables. First of all, the software searches the database for all translations, which are known as translation containers in XML. Subsequently, each translation container is linked with the sense set, which includes several elements: a definition, examples and a headword with part of speech label. The main parameter used for creating basic tables for each language is the definitions, constituting the main attributes of the linked sense, and sense identifiers. Next, the software uses the resulting tables for further parsing. At this step, it identifies translations, which contain commas and semicolons inside the text, and automatically parses them into several parts, divided by these characters. Subsequently, these parts are also turned into separate headwords. The newly-built raw index has the following elements:

- TL translation (turned into headword)
- part of speech
- SL definition
- SL examples (if needed)
- SL senses

Finally, the software links all the sense sets associated with a TL headword. See Figure 1 for the microstructure of a TL entry.

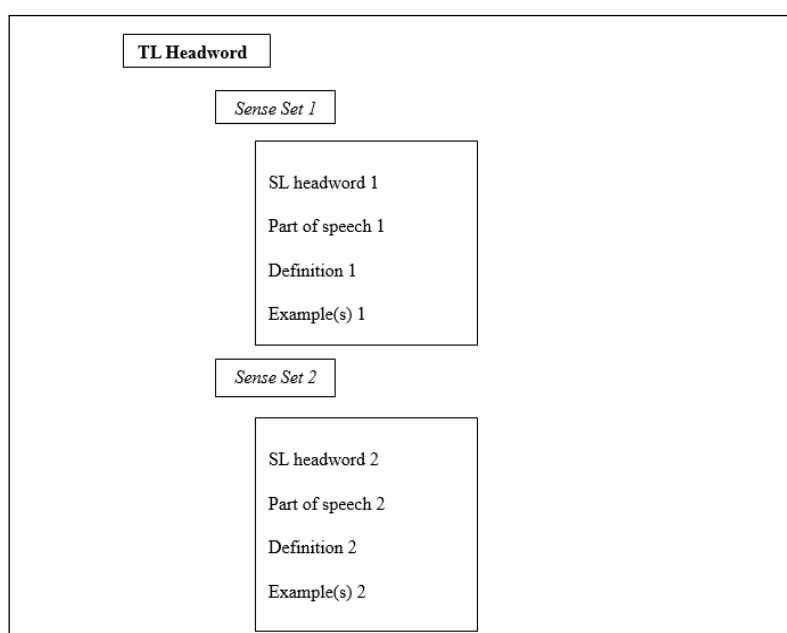


Figure 1: Microstructure of a TL entry

Sorting of the generated index is performed according to the TL alphabet. Subsequently, the editor is provided with the initial index for further editing in KIET.

2. Description of the editing process

The main editing task was to keep the entire structure simple and shape it into a cohesive and comprehensive unit. As the index was intended for Russian speakers, it was important to provide, in one entry, links to all possible English equivalents ('senses') associated with the Russian headword and to make them easily accessible. The entries are displayed in a simple way: corresponding English senses are ordered in a flat structure and followed by definitions (see Figure 2). Examples are not visible in this section. However, when needed, examples of usage and other additional dictionary data can be looked up in full entries.

<p>ВЕКТОР <i>noun</i></p> <p>1. vector <i>noun</i> (mathematics, physics) a quantity such as velocity that has both size and direction</p> <p>2. vector <i>noun</i> (biology) an animal or human cell which is used in genetic engineering to transfer DNA from one cell to another.</p> <hr/> <p>ВЕЛЕТЬ <i>verb</i></p> <p>1. direct <i>verb</i> to order or instruct</p> <p>2. tell <i>verb</i> to order or command; to suggest or warn</p> <hr/> <p>ВЕЛИКАН <i>noun</i></p> <p>1. giant <i>noun</i> (in fairy stories etc) a huge person</p> <p>2. giant <i>noun</i> a person of unusually great height and size</p>
--

Figure 2: Preview of the index

In brief, the editing process of the Russian–English index can be described in four steps:

- (1) modifying, removing and adding the Russian headwords
- (2) adjusting part of speech labels
- (3) revising and reordering the list of related senses
- (4) exporting and proofreading the final index

The following sections of the paper will detail each of these editing stages. First, however, it is necessary to provide a short overview of the tool’s functionality. The majority of the editing was performed in the KIET main screen, which consists of three main parts (see Figure 3). On the left is the list of all headwords. In the middle, the editor can view the list of related senses associated with the headword. The current entry structure is displayed in a dictionary-like form in the entry preview window (on the right). The examples are visible only to the editor to assist in decisions regarding the senses. The icons at the bottom of the main screen (from the left to the right), are used to perform the following actions with the headword list:

1. Edit current Headword
2. Duplicate Headword
3. Add new Headword
4. Remove current Headword
5. Restore current Headword

6. Save changes made to the Headword list

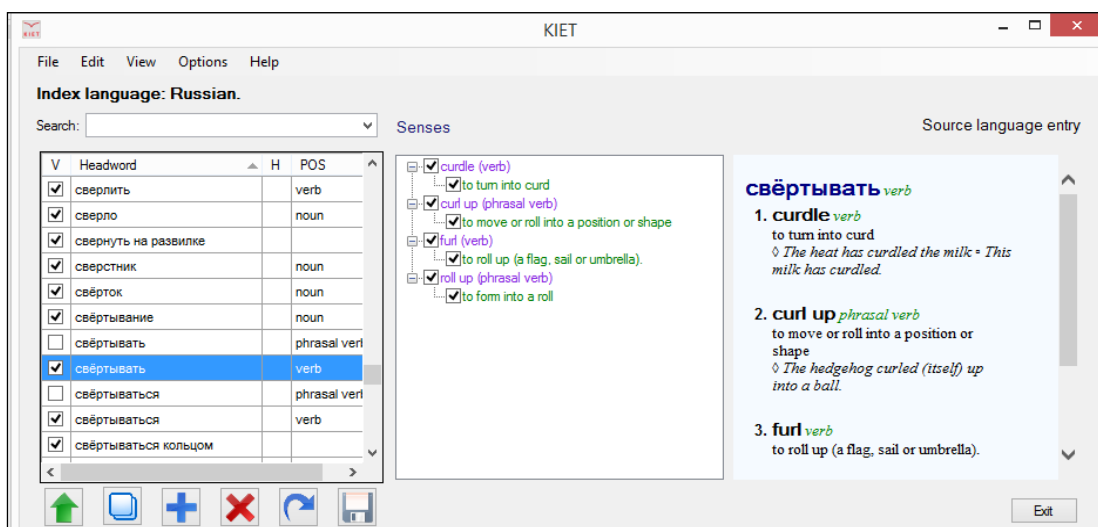


Figure 3: the KIET main screen and its functional buttons.

These functional buttons are used during various stages of editing.

2.1. Modifying, removing and adding headwords

The first editing task concerned reviewing the automatically-generated Russian headword list to check the translations-turned-into-headwords for accuracy and comprehensiveness. The editing was performed in KIET by choosing Select/Unselect a Headword (in the main screen on the left) and checking or unchecking the checkbox preceding it to determine whether or not the headword will be displayed in the dictionary index. In other words, each headword may be set as visible or invisible in the list of selected headwords (e.g. as applied to the redundant headword ‘свёртываться’ (curdle) displayed in Figure 3). Editorial revision at this stage included taking decisions about which headwords should remain unmodified, be modified in different ways, or be removed altogether (buttons ‘Edit entry’ and ‘Remove current entry’, respectively). With KIET it is not possible to physically remove any headword from the initial database but rather it is indicated for later automatic removal by the software from the dataset once editing is complete. It also enables the editor to add new headwords to the headword list if appropriate (buttons ‘Add new entry’ and ‘Duplicate’). In case a newly-modified or added headword happens to already exist elsewhere in the index, KIET displays it to the editor for further consideration.

As the lexical structure of the headword list depended on the *Password* dictionary translation database, there were several types of automatically-formed headwords:

- (1) Direct translations

(2) Approximate translations

(3) Explicative definitions which served as descriptions when there were no equivalents in the TL

Namely, particular challenges were encountered with the second and the third type of translations, in cases when the candidate Russian headword stemmed from them. Such headwords had to be rephrased or shortened into a multi-word expression (if possible) or had certain elements extracted as new headwords to suit the full framework of the edited index and to be comprehensive for its users.

It is important to note that due to the KIET pre-settings the editor was not able to make any corrections in the SL (English) ‘part’ of the dictionary database (including the original source language headword, their part of speech labels, examples and definitions). Only the TL ‘part’ of the database could be edited and modified.

2.1.1. Lexical types of headwords

The Russian headword list consisted of the following types of items: *simple words*, *abbreviations*, *partial words*, and *multi-word expressions* (MWEs). *Simple words* included both *lexical words* (nouns, adjectives, verbs, adverbs and interjections) and *grammatical words* such as prepositions, conjunctions, pronouns, numerals and particles. *Partial words* (productive affixes and combining forms) were also given headword status as many of them are frequently used in Russian: e.g. *про-* (pro-), *недо-* (under-), *два-* (bi-), *авто-* (auto-), etc.

*MWEs*¹ included collocations, fixed and semi-fixed phrases, similes, phrasal idioms, greetings and phatic phrases. Below we give some examples of MWEs from the headword list. As Anokhina (2010) points out, when compiling a bilingual dictionary it is difficult to distinguish between fixed or semi-fixed phrases and collocations, especially those with unconventional translations (even more so for the Russian language, though this is not covered in this paper). Thus, we put first three types of MWEs into one group here:

(1) Collocations and fixed or semi-fixed phrases: e.g. *оказывать влияние* (to bias), *проводить кампанию* (to campaign), *дурное предчувствие* (misgiving, foreboding)

(2) Similes: e.g. *холодный как лёд*² (stone-cold, stone-dead, stone-deaf), *как бешеный* (like fury), *словно живой* (lifelike)

¹ Here we follow the classification of multi-word expressions given by Atkins & Rundell (2008: 166–171).

² Russian similes may be (and usually are) translated with the English equivalents belonging to other types of MWE or even to single-word units.

- (3) Phrasal idioms: e.g. *буря в стакане воды* (a storm in a teacup), *лезть на рожон* (to stick one's neck out), *сводить концы с концами концами* (to make (both) ends meet)
- (4) Greetings: e.g. *Добрый день!* (Good afternoon!), *Здравствуйте!* (hello, hallo)
- (5) Phatic phrases: e.g. *всего хорошего!* (Cheers!), *не беспокойтесь* (never mind)

The bulk of the headwords were common words, but a limited number of proper names was included as well, e.g. *Восток* (the Orient, the East), *Венера* (Venus), *Телец* (Taurus), *Ханука* (Hanukkah), etc.

2.1.2. Homograph headwords

The editorial revision of the headword list included treatment of homographs, since it turned out that the Russian homographs were not identified in the automatic parsing, so it was decided to treat homographs as separate entries. There were two types of homographs to deal with:

- (1) Same spelling but different meaning and pronunciation

e.g. **атлас**¹ (with the stress on the second syllable) (satin) and **атлас**² (with the stress on the first syllable) (a book of maps)

- (2) Same spelling and pronunciation but different meaning and capitalization

e.g. **Весы** (sign of the Zodiac) and **весаы** (a weighing machine)

As a result, homographs with the same spelling but different meaning and pronunciation were duplicated and distinguished by the symbol # and an Arabic numeral (1, 2, etc.). This was performed in KIET by means of clicking on the 'Duplicate' button and making the necessary changes in the list of related senses. As shown in Figure 4, the inappropriate sense (a book of maps) was unchecked from '**атлас#1**'. That sense was linked to the duplicated entry '**атлас#2**' with this meaning. Figure 5 shows a preview of the two entries after changes were made.

The initial processing of the SL translations also did not differentiate between capitalized and non-capitalized homographs with the same part of speech, and these corrections followed manually. If their meanings were different they were also treated as separate headwords but with no homograph number distinction. The capitalization served as a sign that meanings were different (see Figure 6).

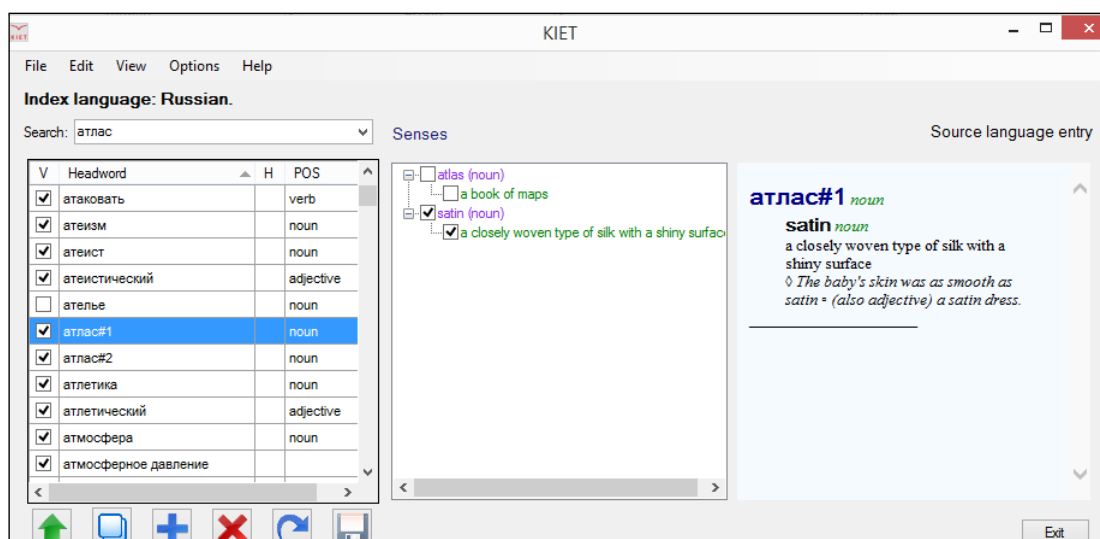


Figure 4: Entry ‘**атлас**¹’ preview

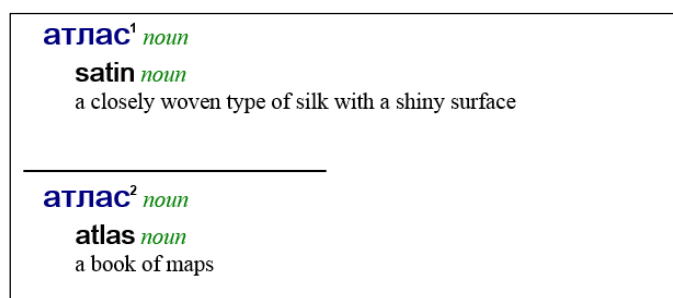


Figure 5: Entries ‘**атлас**¹’, ‘**атлас**²’ preview

For those cases when it was difficult to differentiate homonymy from polysemy – whether it was a plurality of meanings or ‘meaning’ from ‘shade of meanings’ – the headwords were not treated as separate entries. In the case of such difficult decisions, other bilingual and Russian monolingual dictionaries were consulted.

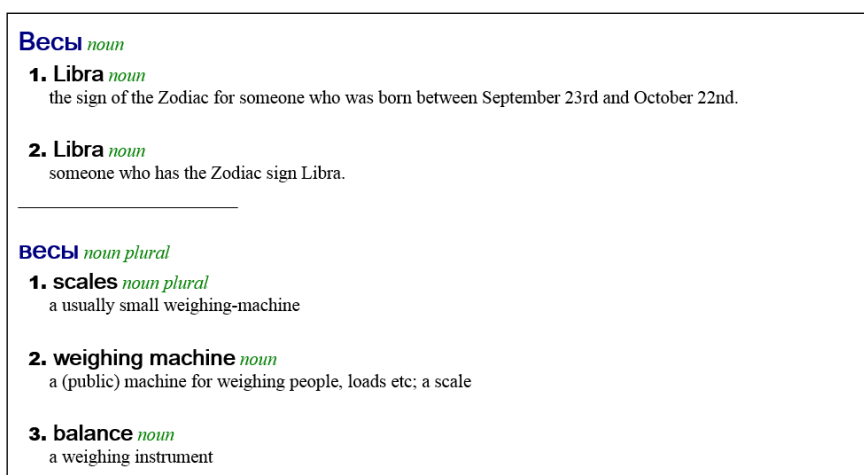


Figure 6: Entries ‘**Весы**’, ‘**весы**’ preview

2.1.3. Making one headword out of several parts

During automatic index generation preceding editing, TL translations that contained commas and semicolons inside the text were parsed by the software and divided by their punctuation settings into separate headwords. This worked well for the translations where a comma or semicolon were used to separate items in a series (e.g. when several synonyms denoting the same thing or object were listed), with each item becoming an independent headword. However, when these punctuations served to introduce a clause in a translation, this rule made a mess. In such cases the translation, which consisted of a complex sentence, was split into two parts that made no sense when used separately. For example, in the translation database the noun *achiever* was translated into Russian as ‘*человек, добивающийся успеха успеха в жизни*’ (literally, ‘a person who achieves success in life’). The second part of the translation, separated by a comma, is a participial phrase, which starts with a Russian present participle ‘*добивающийся*’. As a result of the automatic parsing, there appeared two headwords in the index, ‘*человек*’ (person) and ‘*добивающийся успеха успеха в жизни*’ (someone who achieves success in life), neither of which makes any sense on its own. Subsequently, while revising the headword list, the editor’s task was to find and identify such ‘nonsense’ or inappropriate headwords and reunify the split parts into the corresponding headword (‘*человек, добивающийся успеха успеха в жизни*’).

2.2. Adjusting the part of speech labels

As explained with regards to the lexical structure of the headword list in 2.1.1, both lexical and grammatical words were included in the index. They belonged to the following word-class categories: nouns, adjectives, verbs, adverbs, interjections, prepositions, conjunctions, pronouns, numerals and particles. Due to the overall simplicity of the structure, we did not add grammatical subcategorization in the index. Thus, indications of verb transitivity/intransitivity, of their perfective/imperfective aspects or of various types of pronouns (reflexive, demonstrative, possessive, etc.) were not provided.

According to the pre-established principles, the software automatically attributed the original SL part of speech label to the TL headwords. Subsequently, if the SL equivalent did not belong to the same word-class, the part of speech had to be modified in line with the edited Russian headword or to be removed in the case of MWEs as headwords, which are not labelled at all. In the screen ‘Edit Headword’ the POS label may be changed by selecting from the drop-down menu the necessary word-class marker (see Figure 7). After introducing the changes, the ‘Update’ button was clicked to accept them.

Indeed, in most cases the Russian and English parts of speech did not correspond to each other due to several reasons.

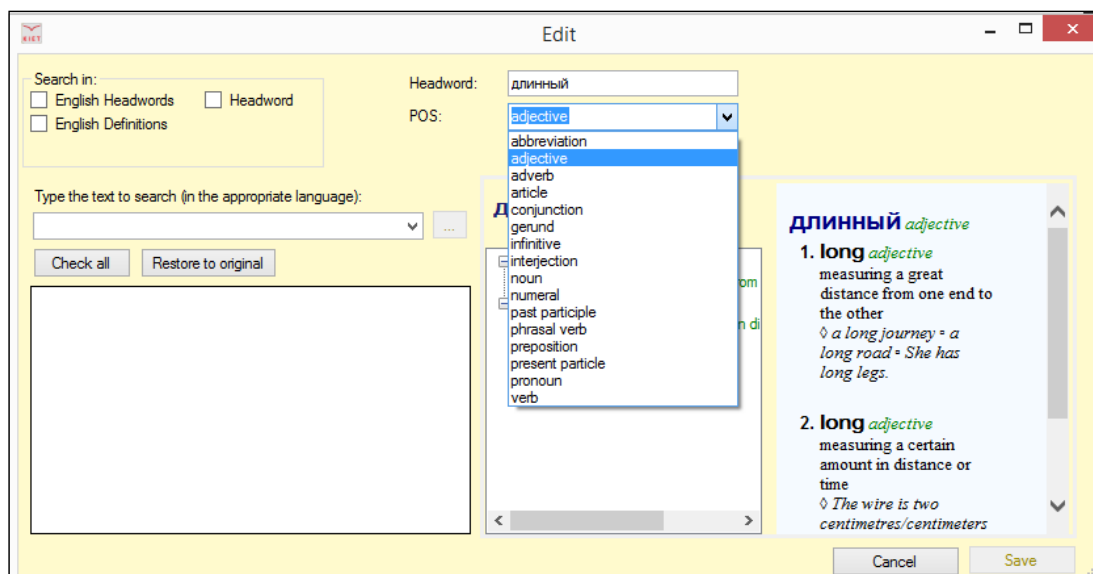


Figure 7: ‘Edit Headword’ screen with a part of speech drop-down menu

First, many English headwords were initially translated into Russian as a MWE or (more rarely) by a different word class. For example, the noun *bookshop* was translated into Russian as *книжный магазин*, which is an adjective + noun fixed phrase (or collocation). Another example is the noun *intermarriage*, which is impossible to translate into Russian as a single-word unit. The typical translation is a phrase of five words of different word-class categories (N. + Prep. + N. + Adj. + N.) such as ‘*брак между людьми разных национальностей/рас*’ depending on the context.

Secondly, some English grammatical categories do not exist in Russian (e.g. articles, gerunds and phrasal verbs). If a headword was automatically attributed this kind of ‘foreign’ word-class marker it had to be adjusted according to Russian grammar. For instance, additional editing was done with ‘*phrasal verb*’ labels, which appeared frequently. Phrasal verbs are usually translated into Russian as verbs with semantically meaningful verbal prefixes (though also depending on the context, see e.g. Yatskovich, 1999; Mudraya et al., 2005). For example, in the dictionary database the phrasal verb *to wake up* was translated as *разбудить* (a single verb with a prefix *раз-*). When the TL translation (*разбудить*) was converted into a headword it still retained the original English-derived part of speech label (*phrasal verb*) and had to be modified into a ‘*verb*’ label. The editor considered all these ‘phrasal verb’ cases in the index and made any necessary changes.

2.3. Revising and reordering the list of related senses

Another main task of the editorial process consisted of attributing the appropriate English equivalents (‘senses’) for each Russian headword and re-arranging them in order. This involved not only fitting the right English translation(s) to the Russian headword, but actually linking the headword to each specific sense of English

polysemous entries that corresponded to it.

If a particular sense was not in the list, the full database was searched. KIET enables the editor to search among the original English entries and definitions or other Russian headwords in the index. A new sense is added by ticking the checkbox that precedes it and the result appears automatically in the preview section.

According to the predefined entry microstructure, the headword senses were presented in a simple flat structure and numbered 1, 2, 3, and so on. The order of the senses could be changed using the mouse to drag the selected sense and drop it in place. This could be done either in the ‘Edit entry’ screen or in the main screen (in the section showing the list of related senses). As Atkins and Rundell point out “...‘dictionary senses’ in a bilingual dictionary are not really senses of the headword at all, but simply the most user-friendly way to structure the material. Bilingual dictionary senses are predicated more on the TL than on the actual meaning of the SL headword” (Atkins & Rundell, 2008: 500). At this stage of editing we stuck to these rules and tried to lay out the senses in a user-friendly way, based on the presumption of which sense the user will look up first. Therefore, we chose the semantic order, putting first the ‘core’ or most common meaning, as judged intuitively. We did not follow the frequency order, as this required a parallel corpus and a frequency analysing software which we lacked. As a rule, the commonest meaning usually consisted of the direct translation of the Russian headword or the most neutral word (in style and register) when selecting among several translation variants from the database. Figure 8 shows the headword ‘**вверх дном**’ (*upside down*) linked to three English senses. The first two are synonyms and the last one is a contextual, indirect translation that was linked with the Russian TL translation in the dictionary database. Therefore, we placed the ‘safest’ meaning (*upside down*) first followed by the less common or stylistically different variants.

вверх дном <i>adverb</i>
1. upside down <i>adverb</i> with the top part underneath
2. topsyturvy, topsyturvey <i>adjective, adverb</i> upside down; in confusion
3. at sixes and sevens in confusion; completely disorganized

Figure 8: Entry ‘**вверх дном**’

In cases when senses that were linked to the headword happened to be regional variants, they were also ordered in the same way. For instance, the Russian ‘*багажная тележка*’ was formed from two ‘senses’ – *luggage cart* in British English and *baggage cart* in American English – that were in fact derived from a single entry.

They were subsequently numbered sense 1 and sense 2, with preference given according to the editorial style guide to the American variant. As a result of this rearrangement, this entry appeared as in Figure 9:

багажная тележка

1. baggage cart *noun*
(American) (alsoluggage cart) a cart used by passengers at an airport etc to carry their luggage.

2. luggage cart *noun*
(British) a cart used by passengers at an airport etc for carrying their luggage; baggage cart(American)

Figure 9: Entry ‘багажная тележка’

The entries that consisted of the full translation equivalent and its contracted or abbreviated form were also presented in a flat structure with the full form always first and the contraction/abbreviation after. For instance, as two English equivalents were linked with the headword ‘*суббота*’ (*Saturday* and *Sat.*), we rearranged their order using the drag-and-drop function and listed *Saturday* as sense 1 and *Sat.* as sense 2 (see Figure 10).

суббота *noun*

1. Saturday *noun*
the seventh day of the week, the day following Friday

2. Sat. *written abbreviation*
short for Saturday

Figure 10: Entry ‘суббота’

2.4. Exporting and proofreading the final index

Finally, after all changes had been saved in the database, the edited index was exported from KIET and the export files were sent for processing. The features of KIET also enable the editor to create HTML files and see all the performed changes and the final result in a user-friendly format. When the data had been processed, the entire index was proofread line by line (in HTML-format on a screen) for spelling and grammar mistakes. The POS-labels and the linked senses were double-checked once again.

3. Conclusion

This paper gave an overview of the functions of KIET that are used for automatic generation of bilingual indices. After editing and proofreading was completed, the size of the Russian–English index changed from 31,666 to 29,039 headwords with 45,929 senses. In other words, at least 2,627 raw headwords were removed altogether (especially explicative definitions, due to their wordiness and a low probability of being looked up). Another part was paraphrased and shortened and some of the

headwords, which were split parts of single translation units, were combined into a single headword. While revising the headword list, we did not add many new headwords; where added, they were basically duplicated entries for the homograph headwords we discussed above.

Editing the Russian–English index was an interesting, challenging and thought-provoking task. Some of the challenges, no doubt, are language-specific and may be explained by the peculiarities and complexity of the Russian language. Major problem areas (such as part of speech tagging) were reported to the KIET technological developers and solved on the run by means of export adjustments in initial data processing. New export algorithms were added to the latest version of KIET. It would be interesting to investigate if the main challenges and problem areas discussed in this paper are also relevant to the editing of other language pairs, and to compare the results of other *Password* indices.

4. Acknowledgements

The author wishes to thank Natalia Kustovinov, of the KIET development team, for her help with the technical description of the editing tool.

5. References

- Adamska-Sałaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. *International Journal of Lexicography*, 26(3), pp. 329-345.
- Anokhina, J. (2010). Lingvo Universal English-Russian Dictionary: Making a Printed Dictionary from an Electronic One. In A. Dykstra & T. Schoonheim (eds.). *Proceedings of the 14 Euralex International Congress. 6-10 July 2010. Leeuwarden/Ljouwert, the Netherlands: Fryske Akademy*, pp. 539-548
- Apresjan, J. D. (1973). ‘Regular Polysemy’, *Linguistics* 12 (142), pp. 5-39.
- Apresjan, J. D. (2003). ‘Lexicographic Concept of the New English Russian Comprehensive Dictionary’. In Apresjan J.D. (ed.) *New English-Russian Comprehensive Dictionary*. Moscow: Russky yazik, v.1., pp. 6-17 [Leksikograficheskaya kontsepsiya Novogo bol’shogo anglo-russkogo slovar’ya]
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- K Index Editing Tool (KIET) User Guide*, KIET version 1.0.0.0., Copyright © 2004-2014, K Dictionaries Ltd.
- Mudraya, O., Piao, S., Lofberg, L., Rayson, P., & Archer, D. (2005). English-Russian-Finnish cross-language comparison of phrasal verb translation equivalents. Accessed at: <http://comp.eprints.lancs.ac.uk/1061/1/phraseology05.pdf>
- Yatskovich, I. (1999). Some ways of translating English phrasal verbs into Russian. *Translation Journal*, Vol.3 (3), July, 1999. Accessed at: <http://translationjournal.net/journal/09russ.htm>

Dictionaries:

Abby Lingvo-Online English-Russian Dictionary. Accessed at: <http://www.lingvo-online.ru/en>

Apresjan, J.D. (2003) *New English-Russian Comprehensive Dictionary*. Moscow: Russky yazik. [Novyy bol'shoy anglo-russkiy slovar']

Katzner's English-Russian, Russian - English Dictionary (1994). Rev. and expanded ed. John Wiley & Sons, Inc.

Oxford Russian Dictionary (1996). Revised and updated by C. Howlett. Oxford: Oxford University Press.

Password English Dictionary for Speakers of Russian. Accessed at: <http://www.kdictionaries.com>

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



A study of the users of an online sign language dictionary

Mireille Vale

Victoria University of Wellington
PO Box 600, Wellington, New Zealand
E-mail: micky.vale@vuw.ac.nz

Abstract

In this paper I report on a mixed method user study of the Online Dictionary of New Zealand Sign Language (ODNZSL). While sign language dictionaries make comparatively full use of the potential offered by the digital format, they have not previously been the focus of much user research and to date there have been no published studies of the usability of electronic dictionary features such as video material, bidirectional search methods and hyperlinked information. This study focuses on broad questions: who the users of the ODNZSL are, their motivation for consulting the dictionary, aspects of their dictionary consultation behaviour and problems that they currently experience.

The study draws on two data sets: firstly, I analysed log data from the ODNZSL website using Google Analytics; and secondly, I carried out a think-aloud protocol and follow-up interview with representatives of potential user groups identified through a pre-compilation user survey. After a brief description of the structure and format of the ODNZSL, results from these two investigations will be discussed along with implications for optimising the ODNZSL's usefulness for its diverse users, and for online dictionaries in general.

Keywords: sign languages; electronic dictionaries; users; log files; think aloud

1. Introduction

Sign language dictionaries are amongst the dictionaries of lesser-resourced languages (Prinsloo, 2012) that arguably stand to benefit the most from the digital revolution. There are two main purposes for creating dictionaries for sign languages: firstly, to document the language and support its preservation and recognition; and secondly, as an aid to people wishing to learn the language (Schermer, 2006; Woll, Sutton-Spence & Elton, 2001). Digital technologies support these purposes, both for the dictionary maker and the user.

In the case of sign languages, some of the capacities of digital dictionary-making are not yet applicable: for example, since there is no accepted sign language orthography there are no large corpora of written texts to draw on. Although video corpora of sign languages are becoming more widespread (see Konrad, 2012 for a survey of current sign language corpora), these are still small compared to spoken and written corpora, partly because of technical limitations but also because in many ways, sign languages are 'young' languages that have until recently been used only in limited domains and that have high levels of polysemy and variation (McKee & McKee, 2013). Structural issues in sign formation also affect lemmatisation, with a large set of productive

morphemes and semi-lexicalised sign forms, but relatively few established lexemes (Johnston & Schembri, 1999, estimate that these number in the thousands rather than the much higher rates of established lexemes found in spoken languages). This means that most (online) sign language dictionaries have a comparatively modest content of around 2,000–5,000 headwords (Zwitserslood, 2010). In other respects, digital technology has significantly facilitated sign language lexicography. In particular, sign language dictionaries can now store video data to represent signs much more effectively than previous static images. The electronic format thus allows for greater visibility and accessibility of sign languages to both the language community and the wider public, raising awareness that may lead to increased recognition of the linguistic and cultural rights of their communities (Schermer, 2006; McKee & McKee, 2013).

An increase in the production of sign language dictionaries in the past decades has been accompanied by these dictionaries becoming the object of research. Within the growing body of articles on sign language lexicography, there has been some focus on the user; however, this has mostly been limited to surveys of potential users prior to the compilation of a sign language dictionary (e.g. Moskovitz, 1994; McKee & Pivac Alexander, 2008) and reviews of existing dictionaries (e.g. Zwitserslood, 2010; Schmaling, 2012). It is generally assumed that sign language dictionaries – especially the first dictionary for any particular sign language – are multifunctional and will serve a wide range of users; indeed, the forewords to many dictionaries mention the sign-language-using deaf community, (hearing) language learners including parents of deaf children, and language professionals such as sign language interpreters. As a result, sign language dictionaries have nearly always been bilingual, and often unidirectional, allowing only for searches by a written word to locate a sign.

There are now a few examples of thematic dictionaries and smaller dictionaries for specific user groups (Schermer, 2006). For most general sign language dictionaries, however, better use might be made of limited resources by using the digital medium to provide customisation of dictionary content for different users and different functions. One example of this is the bidirectional access provided by some of the recent online sign language dictionaries, which allows users to identify a sign by its phonological features to look up spoken or written language equivalents, as well as the more usual word-to-sign search direction (Zwitserslood 2010; Kristoffersen & Troelsgård, 2013). While performing such a search at the moment requires considerable analytical skills from users unfamiliar with sign phonology, there is potential for modern technologies, such as motion recognition, to provide much more accessible user interfaces in the near future. In the same way as Lew & de Schryver (2014) see a future for a dictionary in a pair of glasses, so may there be a sign language dictionary interface in a pair of gloves. Before such adaptations are implemented, however, it is vital that we confirm who the users are and how online sign language dictionaries are used in practice. Kristoffersen & Troelsgård (2012)

point out that there have as yet been no major usability studies of sign language dictionaries.

The current exploratory study may be the first to report on the observed behaviour of actual users of an online sign language dictionary. The study focuses on the Online Dictionary of New Zealand Sign Language (ODNZSL), an example of a recent dictionary that makes use of many of the digital features discussed above. The next section will describe these features in more detail.

2. The Online Dictionary of New Zealand Sign Language

The project to develop the ODNZSL took place from 2008 to 2011. The project built upon existing data that were collected for the earlier print Dictionary of New Zealand Sign Language (Kennedy et al., 1997) and the Concise Dictionary of New Zealand Sign Language (Kennedy et al., 2002). The aim was initially to review and, where necessary, re-validate data from the approximately 4,500 headwords in the 1997 print dictionary and to make these data available online. The ODNZSL was launched in July 2011.

For the purpose of this paper, a brief tour of the ODNZSL website (<http://nzsl.vuw.ac.nz>) will give an idea of the content, structure and format of the ODNZSL as a background to the user study. A comprehensive description of the development of the ODNZSL and a discussion of some of the lexicographical challenges in its creation can be found in McKee & McKee (2013).

2.1 The Home Page

The home page (Figure 1) gives access to the ‘front’ and ‘back’ matter of the dictionary through a series of tabs, providing background information on New Zealand Sign Language (NZSL); grammatical information regarding the number system, fingerspelling alphabet, and the productive classifier morpheme system in NZSL; a help menu which also contains a glossary of terms used in the description of signs in the dictionary; advice for learners with a link to learning exercises; links to relevant organisations; and a contact form which allows users to provide feedback or ask questions.

By clicking the ‘play this page in NZSL’ button, the information on the home page and in the tabs can be viewed in video format signed in NZSL. English and NZSL are therefore both used not only as part of the bilingual dictionary structure but also as metalanguages. Te Reo Māori translations of each headword were added to the ODNZSL in 2013, so that all three official languages of New Zealand are now represented in the dictionary, although Te Reo Māori is not (yet) used as a metalanguage.

A ‘show me a sign’ feature provides a link to a random sign entry, in a similar way to the ‘Word of the Day’ now provided by some online dictionaries.

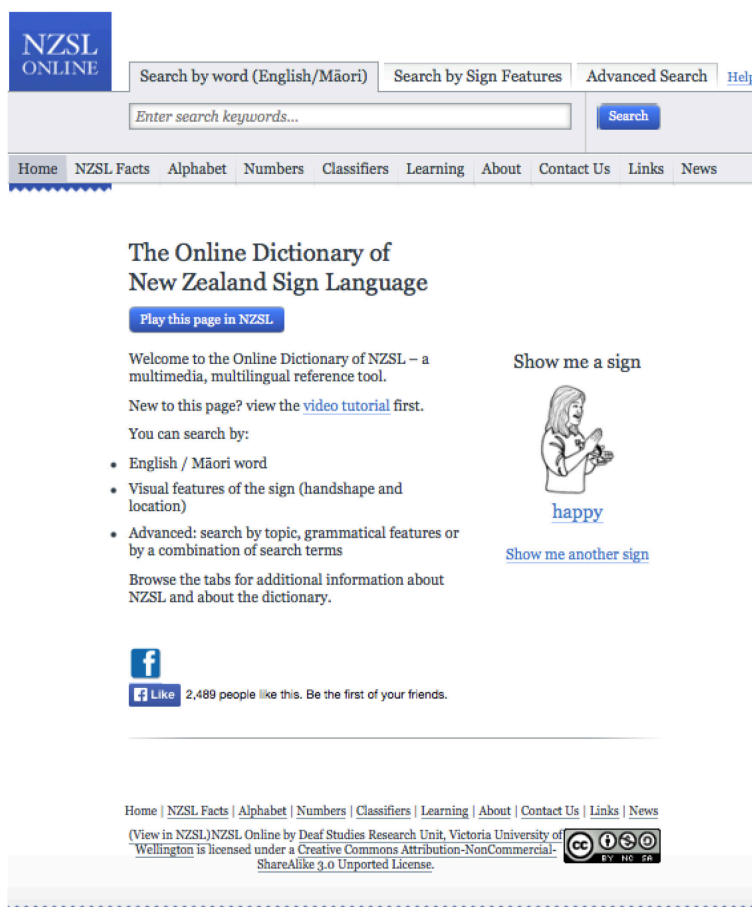


Figure 1: The ODNZSL home page

2.2 Search Methods

Three search methods are available:

- The Search by Word (English/Māori) is a standard search box, which brings up predictive text suggestions of headwords in the dictionary once the user starts typing.
- The Search by Sign Features asks users to select two main phonological features of a sign from a menu of images: the handshape and the location where the sign is produced (see Figure 2)
- The Advanced Search allows for a combination of search criteria from the above two methods, as well as a choice of topics for a thematic search and a list of five usage tags: neologism, archaic, obscene, informal and rare.

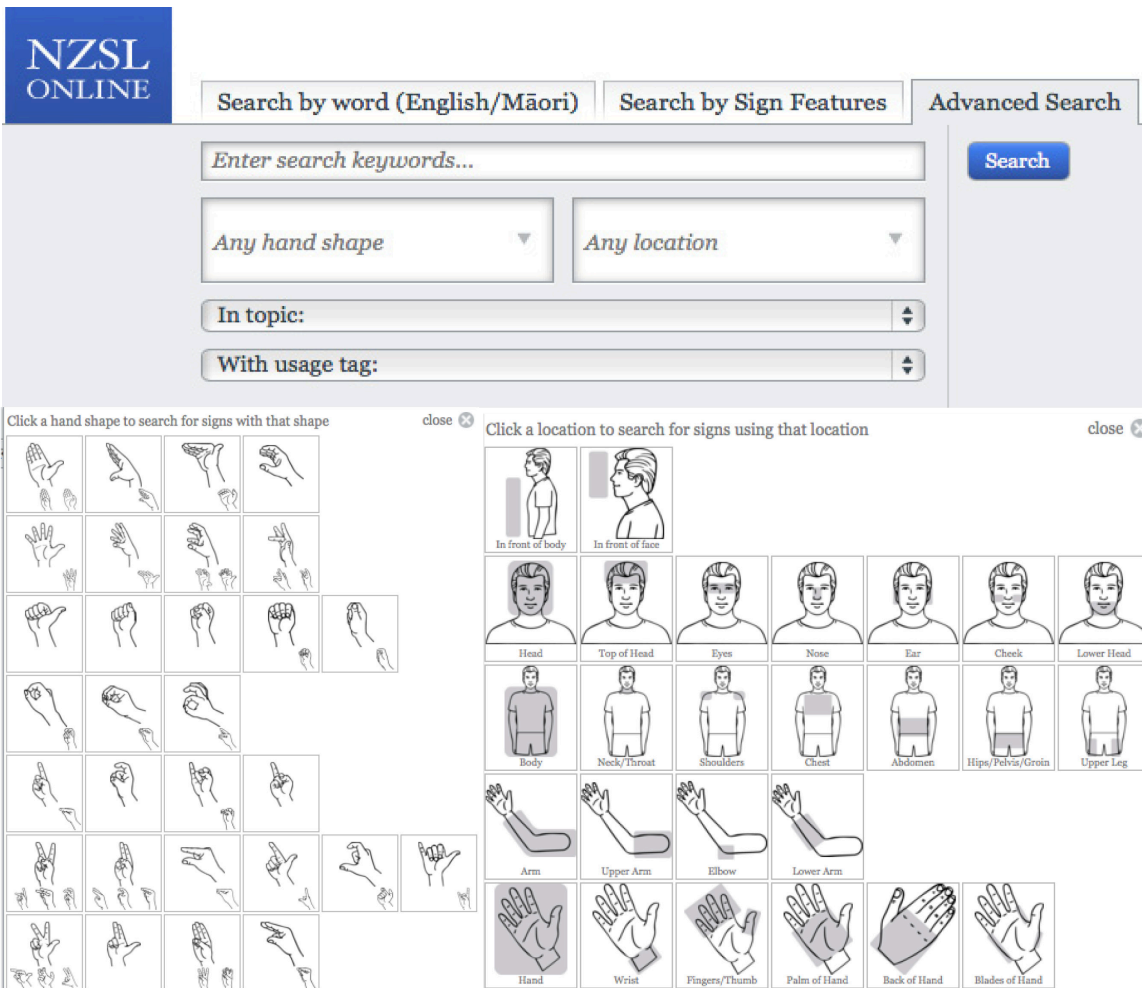


Figure 2: Search methods in the ODNZSL

2.3 Search Results

Information displayed in the search results of the ODNZSL consists of a drawing representing the sign form, followed by glosses in English and Te Reo Māori that capture the main sense(s) of the sign, a series of further translational equivalents in English, and the word class(es) to which the sign belongs. Static representations of the sign are used here instead of video files in order to speed up the loading of the search results. Due to the space the drawings take up, results are paginated with a limit of nine results displayed per page (see Figure 3). Results are displayed in alphabetical order with exact matches for the main gloss displayed first, before exact matches in the translational equivalents and partial matches in both. When there are multiple exact matches, the most frequent sign is displayed first.

The screenshot shows the NZSL ONLINE search interface. At the top left is the NZSL ONLINE logo. Below it are search options: 'Search by word (English/Māori)', 'Search by Sign Features', 'Advanced Search', and 'Help'. A search box contains the word 'grow', with a 'Search' button and a 'clear' link. Below the search box is a navigation menu with links: Home, NZSL Facts, Alphabet, Numbers, Classifiers, Learning, About, Contact Us, Links, and News.

The search results are titled 'Search results for: "grow"' and show '4 results'. Each result includes a drawing of the sign, the English glosses, the Te Reo Māori gloss, and the word class information.

- Result 1:** Drawing of a person with hands in front of their chest. English glosses: **develop, grow**; *mahi, tupu*; development, expand, expansion, growth, increase. Te Reo Māori gloss: *mahi, tupu*. Word class: *verb*.
- Result 2:** Drawing of a person with hands held together in front of their chest. English glosses: **grow**; *tipu haere*; expand, expansion, growth, increase. Te Reo Māori gloss: *tipu haere*. Word class: *verb*.
- Result 3:** Drawing of a person with one hand raised. English glosses: **grow up, lifetime**; *e tipu ake ana, aku rā*; *katoa*; childhood, life. Te Reo Māori gloss: *e tipu ake ana, aku rā*. Word class: *noun*.
- Result 4:** Drawing of a person with hands held out. English glosses: **grow up**; *kia pakeke ngā*; *whakaaro*; act maturely. Te Reo Māori gloss: *kia pakeke ngā*. Word class: *phrase, verb*.

Figure 3: Search results display in the ODNZSL

2.4 The Dictionary Entry



Figure 4 shows the information that is displayed for an individual entry. Each entry contains the following elements (numbered in the figure):

1. Drawings indexing the handshape and location of the sign;
2. One or more English glosses showing the main sense(s) of the sign;
3. A number of further glosses that are either less common senses or common translational equivalents of the sign;
4. A Te Reo Māori gloss;
5. Word class information;
6. Possible inflections, hyperlinked to a glossary in the help menu;
7. A drawing of the sign;
8. A large video showing how the sign is produced;

9. Example sentences, consisting of a signed video accompanied by a translation into English, and a glossed representation of the sentence where each gloss is hyperlinked to the relevant entry in the ODNZSL;
10. A usage note and/or a hint for producing the sign where applicable.

Users also have the option to play any video in slow-motion and to add the sign (in the form of the drawing and English and Te Reo Māori glosses) to a vocabulary sheet to be printed or saved as a PDF.

[Back to search results](#)

1  

2 play


3 player, playing

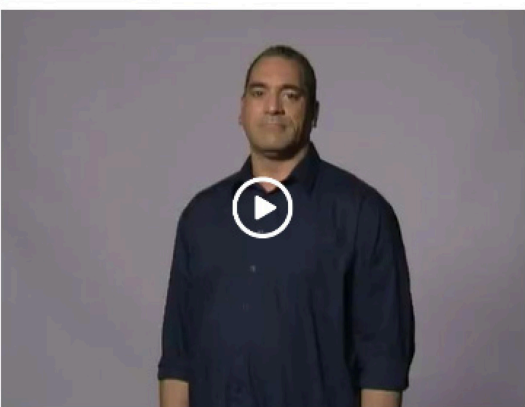
4 tākaro

5 noun, verb

6 temporal inflection

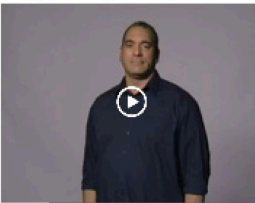
[Add to Vocab Sheet](#)

7 

8 

[Play in slow motion](#)

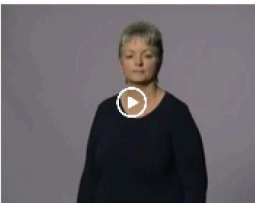
9 Usage Examples



He is very keen on sports and plays soccer and tennis.

[he](#) [keen](#) [sport](#) [play](#) [soccer](#) [tennis](#) [he](#)

[Play in slow motion](#)



It was raining outside. The two sisters had nothing to do and played with toys inside.

[outside](#) [rain](#) [both](#) [sister](#) [twiddle-thumbs](#) [play](#) [toy](#) [inside](#)

[Play in slow motion](#)

10 Notes

Not used for playing a musical instrument.

Figure 4: Individual sign entry in the ODNZSL

3. Research Questions

Since there has been little prior research on the users of online sign language dictionaries, the current study did not specify a particular user group or situation. Instead, it focused on four broad questions similar to those suggested by Tarp (2009) and Nesi (2013) as appropriate for dictionary user research:

- Who uses the Online Dictionary of New Zealand Sign Language?
- What is the users' motivation for using this dictionary?
- How is this dictionary used, and what kinds of information do users look up?
- Do users have particular problems or issues in using this dictionary?

4. Method

Log files are increasingly used as a method in dictionary user research, offering the advantage of unobtrusive observation of real-life behaviour (Tarp, 2009). In their log file based user study, de Schryver & Joffe (2004) show the potential of this method to gather detailed information to the benefit of both immediate improvements to a particular dictionary and a more thorough understanding of user behaviour in general. There are some technical obstacles in the way individual users are tracked and limitations to how log file data findings can be applied when the wider context that prompted the dictionary consultation is unknown (Bergenholtz & Johnsen, 2007; Tarp, 2009; Müller-Spitzer, 2013). For the current study, the advantages of having access to a large number of lookups from all potential users outweigh the shortcomings of using this method.

To gain a more qualitative (if subjective) perspective, the main data from the log files was supplemented with interview questions and a think-aloud protocol to probe into users' motivations and attitudes towards the dictionary, as well as examining particular user problems in more depth. Thus the study attempts to triangulate results through using mixed methods: an approach that is increasingly common in dictionary user studies (Nesi, 2013). This part of the study only involved a small number of participants: larger follow-up studies as well as those employing other methods (such as experiments with particular user groups) will be required to confirm the tentative results reported here.

4.1 Log Files

General website traffic for the ODNZSL has been tracked since its inception in July 2011 using Google Analytics, a widely available web analytics programme.

Standard information tracked by Google Analytics includes the number of visitors, how they arrived on the site, how much time they spent on the site, how many pages they viewed and what site searches they carried out. To track user interaction in more detail, 'Events' were set up to also track:

- the exact search string typed in during a search;
- instances where a user clicked on a video to view it;
- clicks on help items, including the introductory video on how to use the dictionary, the help menu, and hyperlinks to the glossary;
- instances where a user clicked on one of the glossed signs in an example sentence;
- the position of a search result of a sign entry when the user clicked on it.

Since these adaptations to the log files were not implemented until March 2014, the selected time period to collect data comprised three months between April and June 2014, a representative period which includes the most active months of dictionary use during the year. During this period, a total of 31,753 sessions were logged. The number of users was 16,296. The number of page views was 319,662, equating to an average of 10.07 page views per session.

In common with other web analytics programs, Google Analytics relies on the tracking of individual users via ‘cookies’. While this method provides an improvement over logging server side requests (where cached pages, for example, cannot be easily tracked), inaccuracies may occur due to users blocking or periodically deleting cookies, or being misidentified as unique users when logging in from different devices. Google Analytics have recently implemented a ‘unique user’ profile that can distinguish between users from the same IP address, and conversely can trace the use of different devices by the same user. The profile also offers more in-depth demographics. However, there are ethical implications of tracing individuals in this way. If this function is implemented on a website, it is therefore recommended that website visitors are informed that their personal data is gathered from the site and asked for their consent. This may be a deterrent to people using the site. For this reason, and because this function gathers demographic data beyond what was required for the limited purposes of this study, it was decided not to make use of the ‘unique user profile’ function.

4.2 Think-Aloud Protocol

4.2.1 Participants

The selection of participants was based on a number of the potential user groups identified in a survey by McKee & Pivac Alexander (2008) and also reflects the categorisation by Varontola (2002) of dictionary users as:

- 1) Language learners
- 2) Non-professional users
- 3) Professional users

Participants were recruited through existing networks, both through distribution of an information sheet and through personal invitation to relevant groups, such as

networks of sign language interpreters, New Zealand Sign Language classes and the local deaf community. Twelve volunteers were selected. Table 1 shows the selected participants by category and their status in relation to fluency in, and use of, NZSL.

Varontola (2002) dictionary user category	NZSL status	Length of time since learning NZSL	Amount of time spent using NZSL	Number of participants
Language learners	Beginner learner (first year class)	6 weeks of course learning	4-7 hours a week	3
	Intermediate learner (second year class)	1-2 years of course learning	4-7 hours a week	2
Non-professional users	Hearing friends of a deaf person	Minor exposure; no formal learning	Very occasionally	2
	Deaf community	1 since early childhood; 1 since late teens	Daily (main language)	2
Professional users	NZSL tutors / teachers	Since early childhood (before age 3)	Daily (main language)	1
	NZSL interpreters	8-11 years, including course learning for 3-4 years	Daily (work + social)	2

Table 1: Interview / TAP participants

4.2.2 Procedure

The activity consisted of four parts:

- A short pre-interview
- A familiarising exercise
- The TAP exercise
- A follow-up interview

Pre-interview questions focused on the participants' prior language learning and dictionary use and their familiarity with sign language dictionaries.

The need for an orientation phase in the Think-Aloud Protocol is proposed by Okuyama & Igarashi (2007). In the current study, participants were asked to imagine they were in a supermarket on their regular grocery-shopping trip and to describe their thoughts while walking through the supermarket aisles selecting goods.

For the TAP part of the exercise, participants were shown the ODNZSL web page and were directed to use the dictionary as they normally would (or if they were not currently dictionary users, to treat this activity as if they were looking up information in a real situation). No specific task instructions were given, but participants were asked to look up at least three items. Both the screen and the participant were recorded. I remained present in the room during the TAP to deal with technical issues and to prompt participants to ‘keep talking’ if necessary.

Since some of the participants were deaf and would be using New Zealand Sign Language during the TAP, several modifications to the procedure were considered. ‘Thinking aloud’ may not be a feature of sign languages; although there is some evidence for a sign language-based articulatory rehearsal loop equivalent to a ‘phonological loop’ in spoken languages (Wilson & Emmorey, 1997), one’s own signing is most likely not observed as often, or in the same way, as hearing one’s own voice. Also, while navigating through the dictionary, a mouse or keyboard has to be used, which restricts the use of the hands for articulating at the same time. Since I am a fluent NZSL user myself, I sat opposite the deaf participant and provided minimal feedback cues (e.g. head nods) to encourage ongoing talk. I made no other comments. Deaf participants were also encouraged to articulate their thoughts before carrying out an action on the keyboard or mouse. All TAPs were recorded on video.

The follow-up interview probed further into participants’ use of the ODNZSL in this instance and in general. Participants were asked to pinpoint information in the dictionary that they regularly use and that which they do not use at all; they were also prompted to explore any problems that they experienced either during the TAP or during their own use of the ODNZSL. Finally, participants were asked to name features that their ideal dictionary would include.

5. Results and Discussion

5.1 Who Are the Users?

In line with patterns for other online dictionaries (Johnsen, 2005), the ODNZSL experienced growth in both the number of sessions and the number of users every year since its inception. The proportion of new users continued to rise (see Table 2), suggesting that while the ODNZSL attracted further interest, in most cases this did not develop into regular dictionary use. We should bear in mind that the log file data may mistakenly identify returning users as new users because they visit the site from a new device or because they have cleared their cookies. However, there are also

societal factors that may have had an influence on this changing user profile. The 2013 New Zealand Census (Statistics New Zealand, 2013) noted a drop in the number of people who indicated they could have a conversation “about a lot of everyday things” in NZSL (from 24,084 in 2006 to 20,244 in 2013). A number of reasons for this decrease are noted in McKee (2014) and include a lack of support for NZSL in mainstream schools, few opportunities for deaf children to communicate with other deaf peers, and few opportunities for families to learn NZSL. Factors such as these indicate that there may now be fewer learning environments that would support regular dictionary use. Paired with this decrease, however, is a rise in awareness of NZSL by the general public. McKee (2014) also notes an increase in visibility of a ‘Deaf voice’ on the internet.

	Apr - Jun 2012	Apr - Jun 2013	Apr - Jun 2014
New users	8,629 (35.9%)	11,681 (37.2%)	14,567 (45.9%)
Returning users	15,390 (64.1%)	19,690 (62.8%)	17,186 (54.1%)

Table 2: New vs. returning users to <http://nzsl.vuw.ac.nz>

Further support for the dictionary receiving a high level of casual interest but fewer ‘serious’ dictionary consultations comes from an examination of the frequency and page depth statistics. A total of 45.88% of visitors were new users and therefore had only visited the website once. A further 22.07% had visited less than five times, showing that even among visitors logged as ‘returning users’, there are a large number of casual users. The ODNZSL has a smaller number of highly regular users: 2.88% had visited the site more than 200 times, and a further 1.45% had made between 100–200 visits. Returning users viewed more pages per visit than new users (11.14 vs. 8.81, respectively), indicating that on return visits, users engaged with the website in more depth. A total of 28.2% of users left the website after only viewing a single page, and new users were more likely to do so. At the other end of the spectrum, 13.46% of all visits involved viewing 20 or more pages. These in-depth users were more likely to be returning visitors. From the log files, it can be concluded then that although the majority of visitors to the ODNZSL are new users who do not engage with the site in much depth, there exists also a sizeable minority of highly regular users who carry out multiple queries each time they visit.

Similar patterns of usage were reported in the interview data. Non-professional dictionary users who were not involved in formal language learning were aware of the existence of the ODNZSL but had not used the dictionary beyond an occasional browse out of curiosity. Deaf NZSL users said that they very rarely used the ODNZSL to look up signs or English words for themselves, although in their role as language teachers (both teaching classes and informally ‘teaching’ friends, colleagues

or parents of deaf children) they were frequent dictionary users. In this case, they would look up known signs to add to a vocabulary sheet, but would not look at the entry content in any detail. Responses from beginner and intermediate learners in NZSL classes indicated that they were the most regular dictionary users and looked up several signs daily. The two sign language interpreters in this study (who can be seen both as advanced language learners and as professional users) stated that they only occasionally used the ODNZSL.

5.2 Motivations for Using the Dictionary

Although log files cannot directly reveal users' reasons for using a dictionary, some inferences can be made from examining how they arrived at the dictionary website. The largest source of traffic (65.0%) was through the use of search engines, mainly Google. Less than a quarter of visitors arrived at the dictionary website directly (through typing in its URL or having the page bookmarked). Although it may seem more likely that returning users will be more familiar with the website and will therefore access it directly, in fact they were only slightly more likely to do so than new users (22.68% vs. 20.82%). Other traffic showed a sharper contrast, with new users making up the majority of referred (11.79% new vs. 6.98% returning) and social network traffic (6.06% new vs. 2.26% returning).

The search terms that result in a visit to the ODNZSL show that many users may not be looking for the dictionary specifically. 'NZSL dictionary' was only the third most common search term, with the majority of users searching for more generic terms such as 'NZSL' or 'NZ sign language'. Other common search terms were 'learning NZSL', 'basic sign language', 'NZSL alphabet' and various permutations of 'how do you say x in sign language'.

Reasons participants gave for looking up information during the TAP comprised both communicative and cognitive situations (Tarp, 2009). The TAP did not involve a particular task: participants were left to decide which information to look up. This unguided exercise probably encouraged general browsing of the ODNZSL; many searches were sparked by the participant speaking an English word during the TAP and then wondering how this word was expressed in NZSL; others spotted interesting signs that were not related to their original search in the results and followed through. While this was not an authentic dictionary usage situation, participants also mentioned using the ODNZSL in this way outside of the exercise. An often-mentioned 'cognitive situation' was looking up signs that had previously been learned or seen for rehearsal.

Most of the communicative situations involved language production rather than reception. Users mentioned wanting to find vocabulary to have a conversation with a deaf person. For beginners, this involved looking up words or phrases to do with greetings and introductions and themes such as food or family. Intermediate learners

said they often prepared a conversation topic in advance for classes or when they knew they were going to meet a deaf person. They wanted to broaden what they could talk about by looking up new signs around a theme. This included looking up grammatical and variation information as well. One deaf user looked up information in the other language direction, i.e. wanting to express a known sign in English. Looking up signs for reception was limited to classroom situations such as translation exercises or watching a video conversation. In real-life situations, participants said they would usually clarify the meaning with the signer on the spot rather than consulting the dictionary.

The authoritative role that dictionaries have traditionally played was also evident. Many users were aware of the relatively high levels of regional and age variation in NZSL and used the dictionary to confirm whether a sign they had observed or had been taught was in common use. A deaf sign language teacher preferred to choose the particular sign variants in the ODNZSL for inclusion in teaching resources, even when she might use a different variant herself.

5.3 How Is the Dictionary Used?

5.3.1 Searching

One of the original features of the ODNZSL is its choice of search direction, allowing users to either search by word or by sign features. In a pre-compilation survey (McKee & Pivac Alexander, 2008), 45% of potential users said they would use the search by sign features alongside other methods. Log file data show that actual user behaviour is rather different: the overwhelming majority of searches (98%) were a search by English/Māori word. Searches by sign features only accounted for 0.7% of all searches, with the remainder constituting advanced searches. Although the log data does not distinguish between English and Te Reo Māori word searches, there were few of the latter, and the most frequently looked up Māori words are considered to be borrowings into the New Zealand English lexicon such as ‘kia ora’ (a greeting) or ‘whānau’ (extended family).

Together, the top 25 search terms in the ODNZSL (Figure 5) constituted 6.8% of all searches. This figure is slightly higher if misspellings and phrases containing the same words (e.g. ‘my name is’) are included. Beginner participants in the TAP looked up similar words and phrases, as did deaf NZSL teachers preparing for a lesson. The majority of these searches are highly frequent words or phrases in English. De Schryver & Joffe (2004) noted a similar trend in their data.

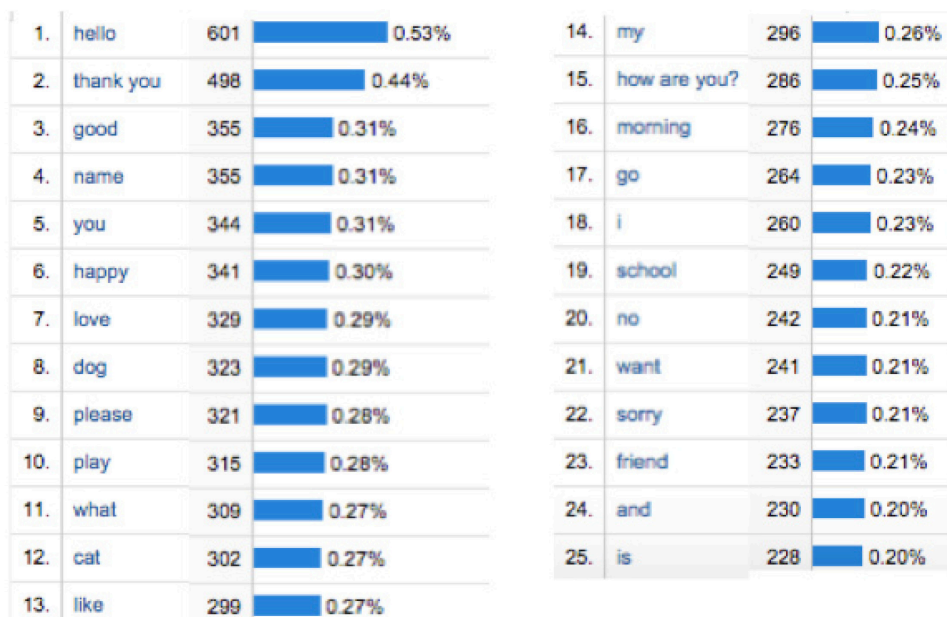


Figure 5: Most frequent search terms

Many TAP participants tried to carry out at least one search by sign features, but said they rarely or never used this search direction in their normal dictionary consultation. The exception was that some learners in classes had been given specific tasks and had been shown by their teacher how to use this search facility. Lack of familiarity with the handshape and location parameters of a sign are a barrier to the effectiveness of this search: beginner learners in the TAP said they did not know where to start, and other participants (including a deaf NZSL user who tried to use this search method to find an English equivalent for a sign) talked about the difficulties of isolating the specific features of a sign in motion.

Taken together, these findings lend further support to the conclusion that the ODNZSL's main user group is (hearing) people with an interest in learning the language, mostly at a beginner level, who mainly consult the dictionary for language production.

Over the three month period, all 4,000 entries in the ODNZSL were visited or showed up in search results at least once. This coverage demonstrates that the current dictionary content, with its focus on the most frequent signs and words, is in line with the needs of its main user group. However, given that more than 21,000 different search terms were looked for, this indicates that there are also unmet needs whereby either the dictionary content does not include the searched-for word, or the search does not identify the target.

5.3.2 Search Results

In line with the findings of other studies (e.g. Lew, Grzelak & Leskowicz, 2014), users of the ODNZSL clicked on the first search result more than half of the time (52.82%), as compared to signs appearing in the second position (20.47%) and third position (9.89%). The number of clicks on signs appearing in lower positions steadily declined. It has to be considered that signs are more likely to be displayed in early positions, since all valid searches will have at least one search result but may not have more. The same behaviour can be seen to occur for individual search results, however. For example, the most popular search query, ‘hello’, returns three different signs. The first search result made up 60% of the clicks, whereas both the second and third search results were selected 20% of the time.

This preference for the first search result in the ODNZSL may not signify a lack of sense discrimination on behalf of the user. For example, the most frequently clicked search result for ‘fine’ was the second sign with the sense ‘alright, ok’ rather than the first sign that has the sense of a monetary fine or punishment.

Interestingly, there is some evidence that dictionary users avoided polysemous signs in favour of signs that have a single sense. An example is that in the search results for the query ‘cat’, the most frequent sign, which also has the general meaning ‘pet’, was not selected at all whereas the second search result was selected 147 times. Similarly, a general questioning sign with the sense ‘what’, ‘where’, or ‘why’ was passed over in the search results in favour of a less frequent sign with the single sense ‘what’.

5.3.3 The Dictionary Entry: Which Information Is Viewed?

Table 3 takes as a typical example the entry for ‘play’, as shown in Figure 4, to examine use of clickable elements in the entry.

As can be seen, not all page views involved further interaction with the more in-depth information on the page. The most used interactive element was the video of the sign in isolation. The ability to show signs dynamically on video rather than as a static image is hailed as one of the greatest advantages of online sign language dictionaries over printed ones (McKee & McKee, 2013). In the light of this it is interesting to find that only just over 36% of page views involved watching the video. This percentage may be somewhat lower than in other cases: the most viewed video (“how are you” – see Figure 5 – was clicked in 55.84% of all page views. Overall, the video showing the sign in isolation was viewed at least once for 93.81% of all entries, showing that this feature is on the whole well used. Example sentences were viewed considerably less often than the sign in isolation, as were slow-motion views of the videos. Hyperlinks to other content in the dictionary were used least often.

Element	Number of views	Percentage of page views that include a view of this element
Page views	263	100.00%
Video showing sign production	97	36.88%
Slow-motion video showing sign production	12	4.56%
Video example 1	9	3.42%
Slow-motion video example 1	6	2.28%
Video example 2	15	5.70%
Slow-motion video example 2	5	1.90%
Inflection hyperlink to glossary	0	0.00%
Hyperlinks to other signs in the example sentences	7	2.66%

Table 3: Views of the different elements for the entry ‘play’

5.4 Problems and Issues

Looking in more detail at the consultation process shows that users experience problems during different parts of the consultation. These problems can be broadly categorised as either having to do with dictionary navigation or dictionary content.

5.4.1 Dictionary Navigation

During the TAP, participants commented extensively on technical issues such as long loading times and glitches with video playback. If a page was not displayed in seconds, participants would lose patience and click on other parts of the page, try to reload, or give up on the search altogether. This behaviour has implications for the technical design of online dictionaries, especially sign language dictionaries that are required to deal with the smooth display of large quantities of videos.

Participants also experienced difficulties as a result of being unfamiliar with the dictionary interface. Problems with using the ‘Search by Sign Features’ interface were

discussed in the previous section; but additional problems were encountered with more usual web navigation devices. One participant who had not used the ODNZSL prior to the TAP spent some time trying to locate the search box and commented in the follow-up interview on the layout of the home page and the need for more prominent search facilities. Other participants missed information because the results display required scrolling down. Pagination of search results was also difficult to navigate. It is significantly of note that nearly all participants indicated that the ODNZSL is the first and only online dictionary they have used; in the context of learning other (spoken) languages, they used print dictionaries, and to look up information about English, a general Google search was used instead of consulting an English dictionary (whether in print or online).

Participants' interactions with the ODNZSL interface are coloured by their more general online experiences. Log file data on search terms entered in the ODNZSL show that users searched for extraneous information, such as song lyrics, names of famous people and other proper nouns; there were also instances of terms in languages other than the three languages of the ODNZSL. Within the boundaries of ODNZSL content, it was evident that participants saw the search box as a way of searching the entire site and not just an individual word. Search terms included semantic categories (e.g. 'Natural disasters'; 'personal qualities'; 'zoo animals') and searches for more general information about NZSL (e.g. 'Fingerspelling chart'; 'numbers').

The influence of generic web searches on dictionary interface expectations can also be seen in the way search terms were entered as natural language queries. Thus, we find searches for whole phrases such as 'my name is', 'you owe me chocolate', or 'the bird flew up in the tree', and searches for inflected word forms such as 'am', 'going', 'made', or 'days'.

A final problem with inputting a search query was misspelled or mistyped information. The ODNZSL uses predictive text in the search box to assist with this issue, and some participants acknowledged that this was an advantage of online dictionaries, although in the TAP the correction suggestions were sometimes overlooked.

5.4.2 Dictionary Content

As mentioned in the Introduction, sign language dictionaries have a relatively small content. The ODNZSL contains just over 4,000 lemmas and mainly covers the most frequently used signs and concepts. It is not surprising, then, that many of the 21,000 logged search terms did not find a match in the ODNZSL. Data on these failed searches can be used to identify so-called 'lemma lacunae' (Bergenholtz & Johnsen, 2007). Indeed, since this user study, several of these 'missing' signs, such as sign equivalents for 'turtle', 'pineapple' and 'slide' have been filmed and are currently being processed for appearing online. Other search terms that failed to bring up a

result may be more difficult to resolve. Firstly, there were searches for auxiliaries, modals and forms of the verb ‘to be’ that do not have a parallel in NZSL. Secondly, lower frequency English words were searched for, including words from more formal and technical registers (e.g. ‘Inebriated’, ‘totalitarian’, ‘prism’). Finally, some search terms were words that have only recently entered the English language and may not (yet) have an accepted equivalent in NZSL: e.g. ‘minecraft’, ‘unfriend’, and ‘onesie’.

TAP participants experienced problems once search results were displayed. All categories of participants, but especially beginner learners, found it difficult to distinguish between sign variants with the same English glosses. In the ODNZSL, the most frequently used sign is shown first in the search results; however, this was not always clear to users. Other variation information (such as age, regional or register variation) is provided, when available, in the notes for an individual entry. This requires users to click on each sign in the results in turn: a somewhat cumbersome process, especially when there are instances when the information on NZSL variation is not complete. This prompted users in the follow-up interview to request more information to be displayed in the search results.

Paired with this, however, is the issue of information cost (Nielsen, 2008). Participants commented on the grammar information in the tabs being too dense and mentioned giving up looking at search results when too many were returned at once (e.g. when searching for a very common topic or handshape).

6. Conclusion

Nesi (2013) states that “the aim of all studies of dictionary use is to discover ways to increase the success of dictionary consultation.” This paper has confirmed the assumption that online sign language dictionaries have diverse user groups and functions, and has looked at these user groups’ consultation behaviour and motivations for using the dictionary. With a better understanding of who the users are and what problems they experience, we can now turn to the question of whether online sign language dictionaries can be improved in order to meet their users’ needs.

Although casual, one-off users were found to make up the majority of ODNZSL visits, it is not towards this user group that possible changes to the dictionary should be aimed. Many of these casual users did not engage with the dictionary content in any great depth, and their visits to the ODNZSL do not reflect an ongoing authentic dictionary usage need. This high level of casual interest may nevertheless contribute to more general aims of sign language dictionaries such as supporting recognition and public awareness of the language.

Looking beyond this casual use, distinct user profiles emerged. While there was a common need of dictionary information for language production, there were also differences in the depth of information users wished to access and the frequency level of the signs they wanted to look up. Beginner language learners looked for common

phrases and frequent vocabulary and were likely to be confused by the dictionary layout and overwhelmed by excessive information. Intermediate learners, by contrast, were the most experienced in navigating the website, but wanted to look up less frequent vocabulary and requested more in-depth information on grammar and variation in order to make sense of the search results. A solution to balancing these conflicting needs would be to explore the possibility of customising the display of dictionary content for different users, as mentioned in the Introduction. By displaying the most looked for information early on in the search results (e.g., by allowing users to play the main sign video directly from the search results without needing to click through), beginner language learners can be shown the essential information in a way that keeps the information cost low. More advanced users can then click through to more detailed information.

Improvements to general navigation of the ODNZSL would also lead to increased success. However, any changes to scrolling, pagination of search results, and video display need to be weighed up against possible increased page loading times.

The ODNZSL search methods may have to be adjusted in acknowledgment of the changing behaviour of dictionary users in the digital age that was also noted by Lew & de Schryver (2014). Users expect to be able to enter natural language queries and inflected forms, for example. Adding lemmatisation of the English glosses in the ODNZSL and allowing searches for other fields (such as topics or grammar information) within the same search box may improve the ‘hit’ rate of search results. Although the ‘Search by Sign Features’ was user-tested before implementation, this search method currently has a very low success rate. Providing training for users to become familiar with this novel search method may be the first step to improvements.

In terms of dictionary content, it is unlikely that users’ desire for additional comprehensive variation and usage information and coverage of technical and infrequent vocabulary can be met in the short term. However, ongoing analysis of log files can identify those missing items that could and should be added to the dictionary.

This paper has shown that user research into online sign language dictionaries has a valuable contribution to make, not only to the dictionary itself but to our knowledge about dictionary users in the digital age and how they interact with novel dictionary formats and features.

7. Acknowledgments

Presentation of this paper has been supported by a grant from the Faculty of Humanities and Social Sciences, Victoria University of Wellington, New Zealand.

8. References

- Bergenholtz, H., & Johnsen, M. (2007). Log files can and should be prepared for a functionalistic approach. *Lexikos*, 17, pp. 1–20.
- De Schryver, G., & Joffe, D. (2004). On how electronic dictionaries are really used. In *Proceedings of the eleventh EURALEX International Congress*, (pp. 187–196). Lorient: Université de Bretagne-Sud.
- Johnsen, M. (2005). *Logfiler som leksikografisk analyseinstrument og hjælpeværktøj*. Masters Thesis, Handelshøjskolen i Århus, Denmark. Retrieved from <http://pure.au.dk/portal-asb-student/files/2040/000139835-139835.pdf>
- Johnston, T. A., & Schembri, A. (1999). On Defining Lexeme in a Signed Language. *Sign Language & Linguistics*, 2(2), 115–185. doi:10.1075/sll.2.2.03joh
- Kennedy, G. D., Arnold, R., Fahey, S., & Moskovitz, D. (Eds.) (1997). *A dictionary of New Zealand Sign Language*. Auckland: Auckland University Press with Bridget Williams Books.
- Kennedy, G. D., McKee, D., Arnold, R., Dugdale, P., Fahey, S., & Moskovitz, D. (eds.) (2002). *A concise dictionary of New Zealand Sign Language*. Wellington: Bridget Williams Books.
- Konrad, R. (2012). *Sign language corpora survey*. Hamburg: Institute for German Sign Language and Communication of the Deaf, University of Hamburg. Retrieved from http://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/SL-Corpora-Survey_update_2012.pdf
- Kristoffersen, J. H., & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 293–318). Oxford: Oxford University Press.
- Lew, R., & De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), pp. 341–359.
- Lew, R., Grzelak, M., & Leszkowicz, M. (2013). How dictionary users choose senses in bilingual dictionary entries: An eye-tracking study. *Lexikos*, 23, pp. 228–254.
- McKee, R. (2014, November). *Assessing the vitality of NZSL*. Paper presented at the Language and Society Conference 2014, University of Waikato, Hamilton, New Zealand.
- McKee, R., & McKee, D. (2013). Making an online dictionary of New Zealand Sign Language. *Lexikos*, 23, pp. 500–531.
- McKee, D., McKee, R., Pivac Alexander, S., Pivac, L., & Vale, M. (2011). *Online Dictionary of New Zealand Sign Language*. Wellington: DSRU, Victoria University of Wellington. Accessed at <http://nzsl.vuw.ac.nz>
- McKee, D., & Pivac Alexander, S. (2008). *NZSL Online Dictionary project 2008 - 2011: User requirements survey report*. Wellington: DSRU, Victoria University of Wellington.

- Moskovitz, D. (1994). The Dictionary of New Zealand Sign Language user requirements survey. In I. Ahlgren, B. Bergman, & M. Brennan (eds.), *Perspectives on sign language: Papers from the Fifth International Symposium on Sign Language Research: held in Salamanca, Spain, 25-30 May 1992 Volume 2: Perspectives on sign language usage* (pp. 421 – 442). Durham: International Sign Linguistics Association / Deaf Studies Research Unit, University of Durham.
- Müller-Spitzer, C. (2013). Contexts of dictionary use. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* (pp. 6–13). Ljubljana / Tallinn: Trojina, Institute for Applied Slovene Studies / Eesti Keele Instituut.
- Nesi, H. (2013). Researching users and uses of dictionaries. In H. Jackson (Ed.), *The Bloomsbury Companion to Lexicography* (pp. 62–74). London: Bloomsbury.
- Nielsen, S. (2008). The effect of lexicographical information costs on dictionary making and use. *Lexikos*, 18, pp. 170–189.
- Okuyama, Y., & Igarashi, H. (2007). Think-aloud protocol on dictionary use by advanced learners of Japanese. *The JALT CALL Journal*, 3(1-2), pp. 45–58.
- Prinsloo, D. J. (2012). Electronic lexicography for lesser-resourced languages: The South African context. In S. Granger & M. Paquot (eds.) *Electronic Lexicography* (pp. 119–144). Oxford: Oxford University Press.
- Schermer, G. M. M. (2006). Sign Language: Lexicography. In *Encyclopedia of Language and Linguistics, 2nd Edition*. Amsterdam: Elsevier Ltd.
- Schmalings, C. H. (2012). Dictionaries of African sign languages: An overview. *Sign Language Studies*, 12(2), pp. 236–278.
- Statistics New Zealand. (2013). 2013 *Census QuickStats about culture and identity*. Retrieved from <http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-culture-identity/languages.aspx>
- Tarp, S. (2009). Reflections on lexicographical user research. *Lexikos*, 19, pp. 275–296.
- Varantola, K. (2002). Use and usability of dictionaries: Common sense and context sensibility? In *Lexicography and natural language processing: A festschrift in honour of BTS Atkins*. Stuttgart: Euralex.
- Wilson, M., & Emmorey, K. (1997). A visuospatial “phonological loop” in working memory: Evidence from American Sign Language. *Memory and Cognition*, 25(3), 313–320.
- Woll, B., Sutton-Spence, R. & Elton, F. (2001). Multilingualism: The global approach to sign languages. In C. Lucas (Ed.), *The sociolinguistics of sign languages* (pp. 8–32). Cambridge: Cambridge University Press.
- Zwitserlood, I. (2010). Sign language lexicography in the early 21st century and a recently published dictionary of Sign Language of the Netherlands. *International Journal of Lexicography*, 23(4), pp. 443–476.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Using a Maximum Entropy Classifier to link “good” corpus examples to dictionary senses

Alexander Geyken¹, Christian Pölitz², Thomas Bartz²

¹ Berlin-Brandenburg Academy of Sciences, Jägerstr. 22/23, D-10117 Berlin, Germany

² Technische Universität Dortmund, Fakultät für Informatik, Otto-Hahn-Str. 14, 44227
Dortmund, Germany

E-mail: geyken@bbaw.de, {poelitz,bartz}@tu-dortmund.de

Abstract

A particular problem of maintaining dictionaries consists of replacing outdated example sentences by corpus examples that are up-to-date. Extraction methods such as the good example finder (GDEX; Kilgarriff, 2008) have been developed to tackle this problem. We extend GDEX to polysemous entries by applying machine learning techniques in order to map the example sentences to the appropriate dictionary senses. The idea is to enrich our knowledge base by computing the set of all collocations and to use a maximum entropy classifier (MEC; Nigam, 1999) to learn the correct mapping between corpus sentence and its correct dictionary sense. Our method is based on hand labeled sense annotations. Results reveal an accuracy of 49.16% (MEC) which is significantly better than the Lesk algorithm (31.17%).

Keywords: WSD; maximum entropy; collocations; legacy dictionaries; example sentences

1. Introduction

Keeping dictionaries up-to-date is a very time consuming task that involves regular checks throughout the entire dictionary for all types of lexicographic information. One particular problem consists of replacing outdated example sentences in the dictionary by suitable corpus examples that are up-to-date or of adding corpus examples to new entries. In general today’s corpora of several billion words of text are too large to allow for regular manual inspection of the entire set of frequent words. Indeed, Moon (2007) states that the 25,000 most frequent words in English all have frequencies higher than one per million tokens. For a one billion word corpus this would amount to analysing 1,000 corpus hits. Since many of today’s corpora exceed 10 billion words, this would quickly result in numbers that are no longer feasible within the budget and time constraints of today’s lexicographic projects. Several methods to automate this task have been developed, the most popular being the “good” example finder (GDEX; Kilgarriff et al., 2008). GDEX is a rule based software tool that suggests “good” corpus examples to the lexicographer according to predefined criteria such as sentence length or word frequency, or lexicogrammatical criteria such as the presence/absence of pronouns or named entities. The goal of GDEX is to reduce the number of corpus examples to be inspected by extracting only the n-“best” examples. The ideas of

GDEX have been used for languages other than English (Kosem et al., 2011, for Slovene) and have given rise to different implementations (Didakowski et al., 2012, for German; Volodina et al., 2012, for Swedish).

The goal of our work is to extend GDEX to polysemous entries. More precisely we attempt to link a given corpus sentence extracted by GDEX to its appropriate dictionary sense (in the case of a polysemous entry). The method we employ is a machine learning technique (cf. section 3). The main hypothesis of our work is that the results of our machine learning approach improve if the linking is not only performed to a dictionary sense represented by a sense number and a definition but rather on the full dictionary sense. In the case of a large reference dictionary this includes the example sentences, citations and the set phrases.

The remainder of this article is structured as follows. In section 2 we present related work in the field of Word Sense Disambiguation. In section 3 we describe the resources we use. The machine learning approach is described in section 4. We then report on an experiment with 100 polysemous and frequently used German words (section 5). The last section discusses the results and presents some ideas for further research.

2. Word Sense Disambiguation

Word Sense Disambiguation (WSD) plays an important role in Natural Language Processing. Many approaches have been carried out in this area. Starting from the pioneer work of Lesk (1986), automatic methods to assign text examples to possible senses given from a dictionary for instance have become increasingly important. The first approaches for assigning senses to given text examples used pure word overlaps between the text and definitions for the senses. These definitions can be for instance from a dictionary or, as proposed by Vasilescu et al. (2004), from synsets from WordNet. Besides pure word overlaps to assign senses to texts, knowledge based methods have also proven successful. Navigli and Velardi (2005) introduce structural and rule based representations of possible senses to efficiently map them to text examples. More recently, machine learning approaches based on supervised methods have emerged in WSD, including Neural Networks (Moony, 1996), Näve Bayes (Patterson, 2007), Ensemble Methods (Escudero, 2000) and Support Vector Machines (Keok & Ng, 2002). A detailed introduction to WSD and a survey on the different methods to solve it can be found in Navigli (2009).

3. Resources

The resources used for the work presented here are threefold: a dictionary, a large database of collocations and GDEX. All these resources are part of the DWDS (Digitales Wörterbuch der deutschen Sprache, Digital Dictionary of the German Language), a project of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). DWDS is a long term project of BBAW. Its goal is to compile a large

aggregated word information system based on large legacy dictionaries, large corpora, word statistics and automated methods to speed up the compilation process (Geyken, 2014).

The dictionary used for our work is the large “Wörterbuch der deutschen Gegenwartssprache” (dictionary of the German contemporary language, WDG, www.dwds.de), a synchronic dictionary of 4,800 pages with 120,000 keywords, compiled between 1961 and 1977. The electronic version of the WDG is encoded in TEI. Each entry consists of a form and a sense part; the sense comprises definitions, diasytematic markers, made-up examples and corpus examples. Relevant to our work are the following components of the sense element: definition, examples made-up by the lexicographer and citations from corpora. We will call these components *dictionary sense* in the remainder of this article. An example for the entry *Leiter* (en. leader, ladder, conductor) drawn from the WDG is given in Table 1. Only sense 2 is fully expanded; for senses 1 and 3, definitions only are provided. The full entry can be looked up at the project’s website (www.dwds.de).

<p>Sense 1: <i>Gerät aus Holz oder Leichtmetall</i> (en.: device made of wood or light metal)</p>
<p>Sense 2: <i>jmd., der etw. leitet, an der Spitze von etw. steht</i> (s.o. who directs sth., who is at the top of sth.)</p> <p>made-up examples and constructions:</p> <p><i>ein technischer, kaufmännischer, künstlerischer, staatlicher, kommissarischer Leiter</i> (a technical, commercial, artistic, governmental, acting director)</p> <p><i>der Leiter einer Baustelle, Abteilung, Schule, Delegation, Touristengruppe, Behörde, Expedition, eines Krankenhauses, Unternehmens</i> (the head of a construction site, department, school, delegation, tourist group, authority, expedition, hospital, company)</p> <p>corpus example:</p> <p><i>Heut bin ich im Funk Leiter vom Dienst</i> (Today I am in the radio manager on duty) [Klepper, J., Schatten, 1960, p. 56]</p>
<p>Sense 3: <i>Stoff, der Energie leitet</i> (substance that passes energy)</p>

Table 1: entry *Leiter* in the WDG

The second resource is the DWDS-Wortprofil (Didakowski & Geyken 2012), an implementation of the sketch engine (Kilgarriff et al., 2004) for German. DWDS-Wortprofil provides co-occurrence lists for twelve different grammatical relations (Tables 2 and 3) and links them to their corpus contexts. The co-occurrence lists and their ordering are based on statistical computations over a German corpus of currently 1.783 billion tokens. For syntactic annotation the rule based dependency

parser SynCoP (Syntactic Constraint Parser; Didakowski, 2008) is used. A grammar for the SynCoP parser was developed which is designed for the specific relation extraction task. Therefore, issues like the attachment of sub-clauses or specific rare syntactic phenomena are not dealt with in this grammar.

syntactic relation	part-of-speech tuples
accusative object	{<verb,noun>}
active subject	{<verb,noun>}
adjective attribute	{<noun,adjective>}
coordination	{<verb,verb>,<noun,noun>,<adjective,adjective>}
dative object	{<verb,noun>}
genitive attribute	{<noun,noun>}
modifying adverbial	{<verb,adverb>,<adjective,adverb>}
passive subject	{<verb,noun>}
predicative complement	{<noun,noun>,<noun,adjective>}
verb prefix	{<verb,prefix>}

Table 2: binary relations

syntactic relation	part-of-speech tuples
comparative conjunction	{<noun,conjunction,noun>,<verb,conjunction,noun>}
prepositional group	{<noun,preposition,noun>,<verb,preposition,noun>}

Table 3: ternary relations

As a result of the statistical computations, the database contains 11,980,910 distinct co-occurrence pairs (types) with a total of 257,402,167 tokens. The DWDS-Wortprofil is part of the web platform of DWDS and is continually extended with new corpora. In its current version it is possible to query 104,704 different lemma/part-of-speech pairs.

The third resource used for this work is a set of corpus sentences. We use an implementation of GDEX for German (Didakowski et al., 2012) to extract the n-best corpus sentences for a given word. The underlying text corpora for this extraction task are the corpora of the DWDS project. The corpora comprise a total of 4 billion words and consist of four subcorpora: 1) the DWDS-Kernkorpus of the 20th/21st century, a balanced reference corpus of 110 million tokens (Geyken, 2007); 2) a balanced historical corpus currently comprising of 120 million tokens for the period from 1600 to 1900, compiled at the BBAW for the project *Deutsches Textarchiv* (DTA, German Text Archive, www.deutschestextarchiv.de); 3) a corpus of ten influential national daily and weekly newspapers, which currently consists of 3.5 billion tokens in 8 million

documents; and 4) several special corpora with a total of 200 million tokens, including a large blog corpus, a corpus of contemporary interviews and a corpus of subtitles.

4. Method

The standard approach by Lesk (1996) to match a text to senses with given definitions is to count the words that both definitions and texts have in common. The higher the number of common words, the more likely that the text will have the corresponding sense. Formally, for a text $t = w_1 \dots w_k \dots w_n$ being the context of a key word w_k , a set of applicable senses $\{s_i\}$ with corresponding definitions $\{d_i = w_1^i \dots w_{m_i}^i\}$, the standard Lesk algorithm calculates the numbers n_i that are the sum of common words from t and d_i . We assign the sense s_j to text t with $n_j = \max_{s_i} n_i$, for all applicable senses s_i . A major drawback of this approach is that for shorter texts and definitions the chance to have overlap decreases.

A simple extension of the Lesk method to lexical databases was proposed by Vasilescu et al. (2004). The authors extend the concept of overlap of words from sense definitions and key word context (i.e. a corpus sentence) to WordNet. A drawback of their approach is that they can only match to WordNet senses and not to arbitrary dictionary entries.

We propose to extend the Lesk algorithm in such a way that we do not only count the number of intersecting words, but also all words that are statistically salient co-occurrences (i.e. with a $\logDice > 0$) in the DWDS Wortprofil, as explained in section 3. These sets of co-occurrences, henceforth called word-profiles, are computed for all content words (nouns, verbs, adjectives and adverbs) of all dictionary senses of a given headword; i.e. the definition, the example sentences and the corpus citations that are part of the legacy dictionary. This results in a mapping from each headword to a list containing all statistically salient co-occurring words from the word profiles together with the corresponding \logDice values. The match from a corpus sentence extracted by GDEX to a dictionary sense is performed by matching all word profiles from the content words in the corpus sentence with the dictionary senses. This means, for each word w_i in the corpus sentence and each word w_l^i in a dictionary sense s_j , we count the number of common words in the two corresponding word profiles weighted by the \logDice from the word profile of the word from the key word context. Finally, we sum up all aggregated \logDice s. The “best” dictionary sense for a given corpus sentence is the one that corresponds to the largest sum (compared to the other dictionary senses). This extension of the Lesk algorithm is henceforth called Lesk_{ext} . An example of how Lesk_{ext} is performed on the dictionary example *Leiter* (cf. Table 1 above) is given in Tables 4 and 5. Table 4 illustrates the \logDice s for the collocations that the two nouns *Spitze* (top) in the dictionary definition and *Verantwortung* (responsibility) in the corpus example have in common. Table 5 displays the total number of collocations as well as the sum of the \logDice values for both, sense 1 and sense 2.

	dictionary definition	Corpus example
	Leiter, sense 2 “ <i>jmd. der etw. leitet, an der Spitze von etwas steht</i> ” (so. who leads, is in the top position of sth.)	“Aufgabe der HI ist es nicht, den Leitern diese Verantwortung abzunehmen.” (It is not the task of the HI, to remove the responsibility from the leaders.)
content words	Spitze	Verantwortung

collocations in common/relation	logDice	e.g. “Spitze”	e.g. “Verantwortung”	logDice
adjective attribute	4.72	<i>international</i> (international)		5.08
	1.79	<i>gesellschaftlich</i> (social)		8.23
	2.93	<i>alleinig</i> (sole)		8.87
	Σ 9.44			Σ 22.18
genitive attribute	6.60	<i>Unternehmen</i> (enterprise)		5.48
	5.99	<i>Aufsichtsrat</i> (directorate)		5.80
	5.29	<i>Politik</i> (politics)		6.00
	Σ 17.88			Σ 17.28
predicative complement	1.68	<i>hoch</i> (high)		3.60
	3.14	<i>deutlich</i> (clear)		3.97
	Σ 4.82			Σ 7.57
		...		

Table 4: Example: Mapping of dictionary examples and corpus sentences (identical senses: head/leader)

	dictionary example	corpus sentence	logDice (sum)
sense 2	head/leader	head/leader	798.22
content words (86 collocations in common)	„Spitze“ (top position)	„Verantwortung“ (responsibility)	
sense 1	ladder	head/leader	62.95
content words (8 collocations in common)	„hoch“ (high)	„Verantwortung“ (responsibility)	

Table 5: Example: Aggregated logDice values

The DWDS-Wortprofil also specifies the syntactic relation between a word and its co-occurrences. We propose to aggregate the logDice values for co-occurrences from the word profiles as before, but now for each of the syntactic relations individually in order to measure the impact on individual syntactic relation. Thus, we can measure the impact on the type of syntactic relation of the matching process to its corresponding dictionary sense. As mentioned above there are 10 binary relations and two ternary relations in the DWDS-Wortprofil. This means we are not getting a single sum after the match of all word profiles but a vector with the sum of the aggregated logDices for each relation. Next, to assign the best weight to each syntactic relation we use a Maximum Entropy Classifier (Nigam et al., 1999) that models the probability distribution of a given context and a given definition from the senses. Formally, the probability of a sense s for a given corpus sentence t is defined as $p(s|t) = e^{\omega' \varphi(s,t)} / Z$ for a feature vector $\varphi(s,t)$, a weight vector ω and the normalization constant Z . Each feature in $\varphi(s,t)$ is the sum of the logDices of the matching words for the dictionary sense s and (sentence) context t for a relation as explained above. We find the optimal weights ω by maximizing the joint probability over a training set $\{(S_k, T_k)\}$ of key word contexts T_k for a given number of key words $w_k \in K$ with hand labeled senses S_k with given definitions. The optimal ω is the parameter vector that maximizes the log likelihood of our given training data. The resulting optimization problem is defined in the following way:

$$= \operatorname{argmax}_{w_k \in K} \left\{ \sum_{w_k \in K} \log(S_k | T_k, \omega) = \sum_{(s,t) \in (S_k, T_k)} \log \left(\frac{e^{\omega' \varphi(s,t)}}{\sum_{s'} e^{\omega' \varphi(s',t)}} \right) \right\}$$

We solve the above optimization problem with a standard BFGS solver (Broyden–Fletcher–Goldfarb–Shanno algorithm) that performs a quasi-Newton optimization as for instance proposed by (Byrd et al., 1995). For the sense association example in Table 4, the MEC classifier provides a probability distribution stating that sense 2 is selected with a probability of 0.9 whereas sense 1 has only a 0.1 chance.

5. Experiment

In an experiment we selected 100 highly polysemous headwords (75 nouns, 25 verbs). These words have a total of 857 fine-grained senses (314 main or coarse-grained senses) in our dictionary (WDG). The list of headwords with English translations of the most prominent sense of the item in parenthesis is the following:

ablösen (supersede), Achse (axis), Adresse (address), Agent (agent), anschließen (connect), Ansicht (view), anstellen (do), Atmosphäre (atmosphere), aufheben (cancel), Aussprache (pronunciation), ausziehen (move out), Bank (bank), beschreiben (describe), Betrieb (operation), Blase (bubble), eingehen (enter), Einheit (unit), Einsatz (use), Eis (ice), eröffnen (open), Fall (case), feststellen (find), Film (movie), finden (find), Flucht (flight), Gehäuse (housing), Gemeinde (community), Gericht

(court), Geschichte (history), Grund (reason), handeln (act), Höhe (height), Interesse (interest), Kapelle (chapel), Kasse (checkout), klappen (fold), Kopf (head), Körper (body), kosten (cost), Leder (leather), Lehre (teaching), Leiter (ladder), lesen (read), Mal (time), Mark (marrow), Markt (market), Masche (stitch), Maschine (machine), Messe (fair), Mine (mine), Mission (mission), Moment (moment), Morgen (morning), Mutter (mother), nachsehen (check), Operation (operation), Parkett (parquet), passen (match), passieren (happen), Passion (passion), Pause (pause), Pension (guesthouse), Phase (phase), Piste (runway), Praxis (practice), Probe (sample), Prozess (process), riechen (smell), Rolle (role), Satz (sentence), Schatz (treasure), Scheibe (disc), scheinen (appear), Schloss (castle), Sitz (seat), sitzen (sit), Sohle (sole), Stärke (strength), Stelle (location), Steuer (tax), Stimme (voice), stimmen (vote), streichen (paint), Strom (current), Tafel (blackboard), Theater (theater), Ton (clay), Tonne (ton), Truppe (troops), Verfahren (method), Verfassung (constitution), Verhältnis (relationship), Vermittlung (mediation), versichern (reassure), versprechen (promise), Vorstellung (representation), Welle (wave), Wende (turn), Zelle (cell), zugeben (admit)

For each headword, we extracted 20 sentences using the GDEX method (Didakowski et al., 2012) applied to the DWDS corpora (www.dwds.de). All 2,000 example sentences were manually annotated with their corresponding dictionary senses by two annotators. We randomly split the example sentences into a training set of 750 sentences and test set of 1,250 sentences and we applied the Lesk algorithm and the Maximum Entropy Classifier method, as described in section 4.

6. Results and Discussion

The results of our experiment show that the Maximum Entropy Classifier significantly improves on the Lesk_{ext} algorithm. Both methods were applied on the same training data using the same resources, including the data from the DWDS-Wortprofil. As stated above, we have an average of 8.57 fine-grained senses. Thus, a random selection as base-line would predict an accuracy rate of 11.67%. With the Lesk algorithm based on intersection of co-occurring words of the DWDS-Wortprofil we achieve an accuracy of 31.17% for the test set. The Maximum Entropy Classifier further optimizes Lesk_{ext} by taking into account the specific syntactic relations as well as the weights provided by the logDice values that are used to compute the co-occurrence strength between the headword and its collocate. The application of the Maximum Entropy Classifier provides an accuracy of 49.16% for fine-grained senses in our test set. There are also differences between the accuracy of nouns (51.8%) and verbs (44.24%). The lower accuracy for verbs is due to the fact that the semantic information of the WDG is poorer for verbs, i.e. it frequently uses only placeholders (such as s.o., sth.) in its sense descriptions.

We have also investigated the impact of the sense granularity. As stated above there are 314 coarse-grained senses for our training set. Hence the base-line would predict an

accuracy of 31.8%. If applied on coarse-grained senses, the accuracy of the Maximum Entropy Classifier augments by about 7%, i.e. 55.74%, instead of 49.1% for fine-grained senses. Again, there are differences between nouns and verbs: MEC provides an accuracy of 58.69% for nouns but only 46.88% for verbs.

Another result concerns the quality of GDEX that we evaluated indirectly by the inter annotator agreement. For our test set we obtain an inter annotator agreement (IAA) of $\kappa = 0.78$ for fine-grained senses. κ for coarse-grained senses rises by 7% to arrive at 0.85. These κ values seem high compared to other WSD tasks. One reason for this finding may be that the examples extracted by our GDEX extractor are more homogeneous than a selection by “chance”. Indeed, for our data we found that the main sense (that occurs most frequently) is attributed to an average of about 11 out of 20, i.e. 55% ($\pm 2\%$ standard deviation), of the examples for each headword. The second most frequent senses cover only about four to five examples ($22.4\% \pm 1.2\%$); the other senses even fewer (0–2 examples, $9.8\% \pm 0.56\%$). The observation that regular senses might be overweighted by GDEX is shared e.g. by Cook et al. (2014: 320) who claim that “example-finding software does not yet routinely achieve the contextual diversity that characterizes example-sets selected by skilled lexicographers.”

Although our MEC improves on the Lesk algorithm it still does not improve to the base-line of always taking the main sense, which in the case of our dictionary consists of the 1st sense. The lines of improvement concern two areas: we plan to enrich the knowledge base with paradigmatic information from the German WordNet (GermaNet, Kunze & Lemnitzer, 2002). Furthermore, we can expect the results of our method to improve with the amount of available example sentences in the dictionary senses. Indeed, example sentences are underrepresented in the WDG as this dictionary was compiled before the era of electronic corpora. Therefore, we plan to repeat our experiments on the basis of the Duden dictionary (Duden-GWDS 1999). Duden has significantly more corpus examples. In the coming months, the Maximum Entropy Classifier will be integrated as a web service in the infrastructure of the Dictionary Writing System of the DWDS project.

7. Acknowledgements

This research has been carried out in the context of the BMBF-funded project KobRA (Korpusbasierte Recherche und Analyse mit Hilfe von Data-Mining, grant ID 01UG1245).

8. References

- Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing* 16, 5 (September 1995), pp. 1190-1208.
- Cook, P., Rundell, M., Lau, J. H. & Baldwin, T. (2014). Applying a Word-sense

- Induction System to the Automatic Extraction of Diverse Dictionary Examples. In: *Proceedings EURALEX*, Bolzano, Italy.
- Didakowski, J. (2008a). 'Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers'. In A. Storrer et al. (eds.), *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing. KONVENS 2008*. Berlin: Mouton de Gruyter, pp. 65-78.
- Didakowski, J., Geyken, A. & Lemnitzer, L. (2012). Automatic example sentence extraction for a contemporary German dictionary. In: *Proceedings EURALEX*, Oslo.
- Didakowski, J. & Geyken, A. (2012). From DWDS corpora to a German Word Profile – methodological problems and solutions. In: *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“*. In *Online publizierte Arbeiten zur Linguistik 2/2012 (OPAL)*, Mannheim: Institut für deutsche Sprache, pp. 43-52.
- Duden-GWDS (1999). *Das große Wörterbuch der deutschen Sprache*. 10 volumes. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.
- Escudero, G., Màrquez, L. & Rigau, G. (2000), 'Boosting Applied to Word Sense Disambiguation', In *Proceedings of the 12th European Conference on Machine Learning, ECML*. Barcelona, Catalonia. 2000.
- Geyken, A. (2014). Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. In U. Heid, S. Schierholz, W. Schweickard, H. E. Wiegand, R. Gouws, & W. Wolski (eds.). *Lexicographica*. Berlin / Boston: de Gruyter, S. pp. 77-112.
- Keok, Y.-L. & Ng, H.-T. (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02)*, Vol. 10. Association for Computational Linguistics, Stroudsburg.
- Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In: G. Williams & S. Vessier (eds.). *Proceedings of the XI. Euralex Conference*. Lorient: Université de Bretagne, pp. 105–116.
- Kilgarriff, A., Husák, M, McAdam, K., Rundell, M & Rychlý, P. (2008): GDEX - Automatically Finding Good Dictionary Examples in a Corpus. In: *Proceedings of the XIII EURALEX International Conference*. Barcelona: Universitat Pompeu Fabra, pp. 425-433.
- Kosem, I., Husák, M., & McCarthy, D. (2011). 'GDEX for Slovene'. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 151-159.
- Kunze, C. & Lemnitzer, L. (2002). GermaNet - representation, visualization, application. In: *Proceedings of LREC 2002, main conference, Vol. V.*, pp. 1485-1491.
- Lau, J. H., Cook, P., McCarthy, D., Gella, S., & Baldwin, T. (2014). Learning Word

- Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA.
- Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*. New York, NY, USA. ACM, pp. 24-26.
- Mooney, R. J. (1996). 'Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning'. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-96)*, Philadelphia, PA, pp. 82-91.
- Navigli, R. (2009), Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2), ACM Press, pp. 1-69.
- Navigli, R. & Velardi, P. (2005), 'Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation', *IEEE Transaction on Pattern Analysis and Machine Intelligence* 27, pp. 671-674.
- Nigam, K, Lafferty, J & McCallum, J. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*.
- Pedersen, T. (2007). 'Learning Probabilistic Models of Word Sense Disambiguation', *CoRR* abs/0707.3972 .
- Rychly, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák, (eds.) *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*. Brno: Masaryk University, pp. 6-9.
- WDG [1961-1977]: Wörterbuch der deutschen Gegenwartssprache in 6 volumes (4,800 pages). Akademie-Verlag: Berlin
- Vasilescu, F.; Langlais, P. & Lapalme, G. (2004). Evaluating Variants of the Lesk Approach for Disambiguating Words. In 'LREC' , European Language Resources Association.
- Volodina, E., Johansson, R. & Johansson-Kokkinakis, S. (2012). Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation. *Workshop on NLP in Computer-Assisted Language Learning. Proceedings of the SLTC 2012 workshop on NLP for CALL*. Linköping Electronic Conference Proceedings 80, pp. 59–70.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Multilingual lexicography for adult immigrant groups: bringing strange bedfellows together

Anna Vacalopoulou, Eleni Efthimiou

Institute for Language and Speech Processing, R.C. “Athena”,
Artemidos 6 & Epidavrou, Maroussi, Greece
E-mail: avacalop@ilsp.gr, eleni_e@ilsp.gr

Abstract

This paper presents a multilingual lexicographic project – expected to be completed by the end of 2015 – which focuses on the development of a set of corpus-based dictionaries for users not previously targeted; namely, adult immigrants in Greece trying to cope with a new reality. The project caters for languages that as of yet remain disjoint and also encompasses a variety of disconnected corpora, relevant to communicative situations with which the target group is most likely to cope.

The ultimate goal of this project is to reduce the linguistic gap between specific disconnected languages and styles as well as set the ground for the development of further relevant electronic language resources and reference works. This endeavour is currently at its final stage, namely the translation of the Greek content into the nine target languages: Albanian, Arabic, Bulgarian, Chinese, English, Polish, Romanian, Russian, and Serbian. This process will result in the compilation of nine bilingual dictionaries – from Greek into each of the aforementioned languages – with more than 15,000 single- and multi-word entries.

Keywords: multilingual lexicography; corpus-based lexicography; lexicography for disjoint languages and disconnected corpora

1. Introduction

This paper describes a multilingual set of dictionaries, which connects language pairs that as of yet remain unconnected, and outlines the approach that was adopted towards its creation. The significance of the user perspective in lexicography has been established and revisited in the bibliography for decades resulting in the continuous creation of significant works in the field (indicative works include Hartmann, R.R.K., 1979; Dolezal, 1999; Tarp, 2008). In this project, the lexicographic team was presented with a double challenge: not only did they have to identify and analyse user requirements, but they had to do so with no prior linguistic, much less lexicographic, work on which they could rely. After explaining the methodology used by the research team to pinpoint user profiles and connect them to specific needs, the paper goes on to describe the lexicographic process itself, in terms of lemma selection and disambiguation, example selection, categorisation of senses into semantic domains and the inclusion of extra information for each dictionary entry. At the end of the paper, the results of this project are summarised, along with some thoughts concerning their exploitation in future work.

2. Methodology of user group identification and analysis

When designing dictionaries, in terms of language coverage, entry selection and presentation mode, the lexicographic team concentrated on the user perspective in attempting to identify the users' reference needs; their proficiency level and background knowledge; their reference skills and strategies; as well as the effectiveness of dictionary use training (Varantola, 2002). Consequently, a needs analysis had to be conducted in order to primarily identify the user group profile(s) and respective needs.

The chief difficulty in conducting such an investigation was the team's inability to follow the methodology set by mainstream lexicographic research (Atkins, 1998). At those early stages of the dictionary-making process, it was not easy to locate the intended users in the first place, much less ask them to participate in any type of survey, since the target group's main concern was to struggle for a living in a new and unfamiliar reality. Additionally, as already mentioned, the specific user group had never previously been targeted, leaving the research team with a substantial gap in the bibliography. Thus, the team decided to postpone actual contact with the target group until a draft of the dictionaries became available online. Members of the target group would then be able to pilot the dictionaries and give valuable feedback while actually using it. This approach follows the so-called "simultaneous feedback" from target users to dictionary compilers (de Schryver et al., 2000). In order to avoid receiving this valuable user feedback too late in the process, which would at best make it useful for implementation in a revised edition of the dictionaries, it was decided to identify prospect user requirements and preferences by piloting an early draft of the dictionaries and receiving feedback through questionnaires. This process is expected to start immediately after the dictionaries are published online, so that compilers can test their hypotheses and be able to make any adjustments or improvements where needed with regards to this feedback.

In the meantime, compilers collected all available data which would enable them to initialise the compilation process; namely official, general-purpose statistical data (Vacalopoulou et al., 2011). The fact is that relevant available data describing the characteristics of immigrants in Greece are very scarce. With the exception of a small number of quantitative and qualitative surveys on immigration (Baldwin-Edwards, 2004; 2008), the only sources available at the time of research into this project were the 2001 census survey data and official data acquired from eurostat (<http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat>). A study of these sources led to the conclusion that the primary immigrant nationalities in Greece were Albanian, Bulgarian, Georgian, Romanian, Russian, Ukrainian, Polish, Pakistani, and Egyptian (in order of multitude). In terms of age, the majority of the immigrant population belonged to the 15–64 years old age group. Another distinct characteristic of the target group was that the main reason for residence permit award (68%) was dependent employment, followed by family reunification and self-employment (about 12% each) and a considerably smaller number of immigrants who moved to Greece in

order to study. The target group profile was completed with the identification of the place that the majority of immigrants occupied in the Greek labour market, revealing building construction, agriculture, industry and tourism as the main activities of males and housekeeping, cleaning, agriculture and tourism as the main activities of females.

For the purposes of dictionary compilation, the target group's level of education and language literacy were also considered. According to the aforementioned sources, the educational level of the vast majority of immigrants in Greece ranged from medium to low. In particular, the statistics suggested the existence of three main categories in terms of education and literacy: (a) people who had completed secondary education before migrating; (b) people who had only attended primary school, and (c) people who were considered illiterate. The first two categories comprised mainly immigrants of European origin (from Albania, Bulgaria, Poland, and Serbia) whereas the third category was populated with immigrants from African and Asian countries. Lastly, the sources revealed that, as expected, the vast majority of all these groups had little or no prior knowledge of Greek. Combining the above data, the research team decided that it was safe to assume that the user group described above had little, if any, experience in dictionary use.

Based on these data, the research team concluded that as diverse as the intended target group was in terms of nationality, level of literacy and language proficiency in Greek, the tendency was towards a lower level. Based on such a user profile, the team pinpointed user needs and requirements as defined by the users' struggle to be included in the Greek society. The dictionaries would thus have to be designed in view of providing basic linguistic knowledge, taking into account the following linguistic and non-linguistic factors: the user group's communicative needs in official settings (e.g. in dealing with the Greek authorities or applying for a green card) and social settings; needs to address everyday issues (e.g. travel and transportation); language learning in formal or informal settings; and familiarization with the general cultural and social context.

3. Lemma Selection

As aforementioned, the dictionaries cover the most common range of foreign languages used and/or understood by the majority of the immigrant community in Greece. Thus, nine bilingual dictionaries for users not previously targeted are being created; specifically Greek–Albanian (EL–AL), Greek–Arabic (EL–AR), Greek–Bulgarian (EL–BG), Greek–Chinese (EL–CH), Greek–English (EL–EN), Greek–Polish (EL–PL), Greek–Romanian (EL–RO), Greek–Russian (EL–RU), and Greek–Serbian (EL–SR). English was selected as one of the target languages to compensate for a lack of languages of less represented immigrant groups in Greece while being an official or widely used language in the countries of several of the respective nationalities (e.g. Pakistan, Bangladesh, the Philippines). At the same time, the Greek–English language pair was included for reasons of lexicographic convenience, as it is generally recognised

as an “international language of communication, a global language [...], which enables speakers of any language to have a common ground with each other [...]” (Kernerman, 2004). Apart from being convenient for users, English also proved a useful means for translators to double-check the rest of the language pairs (i.e. from Greek) which are considerably less frequent.

Each of these bilingual dictionaries consists of more than 15,000 entries covering mainly the basic vocabulary of Greek. Even though a formal complete list of basic Greek vocabulary is still missing from the literature, the basic vocabulary is conceived as one which comprises not only the most frequent items but also less frequent words and phrases that are relative to everyday activities. Thus, a common definition of such a list would be “the set of lexical items in a language that are most resistant to replacement, referring to the most common and universal elements of human experience, such as parts of the body [...], universal features of the environment [...], common activities [...], and the lowest numerals.” (Dictionary.com). For the purposes of this project, the compiling team considered a combination of items which occur with significant frequency in general language corpora, of items representing basic meanings as described in the definition above as well as of items which help interpret the rest of the vocabulary. This last set of items is known in lexicographic practice as a ‘defining vocabulary’ (Atkins et al., 2008).

Apart from the basic vocabulary, another major category of entries is the one often occurring in official, administrative or other documents which the target group is likely to encounter during their stay in the country, as, for instance, when applying for a residence or work permit. To this end, a selection of more technical terms were included as well, pertaining to subject fields that are of utmost interest to the target group. Although technical jargon is generally expected to be part of general language dictionaries (Béjoint, 1988), its scope was limited to those terms that are likely to appear in administrative or other official documents, which were considered more relevant to the user group.

Based on the assumption that the target group would lack basic encyclopaedic information about Greece, the dictionaries also contain proper nouns. These consist of names of geographical entities (i.e. cities, islands, regions etc.), official bodies (i.e. ministries and other state organisations) and geopolitical entities (*Ευρωπαϊκή Ένωση* = *European Union*). Acronyms representing official organisations and geopolitical entities are also included in the entry list.

The dictionaries contain both single- and multi-word entries. Apart from the types of multi-word entities that would usually have entry status in bilingual dictionaries (*ασφάλεια ζωής* = *life insurance*, *χαρτί υγείας* = *toilet paper*), it was decided that the dictionaries would include more types of multi-word entries so as to extend the linguistic coverage (Granger et al., 2012). Thus, entries include several set phrases, such as everyday expressions that would normally appear in tourist phrase books,

collocations and idioms (*χρόνια πολλά* = *happy birthday*, *παίρνω τηλέφωνο* = *make a phone call*, *παίρνω από λόγια* = *listen to reason*). The value of this decision in practice can be understood if one considers that only a few, if any, of these entities could be inferred from word-to-word translation into Greek, as it is often the case (Svensén, 2009). The argument can be further strengthened if one considers the number of disjoint languages and styles this set of dictionary brings together.

Alternative forms of the same lexical item are separate entries interlinked with each other. For instance, *Προαστιακός Σιδηρόδρομος* (*Suburban Railway*) and *Προαστιακός* (*Suburban*) are two separate dictionary entries linking to each other. Similarly, *αντισυλληπτικό χάπι* (*contraceptive pill*) and *αντισυλληπτικό* (*contraceptive*) are treated in the same way. The ‘complete’ form of such lemmas is given main entry status and contains the rest of the information, whereas the secondary entry/entries are cross-referenced to the main entry. In general, when lemmas linked by a cross-reference belong to different registers, the most formal type is given main entry status, as this is the form more likely to occur in official documents. In the case of acronyms, the full name of the entity is given main entry status (*Οργανισμός Ηνωμένων Εθνών* = *United Nations*), with a cross-reference under its acronym (*ΟΗΕ* = *UN*). For reasons of easy reference, acronyms are normalised and thus spelled without full stops between letters.

The process of dictionary compilation was corpus-based; this refers to headword selection, sense disambiguation and extraction of collocations and usage examples. Dictionary entries were semi-automatically selected from a variety of sources, namely (a) a large, POS-tagged and lemmatised general-language corpus of modern Greek (Hatzigeorgiu et al., 2000), known as the Hellenic National Corpus (<http://hnc.ilsp.gr/>), (b) a specialised Greek corpus collected within the framework of the current project, that adheres to pre-defined domains (public administration, culture, education, health, travel, and welfare), and (c) already existing dictionaries, glossaries and travel phrase books, customised to better suit user requirements (communicative situations and relevant vocabulary, etc.). Such resources were previously developed by ILSP for the purpose of other projects and include either published¹ or non-published works.

Furthermore, according to standard practice, the dictionaries include every word in the examples as an entry itself for easy reference; in other words, there is no lexical item in the examples (excluding certain proper names) which does not appear in the dictionaries itself as a separate entry. This led to adding a considerable number of entries to the dictionaries and maintaining a better balance, in terms of content, between everyday vocabulary and the administrative jargon of the public service, thus making sample entries of the two corpora less disconnected. The ultimate goal of this

¹ Two examples of published works are the *Electronic Greek–Turkish Dictionary for Young Learners*, Athens 2004 and *XENION Lexicon*, Athens 2005.

merge was to reconcile “the technical meaning and the everyday meaning [...] and making a concise meaningful representation of the whole to the public” (Hanks, 2010).

4. Lemma Disambiguation

As in most dictionaries of Greek, the main criterion for distinguishing between lemmas is morphology. Therefore, *Δεκέμβριος* and *Δεκέμβρης* (= *December*) are separate entries, as are *φέτος* and *εφέτος* (= *this year*), *κιάλας* and *κιάλα* (= *already*), etc.

The second criterion used for distinguishing between lemmas is part of speech. Therefore, homographs belonging to different parts of speech (*ωραίος, ωραία, ωραίο* = *nice*, *ωραία* = *nicely*) form separate entries. In an attempt to tackle language learning difficulties arising from the fact that “Greek is a highly inflectional language and marks verb suffixes for person and number” (Holton et al., 1997), the past participle of a verb is treated as an adjective. Therefore, past participles form separate entries (*πλυμένος, πλυμένη, πλυμένο* = *washed*, p.p. of the verb *πλένω* = *wash*; *κλειδωμένος, κλειδωμένη, κλειδωμένο* = *locked*, p.p. of the verb *κλειδώνω* = *lock*). Following similar simplification criteria, other types of word derivatives are separate entries in these dictionaries. Therefore, adverbs (*καλά* = *well*; *γρήγορα* = *quickly*) are different entries from the respective adjectives (*καλός, καλή, καλό* = *good*; *γρήγορος, γρήγορη, γρήγορο* = *quick*).

As is standard practice in regular monolingual dictionaries, every single-word entry appears in the base form. As a result, verbs appear in the first person singular present in the active voice; nouns appear in the singular nominative; adjectives and past participles appear in the nominative positive (in this case, in the masculine, feminine and neutral); and adverbs appear in the positive. Exceptions to the above arise when what is considered as the base form is either ungrammatical or particularly infrequent in Greek (*πρέπει* = *it must*, the third instead of the first person, *γυαλιά ηλίου* = *sunglasses*, the plural instead of the singular, *αρρωσταίνω* = *fall ill* instead of *αρρωσταίω* = *cause somebody to fall ill*).

Following the simplification criterion further on, nouns referring to professions or other human activities form two different entries (i.e. masculine and feminine) as, in most cases, their morphology in Greek differs (*αθλητής* and *αθλήτρια* = *athlete*, *καταστηματάρχης* and *καταστηματάρχισσα* = *shop-owner*). Rare exceptions to the above rule include nouns with identical masculine and feminine forms (*ηθοποιός* = *actor* and *actress*; *πολιτικός* = *male or female politician*).

Finally, and along the same lines, the comparative and superlative of a few highly frequent adjectives and adverbs are also given separate entry status. Thus, *καλύτερος, καλύτερη, καλύτερο* = *better* as well as *χειρότερος, χειρότερη, χειρότερο* = *more* appear separately from *καλός* = *good* and *κακός* = *bad*, respectively.

5. Examples of Use as Bearers and Differentiators of Meaning

As aforementioned, this resource does not only bring together disjoint languages but also highly disconnected corpora. In order to meet this double challenge, it was decided that a certain set of rules were to be followed. First, as the dictionaries are mainly targeted towards starter learners of Greek who are in need of speedy learning, it was decided that only basic meanings would be included in them. Meanings are implicitly presented through one or more examples of usage, which, along with their translations, bear the informative load. This makes examples of usage a core element of the dictionaries, playing the additional role of describing each meaning, due to lack of definition. This led to additional difficulty in selecting the right example(s) for each meaning. For instance, a successful example of the verb *αγωνίζομαι* = *struggle* would be *Αγωνίστηκε πολύ για να καταφέρει αυτό που ήθελε* = *She struggled a lot to get what she wanted*, as not only does it include the word in context but it also helps the user to capture its meaning. In general, great care was taken to select examples that would comply with as many items as possible on a list presented in Prinsloo (2013), according to which ‘[g]ood examples disambiguate senses; distinguish one meaning from another; [...] show or indicate the selectional range; place the word in context; specify the semantic range; indicate the collocational behaviour [...]; illustrate the grammatical patterns; specify the word order; give pragmatic uses; note stylistic features; indicate appropriate registers [...]’.

Second, dictionary examples were carefully selected so as to reflect not only different meanings but also the most basic forms of usage, grammar and collocation. Therefore, for instance, the active and passive forms of verbs are presented by separate examples whenever voice differentiates meaning as well; the same process is followed for verbs used with different prepositions, items combined with different collocates etc.

Furthermore, as the lexicographic team’s intent was to include as much information as possible expressed in the most user-friendly way possible, there was a conscious attempt to avoid boring the user. Therefore, while a large number of the examples were extracted from the Hellenic National Corpus, they were usually shortened and/or simplified in order to suit the target group level as is common lexicographic practice (Kilgarriff, 2013). Therefore, examples on the whole are short and contain no excess information. They usually comprise one sentence, although some dialogue is, at times, included in the case of everyday phrases, such as greetings or asking for information. In addition to accelerating the learning process, this brevity principle also simplifies the task of translating the Greek content into nine languages.

Finally, bearing in mind the great variety of target group backgrounds, additional attention was given to political correctness. Dictionary examples are void of any social, political, racial, national, religious or gender bias.

In their attempt to comply with the aforementioned criteria, the lexicographic team

decided to follow the common practice of modifying “corpus sentences which are promising but in some way flawed” when applicable (Kilgarriff, 2013). Such ‘flaws’ included – among others – verbosity, political incorrectness and inclusion of lexical items which were not part of the entry catalogue.

6. Semantic Domains

For easier reference, different meanings of each entry are classified into broad domains reflecting certain communicative contexts. As noted above, this is a highly particular target group in terms of dictionary use, whose communicative needs could be viewed as a combination of the needs of a first-time tourist who is expected to be an active citizen at the same time. Some examples of such needs would be the need to use public transport, to go shopping, to look for a flat, or to register a child in school. As a result, the domains have to be detailed enough to cater for as many different aspects as possible and inclusive enough to facilitate usability. Another reason for classifying dictionary entries into domains was that, according to studies, users of bilingual dictionaries rarely go through the list of senses of each entry to find the appropriate one, as there is a tendency to select the first meaning (Lew, 2004). The team’s assumption was that users would be in a better position to locate the appropriate meaning if senses were tagged for semantic domain. In other words, this classification will hopefully help users to unambiguously retrieve the appropriate information. This assumption, of course, will have to be tested in the piloting stage.

Furthermore, users can simultaneously view different senses of each lemma belonging to different domains, thus being able to compare and contrast among them and gain a better understanding of each word. The communicative domains that were used in the dictionaries are illustrated in Table 1 below, followed by a short description and some indicative examples of entries.

Domain	Description	Examples
<ul style="list-style-type: none"> • Culture, recreation and the media 	<ul style="list-style-type: none"> • vocabulary from the arts; hobbies & spare time; TV & other media 	<ul style="list-style-type: none"> • <i>μουσική</i> = <i>music</i>; <i>μπαλέτο</i> = <i>ballet</i>; <i>μικρές αγγελίες</i> = <i>classified ads</i>
<ul style="list-style-type: none"> • Education 	<ul style="list-style-type: none"> • all aspects 	<ul style="list-style-type: none"> • <i>μάθημα</i> = <i>lesson</i>; <i>νηπιαγωγείο</i> = <i>nursery school</i>
<ul style="list-style-type: none"> • Environment 	<ul style="list-style-type: none"> • flora & fauna; weather; ecology etc. 	<ul style="list-style-type: none"> • <i>λίμνη</i> = <i>lake</i>; <i>μέλισσα</i> = <i>bee</i>; <i>μόλυνση</i> = <i>pollution</i>
<ul style="list-style-type: none"> • Finance 	<ul style="list-style-type: none"> • money & the economy; taxation; bank 	<ul style="list-style-type: none"> • <i>λογαριασμός</i> = <i>bill</i>; <i>μετρητά</i> = <i>cash</i>; <i>ναύλα</i> = <i>fare</i>

	transactions etc.	
• Geography	• countries; nationalities; languages; Greek cities & areas	• <i>Μεσόγειος Θάλασσα = Mediterranean Sea; ήπειρος = continent</i>
• Housing & Accommodation	• parts of the house; furniture & appliances; hotels etc.	• <i>κουζίνα = kitchen; κουζίνα = cooker</i>
• Labour & Insurance	• all aspects	• <i>ανεργία = unemployment; μισθοδοσία = payroll</i>
• Law, Justice & Public Safety	• all aspects	• <i>δικηγόρος = lawyer; παράνομος = illegal</i>
• Physical condition & Health	• parts of the body; diseases; doctors etc	• <i>μελανιά = bruise; μικρόβιο = virus</i>
• Public Administration	• all aspects	• <i>ληξιαρχείο = registry office; πολίτης = citizen</i>
• Greek Holidays & Traditions	• the most common ones	• <i>Πάσχα = Easter; κηδεία = funeral</i>
• Relations & Family	• all aspects	• <i>μητέρα = mother; παντρεμένος = married</i>
• Science & Technology	• widely used terms	• <i>μηχανικός = mechanic; κινητό τηλέφωνο = mobile phone</i>
• Transport & Travel	• urban transport; travelling	• <i>λιμάνι = port; μετρό = metro</i>

Table 1: Dictionary domains

As expected, the most populated domain is general vocabulary. For mainly educational reasons, part of this was further subcategorized into easily grasped vocabulary groups including: numbers, clothing and accessories, food and cooking, time, space, colours, units of measurement, and everyday interaction (informal words and expressions).

7. Additional Entry Information

Excluding entries which are cross-references, each dictionary entry is accompanied by an audio file to exemplify pronunciation, hyphenation, alternative entry types, basic

grammatical information (i.e. the masculine, feminine and neutral type for all adjectives and past participles) and examples of usage. Each example is translated into nine languages, with the entry lemma highlighted in the example.

Concerning pronunciation, audio files also accompany all dictionary examples in Greek and their Bulgarian translations using a synthetic voice. These are expected to support users with vision or literacy problems on the one hand and also help the vast majority of users who are unfamiliar with the Greek script on the other.

Finally, all multi-word entries are linked with each of their components (excluding functional words) through cross references. Apart from facilitating easy reference this feature also bears a pedagogical added value, given that most of the words which form these phrases are inflected types of other entries. It, therefore, becomes easier for users to link each inflected type to the base form of the entry.

8. Results and Future Work

We presented lexicographic work targeted at the development of a set of nine online bilingual dictionaries for immigrants in Greece. This project (which is currently at the translation stage) is expected to be finished by the end of 2015 and its results will be freely available online.

Concerning the exploitation of the results of the project, efforts are being made to come up with as many user friendly ways as possible in which different users will be able to make different searches. Various ways of presenting the results of those searches are also explored. The lexicographic team feels that this is of the essence, as the immigration landscape in Greece keeps changing rapidly largely for reasons relating to the country's financial crisis (Triandafyllidou, 2014). Therefore, if such a linguistic resource aspires to remain useful, exploitable and relevant, it must be flexible enough to cater for as wide an audience as possible.

Lastly, the results of this project will form a valuable multilingual resource in themselves, as this set of bilingual dictionaries will provide a common core lexicon for 10 disjoint languages. Another step to be taken will be the exploitation of these unique dictionaries as corpora for the extraction of more reference works and/or the support of NLP tools which will cater for the specific target group.

9. Acknowledgements

Anna Vacalopoulou & Eleni Efthimiou were supported by the POLYTROPON (KRIPIS-GSRT, MIS: 448306) project.

10. References

- Atkins, S.B.T. (1998). *Using Dictionaries: Studies of Dictionary use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.
- Atkins, S. & Rundell, M. (2008). *Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press, pp. 449-450.
- Baldwin-Edwards, M. (2005). *Statistical Data on Immigrants in Greece*. Athens: Mediterranean Migration Observatory.
- Baldwin-Edwards, M. & Kolios, N. (2008). *Immigrants in Greece: Characteristics and Issues of regional distribution*. Athens, ISTANCE.
- Béjoint, H. (1988). Scientific and technical words in general dictionaries. *International Journal of Lexicography* 1(4), pp. 354-368.
- de Schryver, G.M. & Prinsloo, D.J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, pp. 197-209.
- Dictionary.com: <http://dictionary.reference.com/> (5 July 2015)
- Dolezal, F.T. & McCreary, D. (1999). *Pedagogical Lexicography Today. A Critical Bibliography on Learners' Dictionaries with Special Emphasis on Language Learners and Dictionary Users*. Lexicographica. Series Maior, 96. Tübingen, Max Niemeyer Verlag.
- eurostat: <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat> (5 July 2015)
- Granger, S. & Lefer, M.A. (2012). Towards more and better phrasal entries in bilingual dictionaries. In R. Vatvedt Fjeld & J.M. Torjusen (eds.) *Proceedings of the Fifteenth EURALEX International Congress, EURALEX 2012*. Oslo: UiO, pp. 682-692.
- Hanks, P. (2010). Terminology, Phraseology, and Lexicography. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the Fourteenth EURALEX International Congress, EURALEX 2010*. Ljouwert: Fryske Akademy, pp. 1299-1308.
- Hartmann, R.R.K. (ed.) (1979). *Dictionaries and Their Users* [BAAL Seminar, Exeter 1978] (Exeter Linguistic Studies 4). Exeter: University of Exeter Press.
- Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Spiliotopoulou, A., Vacalopoulou, A. et al. (2000). Design and implementation of the online ILSP Greek Corpus. In *Language Resources and Evaluation Conference, LREC 2000*. Athens, Greece.
- Holton, D., Mackridge, P. & Philippaki-Warbuton, I. (1997). *Greek: A Comprehensive Grammar of the Modern Language*. Routledge, London.
- Kernerman, I.J. (2004). Dictionary Visions, Research and Practice. In H. Gottlieb & J.E. Mogensen (eds.) *Selected Papers from the Twelfth International Symposium on Lexicography*, Copenhagen.
- Kilgarriff, A. (2013). Using corpora [and the web] as data sources for dictionaries. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London,

- Bloomsbury Publishing, pp. 77-96.
- Lew, R. (2004). *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-Bilingual Dictionaries by Polish Learners of English*. Poznań: Motivex.
- Prinsloo, D.J. (2013). New developments in the selection of examples. In R. Gouws (ed.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, Walter de Gruyter GmbH, Berlin/Boston.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Triandafyllidou, A. (2014). Migration in Greece: Recent Developments in 2014. Report prepared for the OECD Network of International Migration Experts, *Global Governance Programme*, Paris, 6-8 November 2014.
- Vacalopoulou, A., Giouli, V., Giagkou, M. & Efthimiou, E. (2011). Online Dictionaries for immigrants in Greece: Overcoming the Communication Barriers. In I. Kosem & K. Kosem (eds.) *Proceedings of the Second Conference "Electronic Lexicography in the 21st Century: New Applications for New users"*, eLEX2011, Bled, Slovenia, pp. 274-279.
- Varantola, K. (2002). Use and Usability of Dictionaries: Common Sense and Context Sensibility? In M.H. Correard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins*. Grenoble, France: EURALEX, 2002.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Overwriting knowledge: analyzing the dynamics of Wikipedia articles

Nathalie Mederake

Deutsches Wörterbuch, Göttingen Academy of Science and Humanities,
Papendiek 14, 37073 Göttingen, Germany
E-mail: nmedera@gwdg.de

Abstract

The popularity of open collaborative content generation such as Wikipedia, while expanding the amount of available information, also poses particular challenges as its user-generated content changes constantly. This paper proposes to study the development of Wikipedia entries and to systematically measure and evaluate this type of user-generated dynamics. The applied approach is able to identify phases of the constant process of content generation. It takes into account the interrelations between dynamics of user contributions and article-related real-world events. A data set spanning article versions and associated discussion pages over two years was analysed. This allowed identifying trigger pulses that drive the articles' development both on qualitative and quantitative levels. For effective planning of online dictionaries that stress the involvement of users or intend to add collaborative components, it is crucial to consider such findings. The approach might also be transferrable to lexicography in terms of analysing the revisions of a collaborative dictionary entry as a signal indicative of lexical change. For that reason, I conclude with a discussion of the results and their relevance for expert lexicographic products.

Keywords: wiki; collaborative lexicography; content generation process

1. Introduction

With the rise of the Web 2.0, users can actively participate in the compilation of online reference works such as dictionaries and encyclopaedias. However, these works can be subdivided into different partial areas of lexicography (each with its own characteristic forms), as they are displayed by Wiegand et al. (2010: 125). Lexicographic products can investigate the respective language or their subjects “when the perspective of the comments is such that one can obtain answers about corresponding non-language objects” (ibid.). According to the distinction made by Wiegand et al., the largest available and fastest growing collaboratively constructed encyclopaedia project Wikipedia is to be defined as a non-scientific lexicographical reference work, predominantly fulfilling the mentioned purposes related to subjects.

Compared to editorial reference works, the collaborative lexicographic process shows significant differences in the steps and phases towards compilation. One of the peculiarities of a collaborative lexicographic process is the iterative writing process that yields multiple revisions of an entry (cf. Meyer, 2013: 53). These revisions can lead to continuous changes in the lexicographic product, for example, when a new article constituent is introduced. Hence, collaborative projects are revision-driven and

not directed to a final closing phase as might be the case with editorial reference works. Users write and edit articles in a collaborative manner and the outcome is published immediately on the web; also, feedback can be instantly given. One might consider it either a problem or actually a benefit that web contents are subject to constant change and that dictionaries or encyclopaedias thus will not remain the 'final products' they used to be for a long time. Of course, the traditional dictionaries or encyclopaedias are also not entirely "final" - there is a discrete number of successive editions representing the major development over longer time periods. In contrast, the fact that wiki entries are updated in a continuous manner, as often as needed or regarded useful, in principal by anyone who wishes to make a change, has made them an integral part of everyday life.

It is not surprising that methods of how to systematically measure or evaluate user-generated contents within the wiki-environment are developing. They are concerned e.g. with the evolution of discussion (Kaltenbrunner & Laniado, 2012), the understanding of the writing process (Kallass, 2015), and the investigation of look-up frequencies (Müller-Spitzer et al., 2015). The research of Stvilia et al. (2005a, b; 2008) and Stvilia & Gasser (2008) discusses the aspects and dynamics of information quality in Wikipedia and gives useful pointers on how the quality assessment and improvement process operates. Their model is concerned with changes in the field of information quality and can actually be used for reasoning about similar dynamics in different settings. In their study, they used the discussion page or talk page and other process-oriented pages within Wikipedia to determine indicators for information quality. Despite these advances, web dynamics continue to be an ongoing challenge for lexicographers (and linguists in general). In addition, lay users are still mostly unaware of the developments that happen in the background of collaborative projects such as wikis and of how contents are changed in the course of a revision.

In fact, since every user benefits from up-to-date content and is given the opportunity to reflect on how content has developed in the page history, it is important to set the starting point there: What changes have been made, which links have been replaced or which illustrations have been chosen at what time? In addition, less compressed forms of presentation, as available in the wiki-interface, result in longer, sometimes less structured articles¹. But what is important or relevant for both the users and the producers in this reference work; what do they deal with, especially in a more narrative structure? I believe that answering these questions will also lead to fruitful findings for institutional or professional lexicography. The research of Müller-Spitzer et al. (2015) for example uses quantitative evaluations of log files to explore general patterns of look-up behaviour in Wikipedia's sibling, German Wiktionary, to

¹ As a side note: The absence of space restrictions in the digital environment altogether will, in the long run, lead to longer dictionary articles, or narrative article structures on word-related information in institutional lexicography as well, like in the examples of so-called Wortgruppenartikel (= entries referring to word group) in *lexiko* or Macmillan's *BuzzWord*.

understand the needs of users and the information they would like to have. Accordingly, I believe that we can only use search results derived from wikis for our own lexicographic products if we fully understand how the collaborative system works and what is important for the active user. I will therefore present a method of how to systematically study the development of Wikipedia entries. The analysis takes into consideration findings from the history page related to the respective article as well as the discussion pages, together with corresponding real-world events. Besides, some light will be shed on the following questions: what kind of information seems to be important for user-generated content in an online encyclopaedia and what are the underlying strategies of revision? I will conclude with findings on regularities in the dynamics induced by the collaborative environment and a discussion of the results within the field of lexicography.

2. Model and distinctive Features

The concept of Wikipedia has been popular for a long time, as has collaborative online editing in general. These processes are being widely used even by information professionals (Lih, 2004; Emigh & Herring, 2005) – and they have also found their way into daily language lexicographic routine. In fact, there seems to be a fruitful coexistence between Wikipedia and more traditional language dictionaries: institutional dictionary projects such as *Algemeen Nederlands Woordenboek* also offer links to Wikipedia in their search results². Similarly, institutional language dictionaries are used as references in Wikipedia’s articles³. Taking the sister project Wiktionary into account, it becomes apparent that the German Wiktionary, for example, relies to a large extent on secondary sources such as Duden online, *Digitales Wörterbuch der deutschen Sprachen* or *Deutsches Wörterbuch von Jacob and Wilhelm Grimm* (cf. Meyer, 2013: 42). However, the variety within primary, secondary and tertiary sources, such as monographs, grammar etc. (cf. Wiegand, 2010: 133), tends to differ according to the specifications of each reference work and also depends on whether it is going to serve language or subject related lexicographic purposes. Likewise, it is argued that open-collaborative contributions (that by definition draw upon very diverse sources) have enormous potential in keeping the contents of a dictionary up to date and ensuring their high quality (cf. Abel & Meyer, 2013: 179), even if most of them are not constituted or controlled by a predefined group of experts. In fact, Wikipedia actually “gets better the more people use it, since more people can contribute more knowledge, or can correct details in existing knowledge for which they are experts” (Vossen & Hagemann, 2007: 47).

² E.g. <http://anw.inl.nl/article/peer>

³ Compare references to Oxford English Dictionary and *Griechisches Etymologisches Wörterbuch* in the German Wikipedia article ‘Birne’:
<https://de.wikipedia.org/wiki/Birnen#Quellen> (6/7/2015)

Therefore, a general model for the better understanding of the collaborative process will be presented. It refers to the Wikipedia system in particular and highlights its distinctive features. Following Bruns' (2008: 102) description of a wiki, “[w]ikis enable their users to create a network of knowledge that is structured ad hoc through multiple interlinkages between individual pieces of information in the knowledge base; they represent, in short, a rapidly changing microcosm of the structures of the wider Web beyond their own technological boundaries”. Based on this, circular movements in the contribution process and complex interactions of endogenous and exogenous factors can be specified (cf. Fig. 1). Such factors correspond to activity peaks that have been observed so far not only in Wikipedia (e.g. Kaltenbrunner & Laniado, 2012; Mayer, 2013: 123–143) but also in other social media platforms such as Youtube (Crane & Sornette, 2008) or Twitter (Lehmann et al., 2012).

One of the endogenous factors for a collaborative encyclopaedia is for example the software platform of Wikipedia, which is built upon a relational database with different search paths. The linking structure also allows for immediate cross-references – even to articles that do not yet exist. Additionally, wiki-based reference systems are usually neither based on fixed (lexicographic) instructions nor do they show a predefined microstructure. One of the main characteristics of wiki software is an extreme reduction of the costs of collaborative content creation, dissemination and upkeep. The structural openness obviously causes inconsistencies in the layout of the articles and their microstructure. But most importantly, users can and do directly modify contributions of other users. The process of production and using is ongoing and is never finished. In fact, the most important result of collaborative editing is a continuous process rather than a static product. This process can generate projects that are richer and more complex than those produced by individuals, which leads us to the most important exogenous factor: A wiki is nothing without its users.

Wikipedia still grows and develops its features, despite the known discrepancy in active and passive user behaviour, e.g. in German Wikipedia (cf. Busemann, 2013: 319). For example it has been shown (cf. Döring, 2010: 177) that passive usage (via page visits etc.) prompts further active participation. Additionally, search engine optimization has had a significant effect on the visibility (and in that, recognition) of web content. In this environment the concept of ‘prosumption’ (i.e. in the most general sense, the creation of products and services by the same people who will ultimately use them) seems to work better than an elaborate and refined product created by experts (such as expert lexicographers). The idea behind the prosumer commodity and thus that of user-generated content (Lew, 2014), and bottom-up-lexicography (Carr, 1997) is that the roles of producers and consumers blur and merge. It is also argued that criteria such as openness, sharing, peering and global outreach increase the value of prosumer participation. Facing the collaborative extension and editing of Wikipedia, Bruns coined the term ‘produsage’ to describe user-led content production within the Web 2.0 environment. He argues that “within the communities which engage in the collaborative creation and extension of information and knowledge [...] the role of

‘consumer’ and even that of ‘end user’ have long disappeared, and the distinctions between producers and users of content have faded into comparative insignificance” (Bruns, 2008: 2).

Therefore, boundaries become transient. The concept of article ownership does not apply as anyone can modify articles at any time. The collaborative process is intermittent and not systematic due to significant interactions. They are fostered on an object level, where article creation (and thus representation of knowledge) takes place, as well as on a meta-level, where the above mentioned concept of ‘produsage’ as well as events and developments over time affect every article. Such interactions also determine the dynamic character of content creation. Because of the ongoing “work in progress” situation the quality of every article can also only be expected to be fluid and transient. Here, the term ‘dynamic’ points to the fact that the articles’ contents and appearances change over time. But is there a pattern?

In their studies about dynamics in information quality, Stvilia et al. (2005b; 2008) and Stvilia & Gasser (2008) agree on the definition about information quality as being the assessment on information’s ‘fitness for use’ (cf. Juran, 1992; Wang & Strong, 1996) in a particular task system or activity system. Regarding information quality in Wikipedia, they observed a number of patterns in the development trajectories for featured articles that appeared to follow the life cycle of the underlying entities. However, besides the articles’ underlying entities or the context of its evaluation (e.g. degree of domain knowledge) and use (also in terms of sociotechnical structure) there is a significant link to the element I described as ‘produser’. In terms of quantification, this means: the number of edits an article may receive is affected by the attention drawn to the article’s entities. Ferron & Massa (2011a, b), Keegan et al. (2011) and Kallass (2015) have identified this kind of intensive participation in revisions and discussions on talk pages as event-related. Additionally, the analysis of Stvilia & Gasser (2008) showed that Wikipedia “would direct community resources to a particular article in anticipation of an event that could change the quality and/or criticality of the article” (ibid.).

This means that the triggering of an article’s development is caused by real-world changes related to its topic as well as by initiatives of the produser-element. Thus, “fitness for use” seems to resemble a negotiation process which is highly context sensitive: Coherence needs to be achieved in terms of the articles’ entities⁴ and the potential contribution of the produser-element to this topic – in short, coherence between the interactions of endogenous and exogenous factors.

⁴ Here, context sensitivity also relates to Wikipedia policies. E.g. in English Wikipedia the avoidance of *recentism*, that is editing an article without a long-term, historical view, and determining proper weight in depth of detail, quantity of text, prominence of placement, etc., belong to the content policies of Wikipedia:
https://en.wikipedia.org/wiki/Category:Wikipedia_content_selection

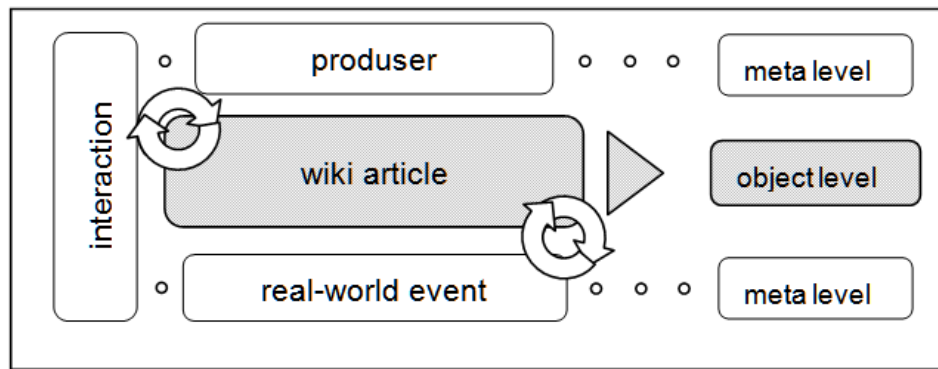


Figure 1: Activity model within a wiki (cf. Mederake, 2014: 239)

3. Data set and methodological approach

Wikis include mechanisms that allow us to follow visible changes made to pages over time, i.e. the display of the related data history⁵, as well as discussion pages or talk pages, which are tied to entries and where various content-related issues can be addressed. As these features are central to the Wikipedia quest in terms of information quality, I will make use of them to see what information is distributed and when.

Wikipedia articles describe or deal with different kinds of entities: people, places, events, concepts, or things. The data set for this study comprised the edit histories of two articles from the German Wikipedia: ‘Zitronenpresse’ (= lemon squeezer)⁶ and ‘Eurokrise’ (= European debt crisis)⁷. Describing 1) a very general object and 2) a current event, these articles are typical examples of article topics in the German Wikipedia; the article ‘lemon squeezer’ also was awarded the label ‘worth reading’ until it was highlighted as ‘excellent’ during the survey and can therefore be qualified as a high-quality article.⁸ Categories like ‘worth reading’ or ‘excellent’ denote article

⁵ In the edit history, meta-data elements can be found containing the following information: data and time, name or IP of the user, comment to clarify the edit purpose. Edit histories are also a source for meta-information about the article (age, time of update, number of times the article has been edited, information about editors and edit type). Such elements of the data history can provide valuable information about the social structure and dynamics of the articles’ content creation. <http://de.wikipedia.org/wiki/Hilfe:Versionen>

⁶ <http://de.wikipedia.org/wiki/Zitronenpresse>

⁷ <http://de.wikipedia.org/wiki/Eurokrise>

⁸ Articles are awarded featured article status after the community has achieved a consensus that the article meets the featured articles criteria (comparable to English Wikipedia; i.e. attributes as well-written, comprehensive, well-researched, neutral, stable, appropriate structure, consistent citation format and so forth). It can be judged that these are general quality dimensions based on respective cultural and social conventions, and characteristics specific to the encyclopedia article genre and the community of Wikipedia. Articles keep their featured article status, even if they get changed again, until they are demoted for lack of meeting the quality requirements. <http://de.wikipedia.org/wiki/Wikipedia:Bewertungen>

status in the German Wikipedia comparable to the ‘featured article’ status, which articles in the English Wikipedia can achieve (after a thorough review process). It should be noted that the objective of the featured article process is to encourage the writing process to evolve and improve, thus increasing quality within Wikipedia.

Over a period of more than two years, a data set of 20 article versions altogether was created, using monthly, bi-monthly or quarterly data points. The first version of every article topic marks the starting point of the survey. Additionally, I looked into the logs of the associated discussion pages or talk pages to allow a more in-depth content analysis of specific incidents within the articles’ development.

In order to observe which instances had been moved or added at what time during the articles’ development, findings in frame semantics (following Konerding, 1993) were applied in a coding procedure to develop a classification scheme. This scheme was then applied to all versions of an article. Coding was performed by using QDA software. Frame semantics⁹ came into play in order to assess the current state of knowledge displayed in the articles’ content and to evaluate what was considered noteworthy at what time in the article. The additional analysis of real-world events (being located on the meta-level, see above) then helped to identify some of the trends and patterns in the articles’ development. Besides qualitative assessment, the focus had been set on data for statistical and quantitative analysis, which was recorded manually for additional results.

Konerding (1993) used findings in frame theory for a study with a lexicographic-lexicological approach. In his approach he redesigned frame theory to “a theory for knowledge representation/realization” (= Theorie der Wissensdarstellung/vergegenwärtigung; translated from Konerding, 1993: 92) and exemplified how linguistic frame analysis can be applied to a variety of purposes by employing frames empirically. In doing so, he developed a method to systematically characterize relevant slots of a frame by using a set of questions. He also invented a procedure called ‘hyperonym type reduction’ including a restricted set of highest hyperonyms to determine potential reference points or slots of any linguistic expression by retracing every one of them to such a highest-level hyperonym (cf. Ziem, 2014: 267). This procedure is used to identify the slots in a frame and is important for the implementation of frames as analytical instruments. As a result, only a relatively small set of German nouns occur as end elements in the reduction chain. In consequence, it is basically the slots in the frame that any lexeme (noun) evokes which correspond to the slots in the frame of a noun specified in Konerding’s approach. Nevertheless, the expression of these lexemes can be retraced via the procedure of hyperonym type reduction.

⁹ I understand frames as conceptual knowledge units that linguistic expressions evoke. They group slots and fillers as structural constituents to define a stereotypical object.

Use of Koneading's approach has been popular in German language research in order to document how concepts (of knowledge) have developed and which aspects of slots are focused in different types of discourse (cf. Ziem, 2014: 16). Therefore, it has been exemplified in several studies how his proposal of linguistic frame analysis can be applied to a variety of purposes by employing frames empirically (ibid.). Due to the wide range of possible applications of frames, they serve as a tool kit in my study to analyse content development in Wikipedia entries. Lexical items, in this case the headword, provide access to a considerable amount of subject knowledge in the corresponding article and display how they have developed over time. For means of my analysis the hyperonyms "artefact" (for 'lemon squeezer') and "event" (for 'European debt crisis') were identified as well as the additional reference points in the frame system of each hyperonym according to Koneading (1993: 309–340). In combination with a systematic question-answer-advance (e.g. for an object-related article, "What are features and characteristics of a lemon squeezer?" "How did this artefact originate?"), an encoding paradigm was defined to study the development of an article with respect to its content. Here, I specifically focused on the use of hyperlinks and their immediate text environment as potential fillers or information units within the systematic frame approach. Hyperlinks do not only act as navigation tools in the network of knowledge unfolded by the articles' editors but also as salient features within a narrative article as they draw the user's attention to specific areas of the text. Additionally, and for the benefit of a more granular analysis of the articles' development, topics from the discussion pages and ongoing real-world events were taken into account.

4. Analysis and discussion

As stated above, frame analysis in combination with a systematic question-answer-approach was used for means of encoding the wiki data. The methodology allowed dissecting and reasoning about the articles' development in the German Wikipedia both conceptually and systematically.

Recurrently analysing the article versions by a code system provided a perspective on the content's diachronic development. Additionally, the hierarchical structure of every article version was taken into account. Connections with endogenous activities of Wikipedia and real-world events, or so-called exogenous activities, could then be traced in the articles' development. These events were called trigger pulses (see above). On a quantitative scale, developments in the article structure became visible whenever a trigger pulse had been identified on the meta-level. Due to the code structure, it was possible to retrace the movement of the information unit around the hyperlinks within the articles' structure. The number of dots in each cell denotes the quantity of fillers or information units per movement type.

date / movement	3/05	6/05	10/0 5	8/06	12/0 6	3/07	4/07	5/07	6/07	7/07	8/07
launch	••••	•				••••				•	
inactive		••••	••••	••••	••••		•••	•••	••••	••••	••••
displaced						••••				•••	
deleted							••	••			
trigger pulses	article launch					writing contest				featured article	

Table 1: Movement of information units in “Zitronenpresse”

The methodological approach and applied code system made it possible to locate selected fillers and allowed statements about changes (i.e. if and when they had been made). As the slot-filler-combination is not likely to change very much in an article that deals with an artefact (Table 1), it changes more likely in an event-related topic (Table 2). Trigger pulses can be identified here, too, but the constant relevance of the topic is noticeable as well in the recurrent launch of new information units. Furthermore, it can be observed how some parts of the article content become more inactive or stable for some time.

date / movement	2/10	5/10	8/10	11/1 0	2/11	5/11	8/11	11/11	2/12
launch	••••	••••	•••	••	•••	••	••	••	••
inactive		•	••	••••	••••	••••	••••	•••	••
displaced		••	••••	•	••	•		•	••••
deleted			•		•	••		••	•••
trigger pulses	media coverage/ sovereign default				Operations by the EFSF				fiscal compact

Table 2: Movement of information units in “Eurokrise”

Trajectories and patterns of an interconnection of endogenous and exogenous factors are, in fact, visible in a feedback loop, e.g. when activity rises due to a featured article process, or ongoing events. As mentioned above, real-world events do affect the number of edits performed on an article; along with these come qualitative changes, which can be qualified on different linguistic levels. Results show that the underlying concept of each article, according to the conceptual frame approach applicable to

either an artefact or an event, is likely to be revised in its components after a trigger pulse.

The development of the articles also showed that the encyclopaedic character of entries (i.e. by stressing information about geographical place-names or names of important persons) evolves only over time. The encoding paradigm helps to set the focus on entities (as can be seen, for example, in numerous references to significant events in time, relevant places, cultural or public figures). Numerous fillers have been identified here, but other reference points or slots were also considered in later versions of an article covering different fields of knowledge representation (Table 3).

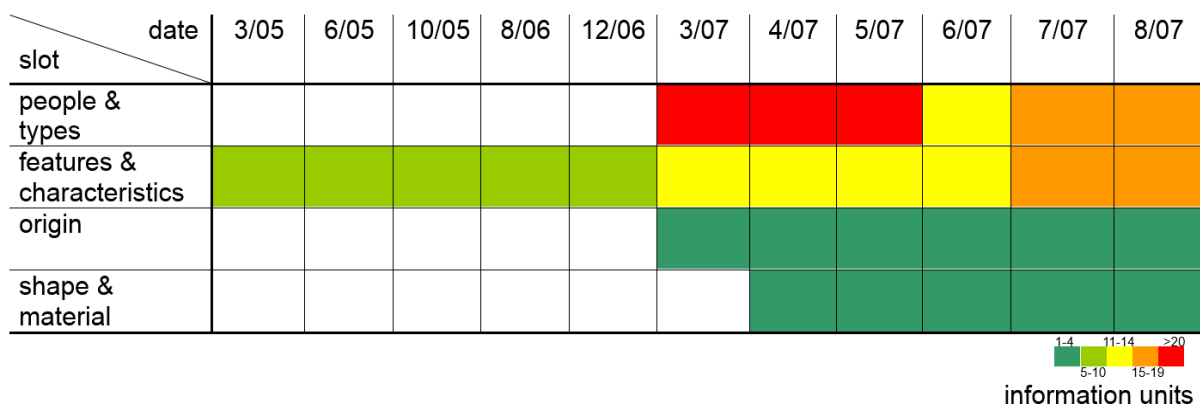


Table 3: Relevant areas of knowledge representation in “Zitronenpresse”

So far, the approach has proved to be useful in the identification of some indicators for interactions and activities within a wiki. However, to understand what information flows into the activated frame and what is relevant for an understanding of a ‘lemon squeezer’ or the ‘European debt crisis’, it can be helpful to enter deeper layers of the information units to identify key elements in the filler-slot structure. As already pointed out, ‘lemon squeezer’ refers to a frame around an artefact that is a kitchen utensil. In terms of this particular frame, high type and token frequencies over a significant period of time within this frame allow us to assume possible stable components or ‘entrenchments’ (cf. Ziem, 2014: 292–299). In fact, the slot ‘features & characteristics’ operated with different fillers: A lemon squeezer is used to make juice; is used for different citrus fruits; is designed to separate the pulp, etc. The phenomenon of a high type-frequency should be considered here as it underlines the importance of the slot ‘features & characteristics’ for the Zitronenpresse frame.

In the article ‘European debt crisis’ the consolidation of a token ‘Greek debt crisis’ was quite noticeable. In the German article versions the filler ‘Greek debt crisis’ could be placed in the slot ‘occurrence’ as Greece was one of the first countries to show a budget deficit. But the filler also matched the slots ‘correlations’ and ‘interference’ as budget crisis in Greece and beyond spread and bailout measures as well as Greece withdrawal

from the Eurozone were discussed. Also, a hyperlink ‘Greek debt crisis’ was recurrently used in the “see also” section as it relates to a topic similar to the discussed one in the article ‘European debt crisis’. However, a high token frequency consolidates the filler but weakens the slot. This means that the answer to a question “What is the European debt crisis?” may include the instance ‘Greek debt crisis’ as a sort of a default value. However, the exact description of this relationship remains open; at least in the examined article versions.

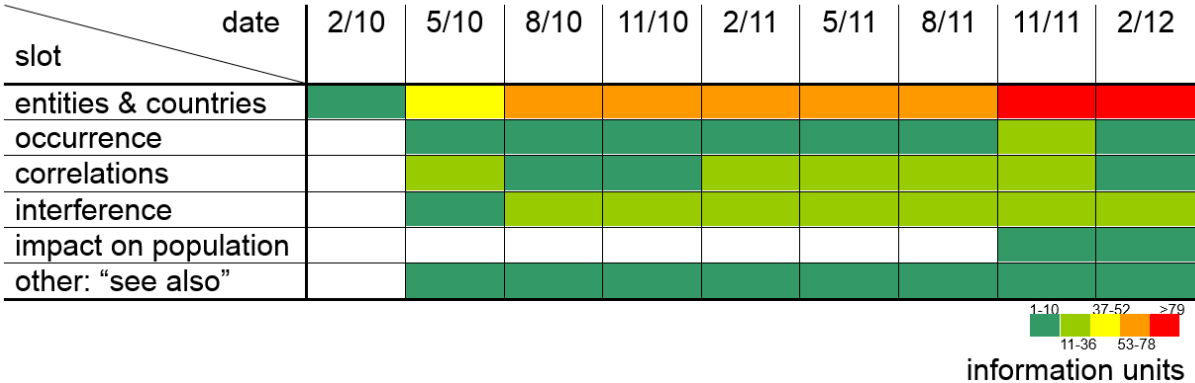


Table 4: Relevant areas of knowledge representation in “Eurokrise”

As we can see, the presented approach takes certain features of Wikipedia’s dynamics into account. Using this approach, I identified phases and interrelations, as well as some aspects of coherence in the interrelation of endogenous and exogenous factors when attention is drawn to the article and its development is triggered. The outlook on possible default values or the process of developing stable components is worth considering. Although the specific patterns of dynamic changes highlighted by this analysis will only be valid for a restricted period until the next edit, the principles derived from this approach should remain relevant and can be applied to other topics or information resources.

Of course, exploring the revisions of an entry is only one step in the multifaceted task of understanding what is important or relevant for both users and producers of a wiki. Certainly, this task needs a broad spectrum of research activities, for example dealing with general patterns of look-up behaviour (Müller-Spitzer et al., 2015), or classifying edits in collaboratively created articles (Daxenberger & Gurevych, 2013).

5. Relevance in lexicography

I previously pointed out that we can only use the results of so-called user generated content or bottom-up-lexicography for our own lexicographic products if we fully understand how the collaborative system works and what is important for the active user. So how can we use the understanding of the collaborative process that we have so far in institutional or professional lexicography? I want to emphasize three possible benefits of analysing Wikipedia dynamics:

- (1) Learning about the collaborative process.
- (2) Using an already existing collaborative product for expert lexicographic purposes.
- (3) Incorporating the collaborative process into an expert lexicographic product (or combining the two).

While pointing out some patterns and trajectories in the life cycle of an article within Wikipedia, we learned that trigger pulses as well as context sensitivity and coherence drive the development. Practical consequences of this are that information flow and information build up is subject to change according to the described factors. The more extensive but also potentially less stable contributions will be associated with whatever seems currently relevant. Thus, relevance has both positive and negative aspects for expert lexicographic purposes. However, in the long run, contributions that are highly contested or on the fringe of a topic will only have a short lifespan and will eventually be ‘overwritten’ during the article’s development, while more general, consensual information will remain. Such facets of the collaborative process should also be taken into account when using it for an expert lexicographic product. As Bon & Nowak (2013) emphasized, a procedure supplying entries with encyclopaedic or world knowledge can support text comprehension as well as (in my point of view) discourse comprehension.

Finally, a combination of direct user contribution via the collaborative process, e.g. in a semi-collaborative dictionary related either to object or language issues, and an expert lexicographic product should point out both more static and more dynamic views on the same topic. Effectively, as Lew (2011: 237) states, the opposition institutional versus collective dictionary may no longer be a sharp one. The discussed examples of Merriam-Webster’s Open Dictionary and the Macmillan Open Dictionary in Lew’s overview on English online dictionaries shows, however, that user-added entries do not meet the criteria for inclusion in the regular edition. However, Lew (2014: 25) and Taganova (2013) might agree on the point that “[t]he cooperation of readers and editors can turn beneficial for the dictionary compilers, as representatives of different interest groups and subcultures can make contribution to the Open Dictionary projects, indicating the words that lexicographers might miss out” (Taganova, 2013: 111). The lexical description of entire vocabularies, however, is a job better suited for language professionals (cf. Lew 2014: 17). A potential outlook is also to transfer the given approach to lexicography, by analysing the revisions of a collaborative dictionary entry as indicative of lexical change. However, additional work needs to be done in order to apply the insights gained from analysing dynamics of encyclopaedic-style Wikipedia entries to environments concerned with information on word-meaning and language comprehension. In any case, with recent advancements in user-generated environments, different views on language become available and may get users more actively interested in lexicographic work in general.

6. Conclusions

In this overview, I presented a short study on the developments in two articles from the German Wikipedia. By means of time series data, a certain pattern was observed which pointed to trajectories between endogenous and exogenous factors within Wikipedia's activity to produce and enhance articles. This pattern appeared to follow a life cycle with regard to the articles' entities.

I believe that this study, in particular the clarification of development patterns within articles, can contribute to a better understanding of collaborative induced dynamics. These results can be utilized when using Wikipedia entries or articles from other wikis for a different lexicographic product. The proposed model can also be used to predict the developments, thus facilitating the use of collaborative products in institutional lexicography. The model may also provide pointers to what is worth taking into account when using user-generated content. Finally, a combination of expert and collaborative knowledge should be considered when thinking about new lexicographic products.

7. Acknowledgement

The author would like to thank the reviewers for their valuable comments and helpful suggestions.

8. References

- Abel, A. & Meyer, Ch. (2013). The dynamics outside the paper: User Contributions to Online Dictionaries. *Proceedings of the 3rd Biennial Conference on Electronic Lexicography* (elex 2013). Ljubljana: Trojina, Institute for Applied Slovene Studies/Tallinn: Eesti Keele Instituut, pp. 179–194.
- Barrett, D. (2009). *MediaWiki*. Beijing: O'Reilly.
- Bon, B. & Nowak, K. (2013). Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki. *Proceedings of the 3rd Biennial Conference on Electronic Lexicography* (elex 2013). Ljubljana: Trojina, Institute for Applied Slovene Studies/Tallinn: Eesti Keele Instituut, pp. 407–420.
- Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond*. From production to produsage. New York: Lang.
- Busemann, K. (2013). Wer nutzt was im Social Web? *Media Perspektiven*, 7-8, pp. 391–399.
- Carr, M. (1997). Internet Dictionaries and Lexicography. *International Journal of Lexicography*, 10(3), pp. 209–230.
- Crane, R. & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *PNAS* 105(41), pp. 15649–15653.
- Daxenberger, J. & Gurevych, I. (2013). Automatically classifying edit categories in Wikipedia revisions. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 578–589.
- Döring, N. (2010). Sozialkontakte online: Identitäten, Beziehungen, Gemeinschaften. In W. Schweiger & K. Beck (eds.) *Handbuch Online-Kommunikation*.

- Wiesbaden: Springer, pp. 159-183.
- Emigh, W. & Herring, S. (2005). Collaborative authoring on the Web: a genre analysis of online encyclopedias. *Proceedings of the 39th Hawaii International Conference on System Sciences* (HICSS). Track 4. Volume 04. Los Alamitos: IEEE Press.
- Ferron, M. & Massa, P. (2011a). Collective memory building in Wikipedia: the case of north African uprisings. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pp. 114–123.
- Ferron, M. & Massa, P. (2011b). Studying collective memories in Wikipedia. *Journal of Social Theory*, 3(4), pp. 449–466.
- Juran, J. (1992). *Juran on quality by design. The new steps for planning quality into goods and services*. New York: Free Press.
- Kallass, K. (2015). *Schreiben in der Wikipedia: Prozesse und Produkte gemeinschaftlicher Textgenese*. Wiesbaden: Springer.
- Kaltenbrunner, A. & Laniado, D. (2012). There is no deadline: time evolution of Wikipedia discussions. *Proceedings of the 8th Annual International Symposium on Wikis and Open Collaboration*, A8.
- Keegan, B. et al. (2011). Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pp. 105–113.
- Konerding, K.-P. (1993). *Frames und lexikalisches Bedeutungswissen*. Tübingen: Niemeyer.
- Lehmann, J. et al. (2012). Dynamical classes of collective attention in Twitter. *Proceedings of the 21st international conference on World Wide Web*, pp. 251–260.
- Lew, R. (2011). Online dictionaries of English. In P.A. Fuertes-Olivera & H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, pp. 230–250.
- Lew, R. (2014). User-generated content (UGC) in online English dictionaries. *OPAL - Online publizierte Arbeiten zur Linguistik* 2014.4, pp. 8–26.
- Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *Paper for the 5th International Symposium on Online Journalism*. University of Texas at Austin.
- Mayer, F. (2013). *Erfolgsfaktoren von Social Media: Wie „funktionieren“ Wikis?* Berlin: LIT.
- Mederake, N. (2014). Artikel der Wikipedia aus lexikografischer und textlinguistischer Perspektive. In M. Mann (ed.) *Digitale Lexikographie*. Hildesheim: Olms, pp. 229–249.
- Meyer, Ch. (2013). *Wiktionary: The Metalexigraphic and the Natural Language Processing Perspective*. Darmstadt.
- Müller-Spitzer, C. et al. (2015). Observing online dictionary users: studies using Wiktionary log files. *International Journal of Lexicography*, 28(1), pp. 1–26.
- Stivilia, B. & Gasser, L. (2008). An activity theoretic model for information quality change. *First Monday*, 13(4). Available at:

<http://firstmonday.org/article/view/2126/1951>.

- Stvilia, B. et al. (2005a). Information quality discussions in Wikipedia. *Technical Report ISRN UIUCLIS--2005/2+CSCW*.
- Stvilia, B. et al. (2005b). Assessing information quality of a community-based encyclopedia, In F. Naumann, M. Gertz, & S. Mednick (eds.). *Proceedings of the International Conference on Information Quality (ICIQ 2005)*. Cambridge, Mass.: MIT, pp. 442–454.
- Stvilia, B. et al. (2008). Information quality work organization in Wikipedia. *JASIST*, 59(6), pp. 983–1001.
- Taganova, T. (2013). New Words in Contemporary Dictionaries of the English Language: Are Words Invented by the Society or is the Society Changed by Words? In O. Karpova & F. Kartashkova (eds.) *Multi-disciplinary Lexicography: Traditions and Challenges of the XXIst Century*. Newcastle upon Tyne: Cambridge Scholars, pp. 103–113.
- Vossen, G. & Hagemann, S. (2007). From Version 1.0 to Version 2.0: A Brief History of the Web. In J. Becker et al. (eds.). *ERCIS Working Papers*, Vol. 4. Available at: https://www.ercis.org/sites/www.ercis.org/files/pages/research/ercis-working-papers/ercis_wp_04.pdf.
- Wang, R. & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), pp. 5–33.
- Wiegand, H. E. et al. (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung*. Berlin: De Gruyter.
- Ziem, A. (2014). *Frames of understanding in text and discourse: theoretical foundations and descriptive applications*. Amsterdam: Benjamins.

Websites:

- Algemeen Nederlands Woordenboek*. Accessed at: <http://anw.inl.nl/> (6 July 2015)
- BuzzWords*. <http://www.macmillandictionary.com/buzzword/> (22 May 2015)
- elexiko*. <http://www.owid.de/wb/elexiko/gruppen/index.html>. (22 May 2015)
- Macmillan's Open Dictionary*. Accessed at: <http://www.macmillandictionary.com/open-dictionary/>. (22 May 2015)
- Merriam-Webster's Open Dictionary*. Accessed at: <http://nws.merriam-webster.com/opendictionary/>. (22 May 2015)
- Wikipedia*. Accessed at <https://de.wikipedia.org/>. (March 2005 – February 2012)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Towards a Pan European Lexicography by Means of Linked (Open) Data

Thierry Declerck¹, Eveline Wandl-Vogt², Karlheinz Mörth²

¹ DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

² ACDH-ÖAW, Sonnenfelsgasse 19, 1010 Vienna, Austria

E-mail: declerck@dfki.de, Eveline.Wandl-Vogt@oeaw.ac.at, Karlheinz.Moerth@oeaw.ac.at

Abstract

In the context of the expanding Linked (Open) Data framework (LOD), work has started to encode linguistic resources in the same format as performed for the data sets present in the LOD, and which represent mainly domain specific knowledge. This approach has been extensively discussed in the W3C Ontology-Lexica Community Group, resulting in the “OntoLex” model, and is also being supported by the European LIDER project, leading for example to extensions of the recently created Linguistic Linked Open Data (LLOD) cloud, and by the European FREME project, applying LLOD principles to various industrial use cases. This development is highly relevant to the goals of the European Network of e-Lexicography (ENeL) COST action, and in this respect we performed a number of experiments to encode lexicographic data of various ENeL partners in a LLOD compliant format. We report in this paper on the first steps taken in the cooperation between ENeL and the other aforementioned projects, providing some detail regarding the encoding model we use: OntoLex.

Keywords: e-Lexicography; Linked Open Data; Multilingualism

1. Introduction

In the context of the European Network of e-Lexicography (ENeL) COST action¹ a question we ask is whether a pan European lexicology and lexicography is conceivable. Concerning the potential European lexicology, this question leads us to searching for commonalities in the structure and the concepts used in the various languages of Europe. Therefore, we need to establish a certain level of interoperability in the description of those languages. Are we able for example to detect and markup shared etymologies between European languages, optimally by automatically consulting machine-readable versions of the dictionaries encoding the properties of the languages? Concerning the potential European lexicography, we aim for example to generate multilingual dictionaries on the basis of the shared concepts or meanings that can be detected between digital versions of monolingual dictionaries. For this we need to have access to a standardized representation of the concepts and meanings used in the different dictionaries for describing their entries. By standardized representation we mean the possibility to anchor the various but similar descriptions of meanings for a headword in different dictionaries into a shared and dereferentiable source on the web.

¹ See <http://www.elexicography.eu/>

Firstly, on this basis, one can attempt to respond to some research questions such as: How many common roots (etymology) are there across European languages, or are there common neologisms²? Are there pan European words, or pan European concepts? How to best utilize pan European multilingual corpora³? Or how to cross-link, and (partially) merge, the authoritative dictionaries that have been developed over the years by many participants of the ENeL COST action?

The recent development of the Linked (Open) Data (LOD) framework⁴ and more specifically of the Linguistic Linked Open Data (LLOD) cloud⁵ seem to offer an ideal environment for solving some of the interoperability issues we mentioned above, while also providing a good platform for linking the content of the authoritative dictionaries to other types of data available on the (semantic) web. We present in the next sections the basic ideas of the LLOD framework and the representation model used for publishing and linking language data in this cloud: OntoLex⁶.

2. Linguistic Linked Open Data

For this paper we adopt the definition of Linked Data given by Wikipedia: “In computing, linked data (often capitalized as Linked Data) describes a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried”⁷. Data sets that have been published in the Linked Data format can be visualized by the so-called Linked Open Data Cloud diagram⁸ or also by other means like the Linked Open Data Graph⁹.

In the context of this further expanding Linked Data framework, work has started to encode linguistic resources in the same format as already existing linked data sets, which primarily consisted of “classical” knowledge objects and entities. In those data sets, language data is mainly used as human readable information encoded for example in the RDF(s) annotation properties “label”, “comment” and the like.

² One can consider expressions such as “Grexit” or “Brexit”, which seem to be used across Europe.

³ Here, we consider, for example, the Europarl Corpus (<http://www.statmt.org/europarl/>)

⁴ See <http://linkeddata.org/> for more details

⁵ See <http://linguistics.okfn.org/tag/lod/> for more details.

⁶ <https://www.w3.org/community/ontolex/>

⁷ http://en.wikipedia.org/wiki/Linked_data. A more technical definition is given at <http://www.w3.org/standards/semanticweb/data>

⁸ <http://lod-cloud.net/>

⁹ <http://inkdroid.org/lod-graph/>

Recently, some researchers¹⁰ in the field of Human Language Technology (HLT) and Semantic Web technologies started to work on models and their implementation that would elevate the language data used in existing LOD data sets to the same type of representation as is the case for the encyclopaedic knowledge they were “commenting” and “labelling”. Cooperation on those topics has been established between, among others, the Working Group on Open Data in Linguistics¹¹ and with the European FP7 Support Action “LIDER”¹². These joint efforts have led to the establishment of a linked data cloud of linguistic resources, which is called Linguistic Linked Open Data (LLOD)¹³ and whose data sets are not only linked to other language data sets, but also to the encyclopedic data sets in the LOD. The Linguistic Linked Open Data cloud is also visualized by an online diagram¹⁴, which itself is derived from information contained in the LingHub repository¹⁵ developed in the context of the LIDER project. More recently, cooperation has been established with the H2020 project “FREME” on the automatic enrichment of digital content¹⁶. In fact, FREME is providing for industrial use cases that are using the LLOD framework. We investigate, in the context of ENeL, if such approaches to LLOD can be applied to authoritative lexicons for (partial) publishing and linking those within this cloud.

The model “OntoLex” is at the core of the publication of language data and linguistic information in the LLOD. This model results from the W3C Ontology-Lexicon community group¹⁷. Since this model was originally based on LMF¹⁸, which is itself the ISO standard for Natural Language Processing (NLP) lexicons and Machine Readable Dictionaries (MRD), it is an appealing model for lexicographers who are seeking to publish their data in the LOD. In the next section, we briefly present the current state of OntoLex.

3. OntoLex

The OntoLex model has been designed using the Semantic Web formal representation languages OWL, RDFS and RDF¹⁹. It also makes use of the SKOS and SKOS-XL

¹⁰ See for example Chiarcos et al. (2013a) and Chiarcos et al. (2013b)

¹¹ See <http://linguistics.okfn.org/> for more details.

¹² See <http://www.lider-project.eu/> for more details.

¹³ See <http://linguistics.okfn.org/tag/llod/> for more details.

¹⁴ <http://linguistic-lod.org/llod-cloud>

¹⁵ See <http://linghub.lider-project.eu/>. LingHub is an open and domain adapted (semantic) repository for language resources. All metadata are available in standardized Semantic Web representation languages.

¹⁶ See <http://www.freme-project.eu/>

¹⁷ See also <https://github.com/cimiano/ontolex>, complementary to <https://www.w3.org/community/ontolex/>

¹⁸ See (Francopoulo et al., 2006) and <http://www.lexicalmarkupframework.org/>

¹⁹ See <http://www.w3.org/TR/owl-semantics/>, <http://www.w3.org/TR/rdf-schema/> and <http://www.w3.org/RDF/> respectively.

vocabularies²⁰. OntoLex is based on the ISO Lexical Markup Framework (LMF) and is an extension of the *lemon* model, which is described in (McCrae et al., 2012). OntoLex describes a modular approach to lexicon specification, thus allowing the e-lexicographer to depart from the “book” view that the headword is the (unique) entry point to information encoded in a dictionary. Senses, usages, concepts, etc. can be independently described, accessed and are all linked to what was considered the headword, and which is now encoded as a virtual entry in a RDF model.

With OntoLex, we can advocate for the fact that all elements of a dictionary entry can be described independently from each other and connected by explicit (typed) relation markers. Now, the components of a dictionary entry can be distributed in a network and linked together by RDF encoded relations/properties. An important aspect of this model is also the relation called “reference”. This represents a property that supports the linking of senses of lexicon entries to knowledge objects available in the LOD cloud. This reflects also our view that the meaning of a lexicon (or dictionary) entry is no longer necessarily encoded in the lexicon (or dictionary) but can be referred to in appropriated resources on the (semantic) web.

In practicality, this means that a dictionary author does not need to describe all components or elements of an entry in detail, but that she/he can also draw on existing elements (e.g. the etymology of a word), and can simply refer to it. We are convinced that these properties of the model can facilitate and support the cooperation between scientific lexicographers, and that this can result in virtual and collaborative research environments in the lexicographical field.

Figure 1 below displays the core model of OntoLex²¹. Boxes represent classes of the model. Arrows with filled heads represent object properties, while arrows with empty heads represent the Sub-Class relations. In arrows labeled 'X/Y', X is the name of the object property and Y the name of the inverse property.

²⁰ SKOS stands for Simple Knowledge Organisation System, see also <http://www.w3.org/2004/02/skos/>

²¹ The figure and the explanations are taken from the wiki page of OntoLex: http://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

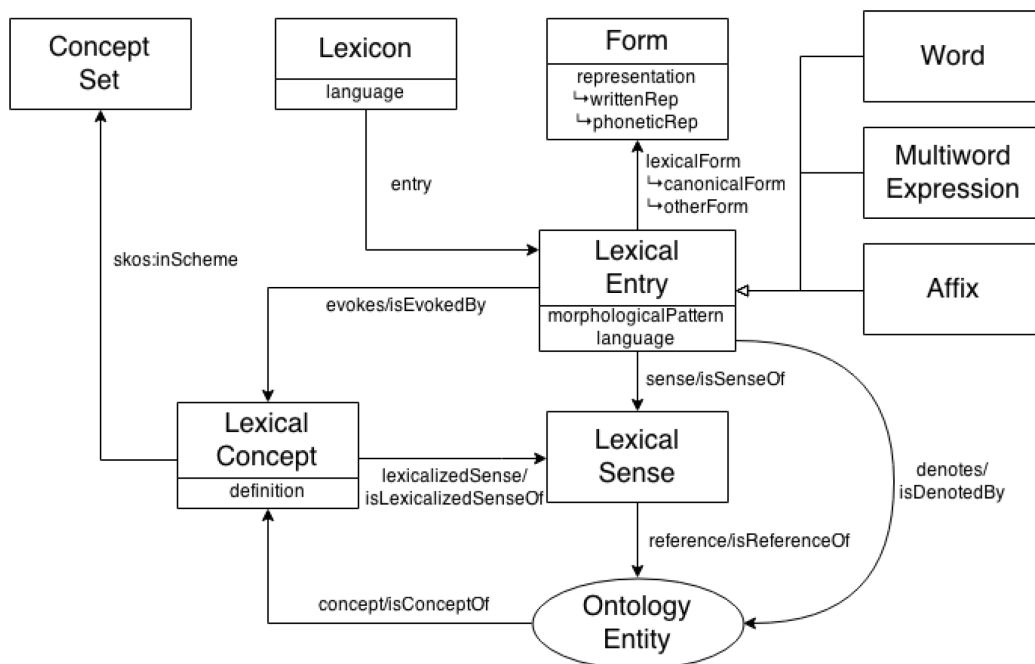


Figure 1: The core model of Ontolex.

Figure created by John P. McCrae for the W3C Ontolex Community Group.

We applied this model on a small list of different types of lexical resources made available by participants of the ENeL network, and we describe this encoding process in the next section.

4. First manual Experiments

In order to test our intuition about the use of OntoLex for the publication of existing authoritative lexicographic resources in the LOD, we provided, as a proof of concept, a manual encoding of some example data provided by ENeL participants in the OntoLex format. The example data we used were taken from:

- 2 Austrian dialect dictionaries (Tustep/XML and Word)
- 1 sample of a Slovak dictionary (XML, + PDF/Word)
- 1 Slovene XML dictionary (XML, based on the LMF standard)
- 2 TEI encoded Arabic dialects (in TEI)
- 1 Sample from a Bask–German dictionary (XML)
- 1 Sample from a French lexicon (extracted from Wiktionary)
- 1 Limburg lexicon (Excel)

- 1 Sample from the KDictionary multilingual source (XML file)
- Sample from the Digital Scottish Lexicon (Old Scottish, html + 1 example in TEI)
- 1 Lexicon extracted from a corpus of “Baroque German”

Every dictionary has been encoded in the OntoLex format as an instance of the `ontolex:lexicon` class, using the `ontolex:entry` object property to indicate inclusion of an entry.²² The class `ontolex:lexicon` thus serves here basically as a container for lexical entries. Below we display the example for the “Wörterbuch der bairischen Mundarten in Österreich” (WBÖ)²³, on which we will focus for the details of the manual encoding in OntoLex²⁴.

```

ontolex:WBÖ
  rdf:type ontolex:Lexicon ;
  rdfs:comment "Dictionary of Bavarian Dialects in Austria"@en ;
  ontolex:entry ontolex:lex_trupp ;
  ontolex:entry ontolex:lex_trüllen ;
  ontolex:entry ontolex:lex_trüsche ;
  ontolex:language "bar"^^xsd:string ;

```

In the code displayed above, the reader can see that the lexicon class is acting as a container, in which original entries (here of the WBÖ) are included via the OntoLex property `ontolex:entry`. The example can be read in natural language as “WBÖ is an instance of the class “Lexicon”, which lists dictionaries and lexicons”. WBÖ deals with the Bavarian Language (“bar”). WBÖ has three entries, “trupp”, “trüllen”, “trüsche”. It is important to note that this instance of a `ontolex:lexicon` class is indexed by an URI. In our case it is a local one (no longer accessible on the web): <http://www.w3.org/ns/lemon/ontolex#wbö>. And this is valid for all instances we will see examples of below: they all have an URI, so that their content can be accessed by any sparql queries²⁵.

In the example above we list only a few examples of entries, as the described experiment was initially performed manually, as a proof of concept.

The entries that are marked in the example of the WBÖ lexicon above in the range of the `ontolex:entry` object property are themselves instances of the `ontolex:LexicalEntry` class. The example for the lexical entry “trupp” is displayed below. The lexical entry

²² All the examples discussed in this section refer to Figure 1.

²³ <http://www.oeaw.ac.at/icl/tt/dinamlex-archiv/WBOE.html>

²⁴ We display all the examples of our OntoLex encoding using the so-called Turtle syntax. Turtle stands for “Terse RDF Triple Language” and is an easily readable serialization of RDF statements. See <http://www.w3.org/TR/turtle/> for more details.

²⁵ SPARQL is a query language defined for RDF triples. See for more details <http://www.w3.org/TR/rdf-sparql-query/>

`ontolex:lex_trupp` also has some features associated with it, all marked by the use of either datatype or object properties²⁶. In the example below, `ontolex:sense` is an example of an object property, while, in the example above, `ontolex:language` is an example of a datatype property.

```
ontolex:lex_trupp
  rdf:type ontolex:LexicalEntry ;
  ontolex:denotes <http://live.dbpedia.org/page/Herd> ;
  ontolex:denotes <http://live.dbpedia.org/page/Social_group> ;
  rdfs:comment "An entry of WB0: Trupp"@en ;
  ontolex:canonicalForm ontolex:form_trupp ;
  ontolex:hasEtymology ontolex:ety_trupp ;
  ontolex:sense ontolex:trupp_sense1 ;
  ontolex:sense ontolex:trupp_sense2 ;
  ontolex:sense ontolex:trupp_sense3 ;
.
```

In the example above, we can see that a “canonical form” is defined for the entry. This is due to the fact that OntoLex is supporting the description of variants (regional, typographical, morphological etc.) that are shared by the same entry²⁷. In the “lex_trupp” example we can also see how OntoLex deals with semantic ambiguities. There are in this example two usages of the `ontolex:denotes` property. Consulting Figure 1 above, the reader can see that the “denotes” property links directly to an object outside of the “lexical domain”. In our case to DBpedia entries, but it could be any domain specific resource. Since we introduced this property twice, we have a clear indication with which we can apply a reference ambiguity. The entry “lex_trupp” also includes three uses of the `ontolex:sense` object property. This property is pointing at objects that are defined as a lexical semantics module within our lexicon space. An example of such a “sense”, as an instance of the class “`ontolex:LexicalSense`” is given below.

```
ontolex:trupp_sense1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "One lexical sense for entry Trupp"@en ;
  ontolex:hasRecord ontolex:rec_trupp1 ;
  ontolex:isSenseOf ontolex:lex_trupp ;
  ontolex:reference <http://live.dbpedia.org/page/Social_group> ;
.
```

As we can see, this object also indicates a DBpedia entry, via the `ontolex:reference` property. The difference between the “denotes” and the “reference” properties is that, in the one case, the domain of the property is an instance of `LexicalEntry` and, in the second case, it is an instance of the `LexicalSense` class. In the second case, we can

²⁶ The distinction between object and datatype properties refers to the fact that a property related to an object can relate either to another object in the ontology (an instance of a class) or to some literal data. See <http://www.w3.org/TR/owl-ref/> for more details.

²⁷ The details of the types of variants currently covered by OntoLex are listed at: http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Properties-and-Relations-of-Entries

establish lexical semantic relations between the instances of the class, and this motivates the introduction of this additional referential mechanism.

For both cases, the fact that we can link an entry or, better, a sense to an external resource, like DBpedia, gives access to related multilingual information that is encoded in such a resource. In the case of accessing “http://live.dbpedia.org/page/Social_group”, we can retrieve related information in many languages (and the potentially related entry in the corresponding language):

- http://fr.dbpedia.org/resource/Groupe_social
- http://de.dbpedia.org/resource/Soziale_Gruppe
- http://cs.dbpedia.org/resource/Sociální_skupina
- http://el.dbpedia.org/resource/Κοινωνική_ομάδα
- http://es.dbpedia.org/resource/Grupo_social
- <http://eu.dbpedia.org/resource/Gizarte-talde>
- http://id.dbpedia.org/resource/Kelompok_sosial
- http://it.dbpedia.org/resource/Gruppo_sociale
- <http://ja.dbpedia.org/resource/社会集団>
- http://ko.dbpedia.org/resource/사회_집단
- http://pl.dbpedia.org/resource/Grupa_społeczna

And we also obtain information regarding related Wikipedia categories, like:

- category:Sociology_index
- category:Social_groups
- category:Social_psychology
- category:Sociological_terminology

Looking at the page http://live.dbpedia.org/page/Social_group, the reader can see that there are many other types of information that can be accessed and linked to.

In the first example of the “lex_trupp” entry above, the reader can additionally see that we introduce a property “hasEtymology”, which is pointing to an instance of the class “ety(mology)”. With this step we further demonstrate how the organization of the digital dictionary can be modularized. All the etymology information contained in the original WBÖ is now contained in a well-defined class of ontology and the instances of this class can be enriched with information from other sources than the WBÖ. The current description of the etymological information included in this WBÖ entry is:

```
ontolex:ety_trupp
  rdf:type ontolex:Etymology_French ;
  rdfs:comment "Instance of a French etymology for the WBÖ entry
\"lex_trupp\" ;
  ontolex:hasCentury 17 ;
  ontolex:hasEtymologyForm "Troupe"@fr ;
  ontolex:isEtymologyOf ontolex:lex_trupp ;
  ontolex:language "French"@en
.
```


This description of the etymology data is very similar to that of the original WBÖ entry “Trupp”, which included the etymology in book form. We can create a specific lexicon for all etymological information contained in the WBÖ, and link the entries of this generated etymological lexicon to other etymological resources, and in fact merge all the compatible information. In this way, we are kind of outsourcing some of the information that is not inherently related to the Bavarian dialect to other sources of information that can be more complete and more accurate, since they were put together by real experts in the field of etymology. In doing so, we have a way to compare many lexicographic sources on their shared etymology data, and hence to establish a more complete list of roots that are shared across dictionaries in the LOD format.

A similar remark can be made on the senses (or meanings) of the original entry “Trupp”. In the instance `ontolex:trupp_sense1` displayed above, the reader can see that we link this particular sense via the “reference” property to an entry in DBpedia: http://live.dbpedia.org/page/Social_group. From there we can access all dictionaries and other sources that point to this URI, and thus establish a relation with those multilingual resources, accessed from now on by senses or meanings that are represented in DBpedia or in RDF versions of WordNet, and the like.

5. Lessons learned

This section is regarding some lessons learned during our manual OntoLex encoding of (aspects of) various lexicographic resources.

2.1 Representation versus Linking of lexicographical data

It very quickly became apparent that there is no need to provide for an OntoLex based representation of the complete information contained in an original dictionary. As in the case of WBÖ, we can be confronted with quite complex information structures, with different levels of embedding. And since such a dictionary has been developed over a number of years, with many different teams involved, internal consistency of the information and the way it has been encoded is not always given. And in general: the aim is not to propose yet another type of representation but to be able to link (and potentially merge) lexical information. We argue that only this type of information that can be linked should be converted in the OntoLex format and so be published in the Linked (Open) Data framework.

As we know, Tim Berners-Lee outlined four principles of linked data, which are listed on his famous page: <http://www.w3.org/DesignIssues/LinkedData.html>:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. So that they can discover more things.

We implemented this strategy, but for now limited it to a partial set of the information included in some of the dictionaries we have been working on and in particular the few examples from WBÖ. This limitation is for practical reasons: we so far encoded in OntoLex only the entries, the associated senses and the listed etymology information. This information, available in LOD compliant codes can be linked to related data sets in the Linked Data cloud. If now a user (a human or a machine) wants to access the full amount of information encoded in the WBÖ, we can for example add the full URL of this information under the `rdfs:see` property to any entry of WBÖ (or other dictionaries) we have been (partially) encoding in OntoLex. Therefore, any data set linking to one of our WBÖ entries encoded in OntoLex will also link to a dereferentiable resource. This will display the original WBÖ entry, as it is encoded in the database version of this dictionary. For example, information about locations that are relevant for an entry can be accessed at <http://wboe.oeaw.ac.at/dboe/indices/ort/A/1>, etc.

2.2 Manual transformation versus automated transformation

While in this paper we have mainly described a manual work for the OntoLex comprising the encoding of a few (complex) examples from different dictionaries, we also gained some insights into which aspects can be easily automated. If the dictionaries possess clear and consistent structures, so that entries, variants and senses can be easily detected and automatically extracted by means of the applications of patterns expressed as regular expressions in a programming language, automatic OntoLex encoding is possible. It is additionally desirable for the data we obtain to be in a structured format, for example Excel, XML and the like. As an example, we automatically mapped a concept-based lexicon for Limburg dialects, dealing with the anatomy of the human body, from its original Excel format into OntoLex. For this, only some lines of codes were necessary. The original data had 75,355 Excel rows. The lexicon lists in the first column (in a repetitive way) the anatomic concepts (mentioned using standard Dutch language), while in the second and third columns we have the lemma of the dialectal forms and lexical variations of those. The original lexicon is very large, since the concepts of interests are repeated in the first column of the Excel file for every possible variation in the dialect forms, but also for the naming of the different regions in which a variation for the basic concept was found.

After transformation in OntoLex, we have a sense lexicon of only 264 instances. Those

correspond in fact to the concepts used in the original lexicon in Excel, and for which 75,355 Excel rows were required. Here, we thus observe the compression power of such a representation in OntoLex (and in RDF in general). In this OntoLex representation, a sense (bovendeel van de rug; *upper part of the back*) has the following form:

```

ontolex:concept_limburg_100
  a    ontolex:LexicalConcept , skos:Concept , ontolex:SenseLexicon ;
  rdfs:comment  "Concept taken from a specific source for the Limburg Language, being
a questionnaire or a dictionary, etc."@en ;
  rdfs:label    "bovendeel van de rug"@nl ;
  ontolex:hasSource    ontolex:source_limburg_4 , ontolex:source_limburg_1 ;
  ontolex:isDenotedBy ontolex:lex_limburg_239 , ontolex:lex_limburg_1833 ,
ontolex:lex_limburg_1846 , ontolex:lex_limburg_1847 , ontolex:lex_limburg_1826 ,
ontolex:lex_limburg_1834 , ontolex:lex_limburg_1853 , ontolex:lex_limburg_1828 ,
ontolex:lex_limburg_1816 , ontolex:lex_limburg_1829 , ontolex:lex_limburg_1841 ,
ontolex:lex_limburg_1845 , ontolex:lex_limburg_1840 , ontolex:lex_limburg_1831 ,
ontolex:lex_limburg_1844 , ontolex:lex_limburg_1832 , ontolex:lex_limburg_1824 ,
ontolex:lex_limburg_1851 , ontolex:lex_limburg_1825 , ontolex:lex_limburg_1855 ,
ontolex:lex_limburg_1838 , ontolex:lex_limburg_1852 , ontolex:lex_limburg_1856 ,
ontolex:lex_limburg_733 , ontolex:lex_limburg_1837 , ontolex:lex_limburg_1827 ,
ontolex:lex_limburg_608 , ontolex:lex_limburg_5 , ontolex:lex_limburg_1839 ,
ontolex:lex_limburg_1843 , ontolex:lex_limburg_1745 , ontolex:lex_limburg_1842 ,
ontolex:lex_limburg_1823 , ontolex:lex_limburg_204 , ontolex:lex_limburg_1830 ,
ontolex:lex_limburg_1822 , ontolex:lex_limburg_1848 , ontolex:lex_limburg_1835 ,
ontolex:lex_limburg_1836 , ontolex:lex_limburg_1849 , ontolex:lex_limburg_1850 ,
ontolex:lex_limburg_1854 , ontolex:lex_limburg_525 , ontolex:lex_limburg_1817 ,
ontolex:lex_limburg_1821 .

```

In this representation, we can see that the sense “concept_limburg_100” has been “denotated_by” (the reverse property of “denotes”) many lexical entries. And this relation is being made explicit in the OntoLex model (and can be quantified), which is also a huge advantage, when compared to the original data.

We have also a total of 4,745 lexical entries, which represent the dialectal variations of the suggested 264 concepts expressed in standard Dutch. An example:

```

ontolex:lex_limburg_1894
  a    ontolex:LexicalEntry ;
  rdfs:label    "staartbot" ;
  ontolex:denotes    ontolex:concept_limburg_103 ;
  ontolex:hasPlace    ontolex:loc_limburg_28 , ontolex:loc_limburg_58 ,
ontolex:loc_limburg_63 .

```

In this example, we can see that a dialectal word “staartbot” is used for denoting the concept “limburg_103”, which is in standard Dutch “stuitbeen” (*coccyx*). We also get the information about the locations in which this word form is used.

To summarize this exercise: the reader can see how all elements of the original Excel file have been encoded as modules in the OntoLex lexicon for Limburg dialects, and that all instances of such modules are linked to each other using explicit and well defined properties. What is missing in our examples are links to external knowledge resources. This is the topic of the next section.

2.3 Linking to external resources

An issue we would like to consider is the possibility of automatically linking to external resources, those being both of linguistic nature or encyclopedic nature. We do not have an answer to this point for the time being. As a heuristic, while knowing that the Limburg lexical data concerns anatomy, and the reference language is standard Dutch, we can automatically query DBpedia for all entries that have a Dutch word marked with the additional “_(anatomy)” extension, such as for example: [http://nl.dbpedia.org/page/Hoofd_\(anatomie\)](http://nl.dbpedia.org/page/Hoofd_(anatomie)). However, this might only offer a very specific solution. We will study the algorithm implemented by BabelNet²⁸ for the automatic cross-linking of language resources in the LOD.

2.4 Quality of the source data

A final point we have to make: In the case of the Limburg lexicon described in this chapter, but also in the case of an automated transformation of two TEI-encoded lexicons of dialectal variants of Arabic into a preliminary version of OntoLex²⁹, we noticed that in a relevant number of cases some fields of the structured data were not correctly filled by those working on the data. In some cases text was added to the TEI slot “sense”, for example “?”, or “correct?”, and it also occurred that two or more values were included in the slot, instead of introducing a new “sense” slot for every meaning to be encoded.

6. Conclusions

We have been testing the use the OntoLex model, with very few additions, for encoding in the LLOD format the lexicographic resources of some participants of the ENeL Network. The next steps will consist of effectively publishing the results in the Linked Data cloud, after curation of some input data and the clarification of copy-rights issues.

Our current work consists of further automatizing the mapping between the original formats of other ENeL dictionaries and investigating more efficient linking strategies

²⁸ See <http://babelnet.org/>

²⁹ See Declerck et al. (2014b)

to encyclopedic sources. We are also extending our work to the encoding of so-called conceptual records used by lexicographers when carrying out field studies: they interview people in certain regions and ask them how they express certain concepts in their language. We started to use the ConceptSet and LexicalConcept constructs of OntoLex for this task.

We also need to establish clear links to temporal information, which is crucial not only for the encoding of etymology, but also for encoding all kinds of examples and publication dates. There is also a need to link certain lexicographic data to location information.

7. Acknowledgements

The work described in this paper is supported in part by the European Union, by the LIDER project (under Grant No. 610782), by the FREME project (under Grant No. 644771) and by the COST Action IS1305 “ENeL”. Our thanks go especially to the participants of the ENeL COST Action who provided for their data and advices.

8. References

- Cimiano, P. & Unger, C. (2014). Multilingualität und Linked Data. In: T. Pellegrini, H. Sack & S. Auer (eds.) *Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien*. Springer, pp. 153-175.
- Declerck, T. & Wandl-Vogt, E. (2014). Cross-linking Austrian dialectal Dictionaries through formalized Meanings. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress*, pp. 329-343.
- Declerck, T., Mörth, K. & Wandl-Vogt, E. (2014b). A SKOS-based Schema for TEI encoded Dictionaries at ICLTT, In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 26-31.
- Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J.-P., Cimiano, P. & Navigli, R. (2014). A Multilingual Semantic Network as Linked Data: *lemon-BabelNet*. In C. Chiarcos, J.-P. McCrae, P. Osenova & C. Vertan (eds.) *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, pp. 71-76.
- McCrae, J.-P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, P., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), pp. 701-719.
- Rehm, G. & Sasaki, F. (2014). Semantische Technologien und Standards für das mehrsprachige Europa. In B. Humm, B. Ege & A. Reibold (eds.) *Corporate Semantic Web*. Springer .
- Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria. (2006). Lexical Markup Framework (LMF). In *Proceedings of the fifth international conference on Language Resources and Evaluation*.

- Chiarcos, C., McCrae, J.-P., Cimiano, P. & Fellbaum, C. (2013a). Towards open data for linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen, L. Qin & and E. Hovy (eds.) *New Trends of Research in Ontologies and Lexical Resources* Springer, Heidelberg.
- Chiarcos, C., Moran, S., Mendes P.-N., Nordhoff, S. & Littauer, R. (2013b). Building a Linked Open Data cloud of linguistic resources: Motivations and developments. In Iryna Gurevych and Jungi Kim (eds.) *The People's Web Meets NLP. Collaboratively Constructed Language Resources*. Springer, Heidelberg.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Spell-checking on the fly?

On the use of a Swedish dictionary app

Louise Holmer, Ann-Kristin Hult, Emma Sköldbberg

Department of Swedish, University of Gothenburg

PO Box 200, SE-405 30 Gothenburg, Sweden

Email: louise.holmer@svenska.gu.se, ann-kristin.hult@svenska.gu.se,

emma.skoeldberg@svenska.gu.se

Abstract

Mobile application software – the app format – offers new ways of using dictionaries. However, so far, only very few user studies of dictionary apps have been conducted. In this article, we present and discuss the results of a web survey on the use of the app version of the monolingual *Svenska Akademiens ordlista* (the Swedish Academy Glossary, 13th edition, 2006), henceforth the SAOL.

The results show that the SAOL app is used mostly for checking spelling. A more surprising result, since the SAOL is not a definition dictionary, is that it is also frequently used for checking the meaning of words. For forthcoming versions of the glossary, the users request more definitions. Regarding the app, users wish for improved search functions, such as wildcard (truncated) search and cross references. The current app (of the 13th edition) is free. A majority of the users state that they are willing to pay a small sum for an app version of the 14th edition of the SAOL.

Keywords: dictionary apps; user study; web survey; app usage; SAOL

1. Introduction

The number of dictionary user studies has rapidly increased since the 1990s. This increase can be ascribed to a keener interest among lexicographers in dictionary users and their opinions, suggestions and needs (cf. Lew, 2011). User response has accordingly become an important factor to consider in the process of dictionary making.

Although dictionaries in the format of applications intended to run on mobile devices have become increasingly common (Gao, 2013), studies of the use of such apps are still scarce. Investigating the use of dictionary apps is important since it is reasonable to expect that this use differs from general dictionary use, in much the same way as mobile apps have changed media consumption in general. Unquestionably, the app format presents both new possibilities and new challenges compared to print and web dictionaries.

In this paper, we present the design and results of a web survey regarding the use of the app version of the *Svenska Akademiens ordlista* (Swedish Academy Glossary), henceforth referred to as the SAOL. The glossary covers general, contemporary Swedish. It includes about 123,000 headwords and provides the (unofficial) norm for spelling and inflection of Swedish words. The mobile app reflects the content of the 13th print edition of the glossary, published in 2006. This off-line app has been developed for several operating systems and can be used on smart phones and tablets. It is free to download and has been downloaded more than half a million times to date (May, 2015), which is a considerable number against the backdrop of Sweden's 9.6 million inhabitants.

The results of the survey are relevant to dictionary app developers and researchers focusing on app user studies. The results are also highly useful to the editorial staff of the glossary (which includes the authors of this paper) for three reasons. Firstly, no user study has previously been performed on *any* version of the glossary (print, CD, online or app), which is remarkable considering the glossary's relatively high status and high sales figures in Sweden. Secondly, a new, fully-revised and updated printed edition of the glossary, number 14, was published in April 2015. The Swedish Academy has announced the release of an app version of the new edition. Bearing this in mind, the editorial staff need to form a picture of the use of the *current* app as well as its strengths and weaknesses. Finally, a related app based on the contemporary dictionary of the Swedish Academy (*Svensk ordbok utgiven av Svenska Akademien*) from 2009, is under development by the same team of lexicographers and developers (see Holmer, von Martens & Sköldberg, 2015), and the outcome of the present survey will clearly be of great value in the design of this particular app.

In the next section, we discuss dictionary apps in general and app user studies. In section 3, we introduce the print and app versions of the SAOL. The results of our web survey are presented in section 4. Finally, in section 5, we conclude with a summary and a brief discussion.

2. Dictionary apps

As previously mentioned, monolingual and bilingual dictionaries are increasingly available via mobile phones and tablets. According to Gao (2013) and Rundell (2013), dictionary apps, as well as online dictionaries, offer major advantages over their traditional, analogue predecessors. For instance, they allow for multimedia presentations of micro-structural information (such as audio pronunciation and animations), cross-references and links to external websites. Apps can also be easily updated, which is beneficial for both producers and users. These features may account for the popularity that many dictionary apps are currently enjoying, in addition to the high accessibility of the dictionary content.

In the app development process, the lexicographic team must confront several

fundamental lexicographic issues. As Simonsen (2014b) points out, dictionary app development should always be based on the following six factors: user, situation, access, task, data and need. But as Holmer & Sköldbërg (2014) argue, there is a need for a more comprehensive discussion of the considerations that go into producing dictionary apps. The authors discuss apps as independent lexicographic resources compared to the printed and/or online dictionaries they are supposed to reflect. Furthermore, they raise the issue of whether the app format is suitable for all kinds of dictionaries.

So far, very few user studies on dictionary apps have been presented. One exception is Marellò (2014), who compares high school students' use of three versions – an Android app, an online version, and a paper copy – of the same bilingual dictionary. Another exception is Simonsen (2014a,b), who focuses on the use of an app version of an extensive medical resource that is widely used in Denmark. Based on his empirical data, Simonsen (2014b: 259–260) draws a number of conclusions regarding the mobile user and the mobile user's situation, a brief summary of which follows. Firstly, the mobile user is active and accesses information while on the move. Secondly, the mobile user's situation is characterized by multi-tasking, e.g. the user is doing several things simultaneously. The mobile user typically double-checks his/her knowledge and performs simple searches. Thirdly, the mobile user navigates the physical world and the user interface of the mobile device at the same time, which calls for a very simple and easy-to-use data access method. Finally, the size of the user interface means that complex data and long text segments are suboptimal.

In order to meet the needs of different user groups, it is required to obtain a deeper understanding of dictionary app users and how, when, and where they in fact use dictionary apps. The most common approach, when it comes to studies on dictionary usage in general, consists of collecting data by using a questionnaire (Tarp, 2008: 15ff.). The strengths and weaknesses of questionnaire surveys are well-known: questionnaires can be distributed to a relatively large number of users and the answers are usually relatively simple to process. The drawback is that this approach relies solely upon how accurate and conscious users are of their own dictionary use. Another relevant aspect is the number of questions that informants can cope with. Swedish media has highlighted the fact that Swedes are increasingly reluctant to answer surveys and questionnaires, which increases the margin of error for various types of statistical surveys carried out by, for instance, Statistics Sweden, a government agency (*Dagens Nyheter* 2015-01-18). Until now, surveys in the form of brief pop-up questions – small windows that emerge relatively discreetly on the user's screen – have not been common in lexicographical user studies, but this question format is common on various commercial sites.

Other research methods include interviews, (traditional) observations and protocols. Interviews, for example, make it possible for the interviewer to explain and expand upon potentially problematic questions. However, these methods are very time-consuming, which often means that the research data will be of limited size.

Finally, the researcher can make use of log files and other forms of web-based statistical tools, which have facilitated the retrieval of data regarding which words are looked up in a dictionary and how frequently (see e.g. Hult, 2012; Lorentzen & Theilgaard, 2012). In the process of dictionary making, this kind of data has been widely welcomed as a way of discovering lemma lacunae (Bergenholtz & Johnsen, 2005). The greatest advantage of the log file method is the large amount of relatively easily processed data that can be generated. Another advantage is that user activities are observed without the presence of a researcher; i.e., the phenomenon of the “observer’s paradox” is not an issue here. On the other hand, log files give no information about users. Consequently, researchers are left in the dark about customary background information and relevant issues concerning users’ lexicographical needs and preferences.

Log files and server based statistics make it possible to gain knowledge of the use of *online* dictionaries. App developers and lexicographers seeking insight into user behavior of *off-line* dictionary apps may be supported by mobile app measurement and advertising platforms like Flurry Analytics from Yahoo! (<http://www.flurry.com/>). Today, Flurry tracks more than 540,000 apps, including Skype and Snapchat. This platform allows the lexicographic team to gain a deeper understanding of which app versions and operating systems are used, which iOS versions and device models are running, etc. as well as how often the app is used and the length of the average session. In addition, the developers get information about which headwords are frequently looked up, and about spell-check use. It should be said that the SAOL app was not equipped with such statistical software at the time of the survey.

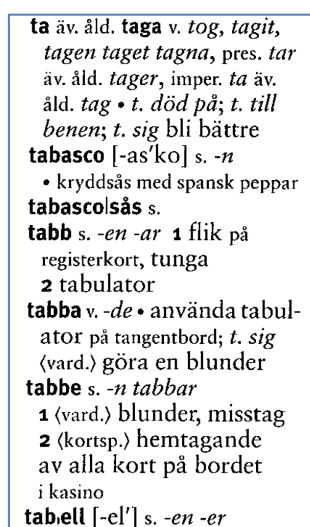
According to Tarp (2008), the best way to gain a deeper insight into user behaviour, is to combine different types of research methods. See e.g. Hult (2012) who combines a web questionnaire with log files, Lorentzen & Theilgaard (2012) who combine data from Google Analytics and log files, and Holmer & Sköldbberg (in press), who make use of Google Analytics combined with a pop-up question survey, examining the use of a Swedish, commercial synonym dictionary site.

3. The 13th edition of the *Swedish Academy Glossary*

The SAOL is financed by the Swedish Academy, and the editors are employed by the Department of Swedish at the University of Gothenburg, Sweden. The very first edition was published as early as 1874. A fully revised and updated edition of the glossary has since been published about every 10th year. The 13th edition was published as a printed book in 2006. In 2007, a CD version of the same edition, *SAOL Plus*, was released. The electronic format was used to provide all semantically motivated inflected forms for every headword (cf. Berg, Holmer & Hult, 2008). The CD also featured an advanced fuzzy search and a full-text search function. The 13th edition of

the glossary was published online in 2009, but only as a facsimile.¹

As previously stated, the glossary holds about 123,000 headwords and provides information on spelling, inflection and part of speech for each headword. About one fifth of the headwords are briefly defined, commented on or syntactically exemplified (Berg, Holmer & Sköldberg, 2010). For solid compound lemmas, only the part of speech is given, usually in abbreviated form (“v.” for ‘verb’, etc.). Irregular verbs are presented with their full inflectional forms. Some of these features can be seen in Figure 1.



ta äv. åld. **taga** v. *tog, tagit, tagen taget tagna*, pres. *tar*
äv. åld. *tager*, imper. *ta* äv.
åld. *tag* • *t. död på*; *t. till benen*; *t. sig bli bättre*
tabasco [-as'ko] s. -n
• kryddsås med spansk peppar
tabascosås s.
tabb s. -en -ar **1** flik på registerkort, tunga
2 tabulator
tabba v. -de • använda tabulator på tangentbord; *t. sig (vard.) göra en blunder*
tabbe s. -n *tabbar*
1 (vard.) blunder, misstag
2 (kortsp.) hemtagande av alla kort på bordet i kasino
tabell [-el'] s. -en -er

Figure 1: An example from the print version of the SAOL 13 including the verb *ta* (‘to take’) and the noun *tabasco* (‘tabasco’)

An app version of SAOL 13 was contracted and financed by the Swedish Academy and developed by the Swedish app development agency Isolve AB. The editors and system developers of the SAOL were mainly involved in the final test stage of the app development process. The app version of the SAOL 13 was derived from the aforementioned digital CD-version of SAOL, *SAOL Plus*, thus providing the full set of inflected forms for each headword. In comparison to the CD, all inflected forms in the app are displayed by default, which is not the case in *SAOL Plus*, where this setting is optional.

The app was released in November 2011 and was initially available only for iOS and Android phones and tablets. There was a subsequent release for Windows Phone and Nokia Symbian. The app works off-line, is free of charge and, as previously stated, has been downloaded more than half a million times (although, of course, the number of

¹ Since a few years ago, the different editions of the glossary can also be accessed through an advanced search interface (SAOLhist.se), which is mainly used by scholars.

active users is lower). Some of the downloads can be ascribed to the popularity of word games such as Scrabble and WordFeud, where the SAOL lemma list and inflectional rule set are, or can be used as, standard.

The main functions of the app consist of simple word search and crossword assistance. In addition to that, users can share entries via email and messaging, and use bookmark and history functions. The app also contains miscellaneous information such as a selection of new and excluded lemmas in the SAOL 13 as compared to previous editions, and information about the Swedish Academy. The “*More*”-section contains user instructions, abbreviations used in the SAOL and an email address that allows users to contact the developers.

The SAOL app is simple in its design (for a review of the app, see Hoel, 2012). For example, there are no hyperlinks, and wildcard search or full-text search functions are not available. See Figure 2 for screenshots of the SAOL app start page and samples of entries.

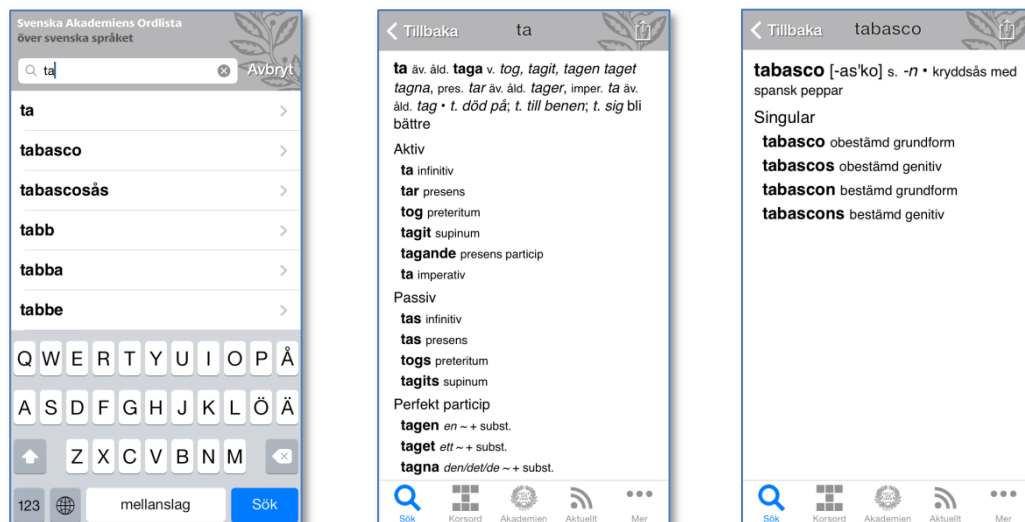


Figure 2: Left: screenshot of the lemma list of the SAOL app on an iPhone. Middle: the entry *ta* (‘to take’) with inflected forms. Right: the entry *tabasco* (‘tabasco’) with inflected forms

Lew (in press), makes an important distinction between *storage space* and *presentation space*, which is highly relevant in the app context. When it comes to the SAOL, a majority of the entries are rather short (see the entry *tabasco* in Figure 2). In that respect, the glossary is well suited for the app format.

4. The SAOL app web survey: method description and results

A web survey was considered the best option for our purposes. First, a pilot study was performed to test the questions and multiple choice answers. The pilot study consisted of 20 questions and was performed in December 2014. We received 44 responses, mainly from our colleagues and students at the Department of Swedish at the University of Gothenburg. Based on the results and the comments from pilot

respondents, the questionnaire was modified and some additional questions were included.

The final questionnaire consisted of 24 questions in Swedish intended to cover four main areas:

- User behaviour – frequency of use, typical function, typical use of app features, etc.
- Design and layout of the app
- Future development – suggestions and preferences for forthcoming versions
- Background information about the respondents

We considered it highly important to keep our questions brief and concise as well as to keep the number of questions to a minimum. Our aim was to limit participation in the study to five minutes (cf. Müller-Spitzer, Koplenig & Töpel, 2012: 429). There were many possibilities for users to add comments and no question was mandatory. A respondent could therefore skip a question (the downside being that there was no reminder function if the respondent had forgotten to reply to a question). The survey was distributed with the aim of reaching the target user group: people who actually use the app version of the SAOL. The web survey link was spread mainly via social media, such as Twitter and Facebook, and was published on some University web pages and in a well-known online Swedish language magazine (*Språktidningen*). The link to the questionnaire was open for about a month. Full anonymity was guaranteed (no IP-logging or other logging of browsers, devices, etc.). The web survey was powered by Webropol.

Altogether 264 questionnaires were submitted. The internal dropout rate was very low, that is, almost everyone answered all 24 questions. Moreover, many respondents took advantage of the several opportunities to add comments, which resulted in a great deal of very useful feedback about the SAOL in general and on specific app issues.

The following sections present the results of the respondents' background information, usage of the app, lookups, suggestions for a future version of the app and pricing. Finally, some examples of useful comments from the submitted questionnaires are highlighted.

4.1 Respondents: background information

The respondents were asked background questions about year of birth, gender, native language, level of education and principal occupation. Their answers show that they were between 20 and 89 years old. The mean age was 43 and the median age was 41 years old. Gender distribution was about 60% women and 36% men; the remaining

percentage answered “other”. More than 90% of the respondents were native speakers of Swedish. The other languages mentioned more than once were Finnish, Polish and German. The respondents were highly educated: more than 80% held a university degree, of which about 10% reached postgraduate level. Nearly 70% of the respondents were employed, about 17% were students and 10% retired. To summarise, the typical respondent involved in the study is a highly-educated professional woman in her early 40s whose native language is Swedish. However, based on this information alone, we are hesitant to draw definitive conclusions concerning the *typical* user of the SAOL app, as we assume that certain users are more likely than others to respond to surveys.

4.2 App usage: frequency and sought information

As mentioned in section 4, the target user group consisted of persons who actually use the SAOL app. The results show that more than 50% of respondents use the app on a weekly basis and an additional 28% use it every month. We also learned that the majority of the respondents have not read the SAOL app user instructions, which is not very surprising. Svensén (2009: 459) states that “it is a truth universally acknowledged in lexicographic circles that user’s guides are very seldom consulted”. However, although a majority of the respondents had not read the instructions, 23% had done so. Considering this fact, there are good reasons to include both user instructions and information about the dictionary itself in the app.

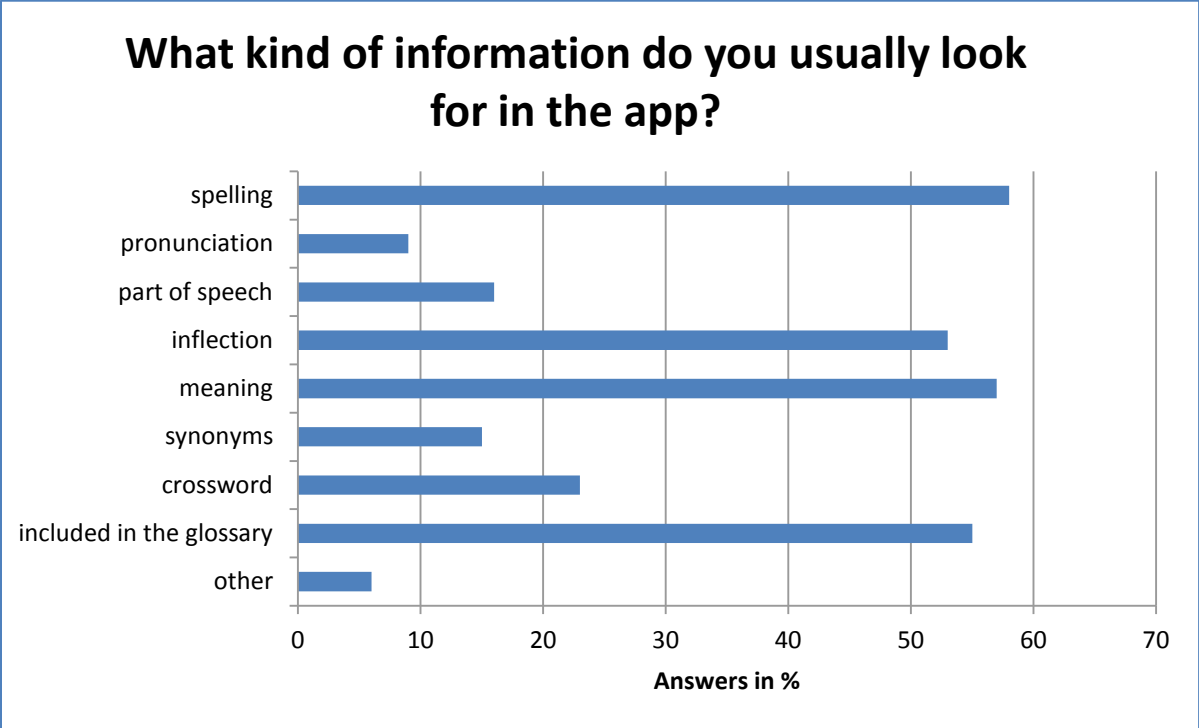


Figure 3: Answers to the question “What kind of information do you usually look for in the app?” (our translation). (Respondents could select more than one option)

One of the most important questions for the editorial staff concerned what kind of information the respondents most commonly search for. As Figure 3 shows, about 57% of respondents mostly use the app to check spelling or meaning. About 54% use it to check “if the word is included in the glossary”, which may be related to the important role of the glossary as a key for word games like Scrabble. In the fourth major category, 53% look for “inflection”. This supports the editorial decision to emphasize the full set of inflected forms by default in the app, compared to the limited information given in the print version.

Another question was: “How often do you find the information you are looking for in the app?”. About 28% answered “always”, roughly 70% answered “often”, and about 2% stated “sometimes”. No respondent answered “seldom” or “never”. To sum up, a vast majority of the respondents always or often find the information they are looking for in the app.

The responses to the two questions above may be inter-related. A cross-tabulation between the two questions shows that a majority of the respondents using the app for spelling, “often” or “always” find the information they are looking for. The same applies to respondents looking for information on inflection, as well as, surprisingly, those who are looking for meaning. This was a rather unexpected result since meaning is not one of the main information categories, although about a fifth of the lemmas have some kind of, usually very brief, explanation. The fact that so many users search for information on meaning in the glossary is not unexpected *per se*. A majority of the users are in all likelihood unaware of the difference between a glossary and a dictionary containing more extensive definitions. It is, however, striking that such a large number of respondents are satisfied with the information concerning meaning with which they are provided. This can possibly be related to the specific group of respondents in the study and the words they look up (see section 4.4 below).

4.3 App usage: when and where?

As referred to in section 2, Simonsen (2014b) states that the mobile user typically performs simple searches. According to his findings, dictionary app users are frequently on the move while using the device. Based on our data, we are hesitant to draw major conclusions concerning the typical mobile user situation. The glossary includes a large number of headwords but the information provided for each word is strictly limited and does not constitute a challenge to the user from a cognitive perspective. A clear majority (about 75%) of the respondents stated that they use the app when they are writing a text, i.e. in productive situations. This result was expected *a priori*, given the information that the glossary offers regarding spelling and inflection. However, as many as 35% of respondents claimed to consult the app while they are reading; i.e. in receptive situations. Finally, about 45% of respondents mentioned that they also look words up during conversations. We find it likely that

they consult the glossary with the intention of checking if a specific word or inflected form is “accepted” by the Swedish Academy. To summarise, the responses concerning typical user situations are consistent with the answers concerning what kind of information is typically sought when using the dictionary app.

Another question asked where the dictionary app was typically used. With reference to the question posed in the title of this paper, only a few respondents (about 16%) answered that they use the app on the fly; e.g. when walking down the street. Almost the same percentage of users responded that they consult the SAOL app in cafés, restaurants, etc. However, a clear majority of lookups take place at home or at work.

A majority of the respondents, about 64%, use the app on an iPhone and about 35% use it on another phone. The option “other phone” may seem a bit vague, but our background knowledge from the app developers tells us that Android is the second most common operating system, although there are also some Nokia Symbian and Windows Phone users as well. It is much more common to run the app on phones than on tablets; only 23% use tablets. This may be a result of the general relative abundance of phones.

4.4 Lookups

The editorial staff of the SAOL was naturally interested in what kind of words users want to look up when accessing the app. We therefore asked the following question in the survey: “Which word did you last look up in the app (regardless of whether or not it is included in the glossary)?” We are aware of the problems related to this question. First, this is the question with the highest dropout rate. About 200 answers were submitted; of these, about 50 respondents answered “I don’t remember”. Also, respondents may not want to share their lookups with others.

However, it is possible to draw some conclusions from the nearly 150 words (and comments) given by the respondents, especially when the motive is explicitly expressed. The lookups consist of mainly foreign, low-frequency words. A clear majority cannot be considered to belong to basic Swedish vocabulary. The majority of the words in the list are nouns. Some examples are *abderitisk* (‘abderian’), *allegat* (‘voucher’), *befryndad* (‘allied’, ‘kindred’), *chimär* (‘chimera’), *courtage* (‘brokerage’) and *draksådd* (‘a sowing of dragon’s teeth’).

In section 4.2, we discussed the reasons for consulting the app in general. But why did the respondents look up the words specified in the answers? Some respondents went into detail about this in their comments (our translation):

- (1) **cp-skada** (för att se om det skulle vara versaler eller gemener) (‘cerebral palsy injury’, to see if the abbreviation should be written with upper or lower case letters)

(2) **understrecka** (blev osäker på om det skrivs med ä eller e) ('to underscore', was not sure if it is spelled with an 'ä' or 'e')

(3) **Minns inte**, det kan ha varit *hen* (för att kolla objektsformen) (Don't remember. It might have been *hen* (to check the direct object form))

Examples (1) and (2) concern production. Example (3) is about the new gender neutral pronoun *hen* (which has even attracted international attention; see e.g. *The Guardian* 2015-03-24). The motive may have been to see which direct object form (out of two possible ones) is recommended by the Swedish Academy.

4.5 Suggestions for a future version of the app

Yet another purpose of the survey was to obtain information concerning what additional functionality the respondents would like to include in future versions of the app. The diagram in Figure 4 shows the responses.

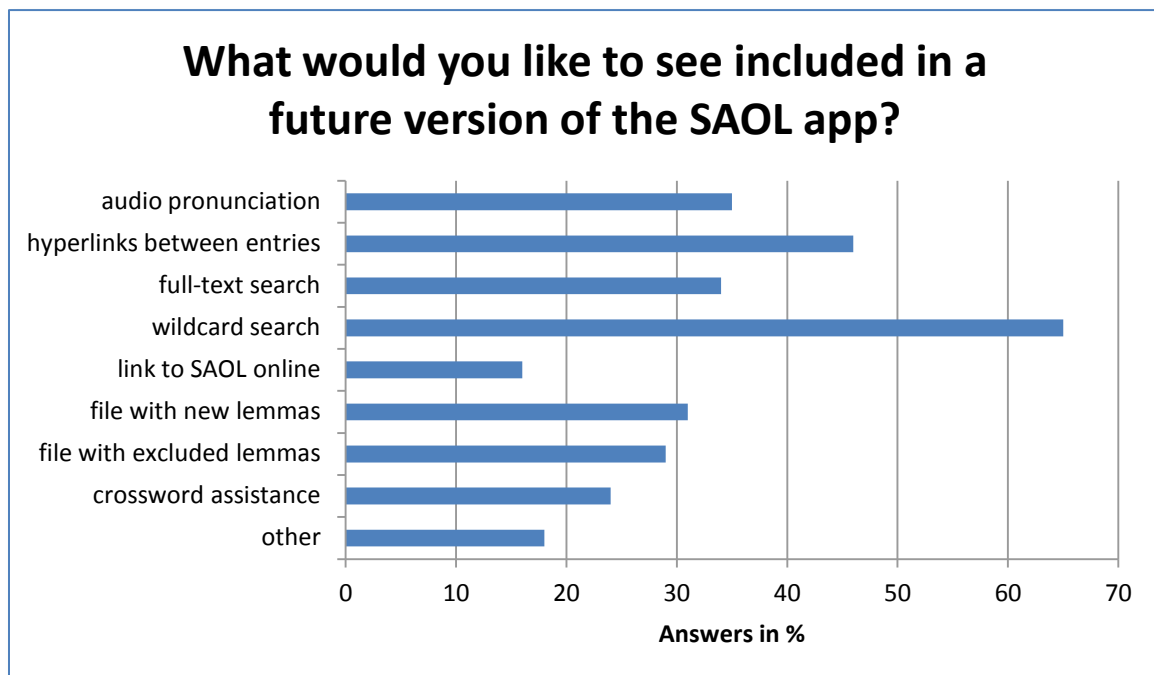


Figure 4: Answers to the question “What would you like to see included in a future version of the SAOL app?” (our translation). (Respondents could select more than one option)

Interestingly, most respondents answered “wildcard search” and “hyperlinks between entries”, with “audio pronunciation” being the third most frequent answer. Both wildcard search and hyperlinks between entries are relatively easy to include in the app considering the digital format and the underlying database structure of the glossary. We clearly should consider this possibility in our future work. Regarding audio pronunciation, at present we have to direct users to the forthcoming dictionary app for the contemporary dictionary of the Swedish Academy, which will include this function.

Those who selected “other” and left a comment suggested improvements on the glossary content rather than on the app functionality. In the app, they requested an improved history function (there is one, but it is evidently hard to find). In the glossary, they suggest definitions, synonyms, etymology and phrasal verbs, etc. According to Malmgren (2014), the 14th edition of the SAOL provides more information on meaning, both explicitly and implicitly. The glossary also includes phrasal verbs as sublemmas. In that sense, the new dictionary content is a solid basis for such improvements in a forthcoming app.

4.6 Pricing

Dictionary sales in Sweden have fallen sharply since the mid-2000s and many publishers have consequently reduced the publishing rate of their dictionaries. Many users now expect linguistic information to be available free of charge (see also Marella 2014: 79). As mentioned, the present SAOL app is free to download, which has had in all probability a substantial impact on the number of downloads. In light of this, it is interesting to see how much the informants are willing to pay for a future version of the glossary app. See Figure 5.

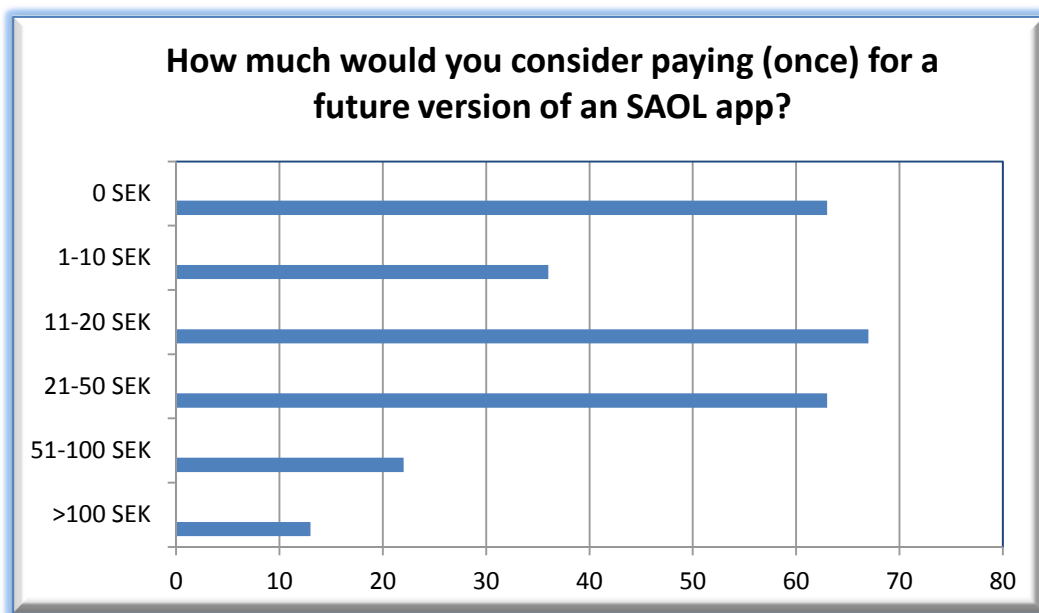


Figure 5: Answers to the question “How much would you consider paying (once) for a future version of the app?” (our translation, 1 SEK = 0.11 EUR)

About 24% say that they are not interested in paying for a new version of the app. Along with those who responded that they are willing to pay the nominal sum of a maximum of 10 Swedish kronor (1.10 Euros), this group constitutes 38% of the respondents. As shown in Figure 5, 25% are willing to pay between 11–20 Swedish kronor. Nearly 5% would be willing to pay more than 100 Swedish kronor, i.e. ca. 11 Euros, which is a hefty sum in the context of apps.

Combining the answers above with the age groups of respondents reveals some correlations. Older respondents appear more willing to pay than younger ones. Respondents aged 40–49 years old are the most willing to pay for an enhanced app. There are also some correlations between user satisfaction and willingness to pay. The more satisfied users are, the more willing they are to pay – but only up to a certain amount (50 Swedish kronor). However, many respondents still request that the app should be free.

4.7 Highly pertinent comments

The various comments offer a wide spectrum of views upon the app from the lexicographical and technical perspectives, on the SAOL as a whole and on dictionary use in a broad sense. The opinions on the app include, for example (our translation):

- (4) “I work with language and I am willing to pay quite a lot for the app – it is amazing! But my students would turn to Google if it started to cost money.”
- (5) “[...] The ‘online version’ available today is not very good; it has to be adjusted more to the web. If the webpage or the web-based SAOL service had a responsive design, it wouldn’t matter if you used it on the computer or your smart phone.”

Considering external links, some of the respondents requested links to other dictionaries, a function that is now included in the dictionary apps published by the Society for Danish Language and Literature (cf. Holmer & Sköldbberg, 2014):

- (6) “It would be fun with a link to the entry in SAOB [The historical dictionary of the Swedish Academy, 1893–], for the words included in that dictionary.”

And, for the SAOL as a whole, we received many comments:

- (7) “I would like more definitions or synonyms for more of the entries.”
- (8) “Both the print book, the app and the web page have their pros, respectively.”
- (9) “My students use the SAOL mainly to look up inflection. They would benefit from more synonyms and hyperlinks between different parts of speech from the same field, for example *thieve* – *thief* – *theft*.”

The overall comments also reveal that there is frustration among online users since the online version is not a database but only a facsimile version of the book. Some app users, such as in example (5), would use the online version if it offered better search options (compared to the now existing facsimile). They seem to use the app as a substitute.

5. Concluding remarks

This article presents the design and results of a web survey regarding the use of the app version of the SAOL, the *Swedish Academy Glossary*, which provides the (unofficial) norm for spelling and inflection of contemporary Swedish words. The survey was directed at people who use the app on a regular basis and consisted of 24 questions covering app usage, design and layout, suggestions for a forthcoming version and respondents' background information. Altogether 264 questionnaires were submitted. Many respondents took advantage of the numerous opportunities to add comments, which resulted in a great deal of highly useful feedback about the SAOL in general and on specific app issues.

The study shows that a clear majority of respondents (about 75%) use the app when they write a text. But as many as about 35% of respondents consult the app also when reading. The respondents are particularly interested in three information categories: spelling, meaning and inflection. In general, their searches consist mainly of foreign, low-frequency nouns. Regarding typical locations for using the dictionary app, few respondents (about 16%) answered that they use the app while on the move. Almost the same percentage of users responded that they consult the SAOL app in cafés, restaurants, etc. The clear majority of entry lookups take place at home or at work.

The results from the survey are of great importance, for example in planning the app version of the recently published 14th edition of the SAOL. It has already been decided (by the Swedish Academy) that the statistical tool Flurry Analytics (see section 2) will be running in the new version, and the editorial staff hope to gain even deeper insights into glossary users and app performance through the use of this new tool. However, the implementation of the Flurry Analytics tool will not eliminate the need for surveys. Surveys may still provide data that are not possible to obtain via statistical tools.

Taking the future of the SAOL app into consideration – as well as that of Swedish dictionary apps in general – knowledge of user willingness to purchase future versions of the app is important. Even though the majority of respondents, in one way or another, use the app in connection with their work, relatively few are willing to pay – and those who are, do not wish to pay much. The unwillingness to pay for dictionary apps and online versions of dictionaries among (Swedish) users has had serious consequences for dictionary publishers in Sweden. This, we believe, mirrors an almost global development concerning traditional dictionaries. Dictionary projects (including app development) are costly and from our professional stance we find it reasonable for users to pay, at least a nominal sum, for these resources. However, convincing users of this is a true challenge, at least in Sweden.

6. References

- Berg, S., Holmer, L. & Hult, A.-K. (2008). SAOL Plus – a new Swedish electronic dictionary. In E. Bernal & J. DeCesaris (eds.), *Proceedings of the XIII Euralex International Congress, Barcelona 15–19 July 2008*. (Institut universitari de lingüística aplicada, Sèrie activitats 20). Barcelona, pp. 1421–1432.
- Berg, S., Holmer, L. & Sköldbberg, E. (2010). Time to say goodbye? On the exclusion of solid compounds from the Swedish Academy Glossary (SAOL). In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, Leeuwarden 6-10 July 2010*. Ljouwert, pp. 567–576.
- Bergenholtz, H. & Johnsen, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. *Hermes – Journal of Linguistics* 34, pp. 117–141.
- Dagens Nyheter* (2015-01-18). Sveriges officiella statistik hotar att bli missvisande. Accessed at: <http://www.dn.se/nyheter/sverige/sveriges-officiella-statistik-hotar-att-bli-missvisande/> (23 May 2015)
- Flurry Analytics. Accessed at: <http://www.flurry.com/>. (23 May 2015)
- Gao, Y. (2013). The Appification of Dictionaries: From a Chinese Perspective. In I. Kosem et al. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*. Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 213–224.
- Hoel, J. (2012). Appsolutt fingerferdig! En anmeldelse av ordbokappene RO og SAOL. *LexicoNordica* 19, pp. 255–271.
- Holmer, L., von Martens, M. & Sköldbberg, E. (2015). Making a dictionary app from a lexical database: a case of the Contemporary Dictionary of the Swedish Academy. Proceedings of eLex conference 11–13 Aug. 2015.
- Holmer, L. & Sköldbberg, E. (2014). Appifiering till allas lycka? Om danska ordboksappar med särskilt fokus på DDO. *LexicoNordica* 21, pp. 235–252.
- Holmer, L. & Sköldbberg, E. (in press). Synonymer.se i fokus – om användningen av en svensk ordbokssajt. In *Svenskans beskrivning* 34. Lund.
- Hult, A.-K. (2012). Old and New User Study Methods Combined Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. In R. Vatvedt Fjeld & J. Mathilde Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress, 7–11 August, 2012*, Oslo, pp. 922–928.
- Lew, R. (2011). Studies in dictionary use: recent developments. *International Journal of Lexicography*, 24 (1), pp. 1–4.
- Lew, R. (in press). Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries. In P. Bański & B. Wójtowicz (eds.) *Issues in Modern Lexicography*. München: Lincom Europa.
- Lorentzen, H. & Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have? In R. Vatvedt Fjeld & J. Mathilde Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress, 7–11 August, 2012*, Oslo, pp. 654–660.

- Malmgren, S.-G. (2014). Svenska Akademiens ordlista genom 140år: mot fjortonde upplagan. *LexicoNordica* 21, 81–98.
- Marello, C. (2014). Using Mobile Bilingual Dictionaries in an EFL Class. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014*. Bolzano/Bozen, pp. 63–83.
- Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2012). Online dictionary use: Key findings from an empirical research project. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 425–457.
- Rundell, M. (2013). Redefining the dictionary: From print to digital. In *Kernerman Dictionary News* 21. Accessed at: <http://kictionaries.com/kdn/kdn21.pdf> (23 May 2015).
- SAOB: *Svenska Akademiens ordbok*. 1893–. Lund: Gleerups.
- SAOL13: *Svenska Akademiens ordlista*. (2006). 13th edition. Stockholm: Norstedts.
- SAOL14: *Svenska Akademiens ordlista* (2015). 14th edition. Stockholm: Norstedts.
- Simonsen, H. Køhler (2014a). Brugerne er allerede mobile! In R. Vatvedt Fjeld & M. Hovdenak (eds.): *Nordiske studier i leksikografi* 12. Oslo: Novus, pp. 416–429.
- Simonsen, H. Køhler (2014b). Mobile Lexicography: A Survey of the Mobile User Situation. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014*. Bolzano/Bozen, pp. 249–261.
- Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- Tarp, S. (2008). Kan brugerundersøgelser overhovedet afdække brugernes leksikografiske behov? *LexicoNordica* 15, pp. 5–32.
- The Guardian* (15-03-24). Sweden adds gender neutral pronoun to dictionary. Accessed at: <http://www.theguardian.com/world/2015/mar/24/sweden-adds-gender-neutral-pronoun-to-dictionary>. (23 May 2015)

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



A multilingual trilogy:

Developing three multi-language lexicographic datasets

Ilan Kernerman

K Dictionaries, 8 Nahum Hanavi Street, 63503 Tel Aviv

E-mail: ilan@kdictionaries.com

Abstract

This paper offers a brief overview of three multilingual developments by K Dictionaries and highlights the main editorial procedures involved and technical tools applied. The first regards an English multilingual dictionary bringing together 43 language versions of Password semi-bilingual dictionary. The second stems from the first, semi-automatically generating multilingual glossaries for any one of those languages to all others via detailed bilingual L2-English indexes. The third is part of the Global series and consists of monolingual datasets for over 20 languages that serve to create various bilingual and multilingual versions and multi-layered combinations. Further steps are anticipated in order to interlink and unify the different resources and processes, such as by associating translations in one lexicographic set to corresponding entries in others and thereby to more translations in other languages, and to converting the data to RDF format for interoperability with Linked Data and Semantic Web technologies.

Keywords: multilingual; dictionary; dataset; semi-automatic generation; linked data

1. Introduction

Multilingual linguistic resources are becoming exceedingly available, diversified and richly generated and used. Applying smart tools to their development and dissemination improves their quality and forms of usage, and increases their accessibility and popularity in a world opening up to cross-linking ever more languages. K Dictionaries (KD) first became involved in multi-language lexicography at the turn of the century with an English multilingual dictionary (EMD) project, and in recent years we have gone deeper into creating resources multilingually. This paper overviews three of our recent multilingual dictionary/lexicography processes, two of which are interrelated, and prospects for enhancing their interoperability both internally and externally for better technological application. First attempts to interconnect the KD data to Linked Data and integrate with Semantic Web technologies were undertaken last year, and more steps will include further multilingual adjustment of the different layers, resources and processes.

2. English multilingual dictionary

The first version of an EMD that assembles a number of semi-bilingual dictionaries for

learners of English was initiated in 2000 by Kielikone, a language technology company from Finland with experience in electronic dictionaries since the late 1980's (Herpiö, 2001). They used 20 language versions of Password dictionary, 18 of them sharing one common English core (based on *Chambers Concise Usage Dictionary*, CCUD) and two based on another (*Harrap's Easy English Dictionary*, HEED)¹, to publish GlobalDix as part of their MOT Dictionary Shelf and as a stand-alone product on CD-ROM and online, including platforms for Windows, Mac, Unix and Linux, intranet and mobile phone².

The semi-bilingual dictionary was launched by Kernerman Publishing in Israel in the mid-1980s for non-native learners of English and was later also known as the *bilingualized* dictionary (cf. Reif, 1987; Kernerman, L., 1994; Nakamoto, 1994). Its main innovation was to use the core of an English monolingual learner's dictionary with the addition of brief translation equivalents in the learner's native language for each sense of the entry. The first edition published for speakers of Hebrew (*Oxford Student's Dictionary for Hebrew Speakers*, 1986) was based on *Oxford Student's Dictionary of Current English* (1978), and the second for speakers of Arabic was based on HEED (*Harrap's English Dictionary for Speakers of Arabic*, 1987), which also served as a base for a few more languages. However, most semi-bilingual versions that followed in cooperation with local publishers worldwide were based on CCUD.

The beauty of GlobalDix was to present side by side translation equivalents for each specific sense of an English word or phrase (including definition and example) from semi-bilingual dictionaries for different languages, enabling the user to compare languages indirectly through the English intermediary. It thus served as a hybrid link for bilingual and multilingual matching, yet lacked full harmony among all the languages because of its reliance on two separate English layers. Another drawback was that while users could look up words in any of the languages, this search was restricted to the list of translations rather than to having a decent headword list for any of the languages.

Over the years KD proceeded to add new language versions to the EMD dataset, unified the English core around a single (CCUD-updated) base for all the language translations, introduced word-to-word reverse indexes for many of the languages to English and combined morphological links for English and certain languages (thus enhancing their searchability), and also upgraded the XML structure overall. The data has since been used in multiple forms and formats by different publishing partners worldwide, such as online dictionaries offering multi-language translations to English

¹ Chinese Simplified, Dutch, Finnish (WSOY), French, German, Hungarian, Icelandic (EDDA), Italian*, Japanese, Latvian (Zvaigzne), Lithuanian (Alma Littera), Norwegian (Aschehoug), Polish, Portuguese Brazil (Martins Fontes), Portuguese Portugal, Russian (Russky Yazik), Slovak (SPN), Spanish*, Swedish (Studentlitteratur), Turkish; language versions marked * are based on HEED, and all others on CCUD.

² cf. The 21-Language GlobalDix. *Kernerman Dictionary News*, 10, p. 3. (2002).

native speakers and foreign users (Dictionary.com, TheFreeDictionary) or semi-trilingual mobile apps including Korean and one more language equivalent to the English lemma for Korean speakers and foreign users (Daol), etc. Figure 1 presents an extract of an entry from a draft online 42-language version.

away  [ə'weɪ]

◆ adverb

1 to or at a distance from the person speaking or the person or thing spoken about: *He lives three miles away (from the town); Go away!; Take it away!*

Afrikaans weg	Greek μακριά, σε απόσταση	Portuguese Portugal longe
Arabic بعيداً، في مكان بعيد	Hebrew רָחוֹק מ־	Romanian departe
Bulgarian разстояние	Hindi दूर	Russian на таком-то расстоянии; далеко
Catalan a; lluny	Hungarian el, messzire	Serbian odavde
Chinese Simplified 离	Icelandic þurt	Slovak odtiaľ; preč
Chinese Traditional 離	Indonesian jauh	Slovenian stran
Croatian dalje	Italian di distanza; via	Spanish a; lejos
Czech daleko; pryč	Japanese 離れた	Swedish från, ort[ifrån], i väg, undan
Danish væk; bort(e)	Korean 멀어지	Thai ไปทิศทางอื่น
Dutch weg	Latvian prolām	Turkish uzakta, uzağa, başka yere
Estonian eemal(e), ära	Lithuanian tolį, šalin	Ukrainian віддалік; геть
Farsi دور؛ دور از	Malay jauh	Urdu دور، دور
Finnish poissa	Norwegian bort(e)	Vietnamese xa
French (au) loin	Polish stad	
German weg	Portuguese Brazil longe	

2 in the opposite direction: *She turned away so that he would not see her tears.*

Afrikaans weg	Greek προς την αντίθετητεύθυνση	Portuguese Portugal de costas
Arabic إلى الجهة المعاكسة	Hebrew לְיָחוּד אֶחָד	Romanian în altă parte
Bulgarian в друга посока	Hindi विपरीत दिशा	Russian в сторону
Catalan cap a l'altre costat, en una altra direcció	Hungarian el(fordul)	Serbian u suprotnom pravcu
Chinese Simplified 掉转	Icelandic í burtu, undan	Slovak stranou, nabok
Chinese Traditional 掉轉	Indonesian pada arah yang berlawanan	Slovenian vstran
Croatian od	Italian dall'altra parte	Spanish hacia el otro lado, para otra parte
Czech stranou	Japanese あちらへ	Swedish bort, om
Danish væk; den anden vej	Korean 다른 방향으로	Thai ไปทิศทางตรงกันข้าม
Dutch weg	Latvian prom; prolām	Turkish başka yöne, başka tarafa
Estonian kõrvale	Lithuanian į šalį	Ukrainian геть
Farsi به سوی دیگر	Malay ke arah lain	Urdu مخالف سمت پر
Finnish vastakkaiseen suuntaan	Norwegian bort, vekk, unna	Vietnamese tránh xa
French de l'autre côté	Polish w drugą stronę	
German weg	Portuguese Brazil de costas	

Figure 1: Extract of an entry from a draft online 42-language version of the EMD

In 2013–2014 KD has undertaken a new round of thorough editorial revision and update of the (CCUD-based) English dictionary core, pursued by the translation of over 2,000 new entries in most of the language versions available then. The ensuing new EMD dataset currently contains a total of approximately 1.7 million translations in 43 languages, referring to 30,000 English entries (i.e. words and phrases) that include 39,000 senses with 38,000 examples of usage.

3. Multilingual glossaries

The EMD revision was succeeded since the end of 2014 by the development of newly refined reverse L2-English indexes that became the base for multilingual glossaries³. In the past, such indexes consisted simply of word-to-word lists, some including the part of speech of the L2 headword. The headwords were derived from the list of translations in the original semi-bilingual English dictionary for the particular L2, and were manually revised to keep, adjust or remove any item and to edit its matching English headword-turned-into-translation. The new indexes, however, were conceived to link the L2 headword precisely to each specific corresponding sense in polysemous entries of the original English dictionary core, rather than to the English headword, and finally list these English equivalents according to frequency and importance rather than in alphabetical order. Consequently, once a new L2-English index is ready it can be automatically turned into a multilingual glossary by associating the translations in all other languages for each sense of the English entry (now a translation). In this way, if N reverse indexes are made then $N*N-1$ new connections can be obtained. The following three simple steps can serve to portray the general process:

1. Have EN>DE, EN>ES, EN>FR, EN>RU (etc.)
2. Add FR>EN
3. Obtain FR>EN>DE, FR>EN>ES, FR>EN>RU (etc.)

The raw index is produced by automatic processing of the original English-L2 data, a process that incorporates some basic rules meant to help manipulate more complex data, for example pertaining to headwords and translations that happen to have variations (particularly regarding punctuation marks e.g. slash, brackets, comma).

Technically, the program first parses the EMD's XML files and creates basic tables. It searches all the Translation containers and compounds and associates each one with its Sense. The Sense set includes the following components:

- Translations for all the languages
- Definition
- Example(s) of usage

³ The languages indexed and multilingualized so far include Catalan, Chinese Simplified, Danish, Dutch, Estonian, French, German, Hungarian, Indonesian, Italian, Japanese, Polish, Portuguese Brazil, Portuguese Portugal, Russian, Slovene, Spanish and Swedish.

- Headword and part of speech

The outcome of the initial parsing is illustrated in Figure 2:

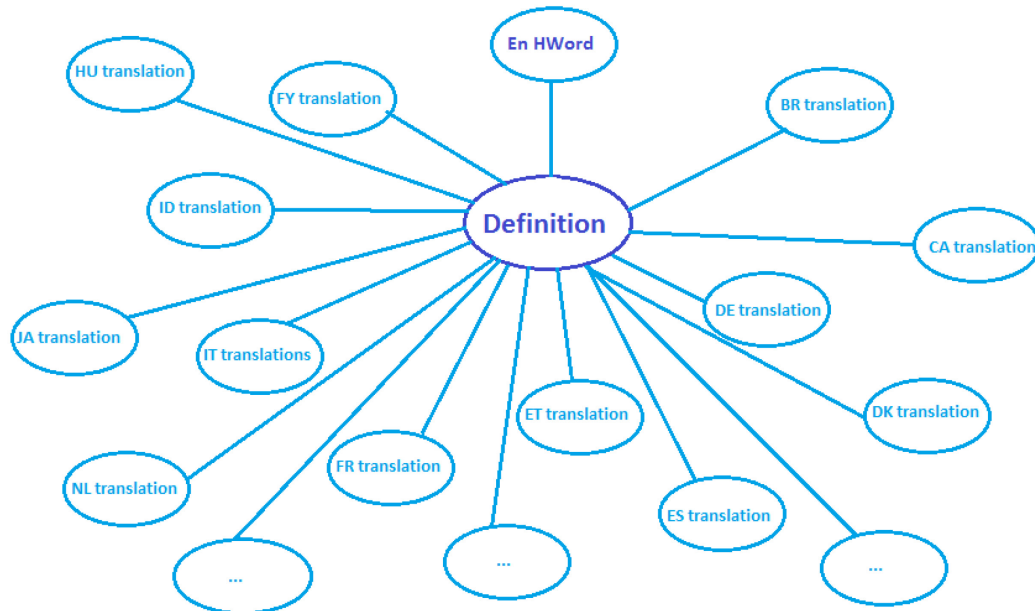


Figure 2: Parsing the XML data and preparing translations in different languages

The main characteristics of the Sense set consist of the Definition and the associated L2 Translation. Each Sense has an identifier, which will serve to generate the multilingual glossary. The software also generates translation tables for all the languages, which will eventually serve the multilingualization process.

At this preliminary stage, the program can generate the raw L2-English index. First, it creates a temporary L2 index by parsing the Translations from the EMD and building a table that includes the following components:

- L2 Translation
- Part of Speech
- English Headword
- (English) Definition
- (English) Example of usage (if appropriate)

As a result, the L2 Translation (from EMD) becomes an L2 Headword. Now the program brings together all the Senses in the EMD that were associated with it as a Translation and lists them alphabetically (according to the original English Headword and Sense number). Subsequently, the L2 Headword is composed as follows:

- Sense set 1
 - English Headword 1
 - Part of speech 1
 - Definition 1
 - Example of usage 1
- Sense set 2
 - English Headword 2
 - Part of speech 2
 - Definition 2
 - Example of usage 2
- Etc.

The ensuing raw index then undergoes thorough manual editing, using an especially dedicated software tool. In general, the editor reviews the L2 translations-turned-into-headwords to decide which items to keep intact, change into appropriate headwords or remove if not relevant, and adjusts their automatically allocated parts of speech. As for the English translation equivalents, the editor removes inappropriate ones and adds others, as well as rearranging them according to frequency and importance⁴. Figures 3 and 4 present sample screenshots of editing the index using this special tool.

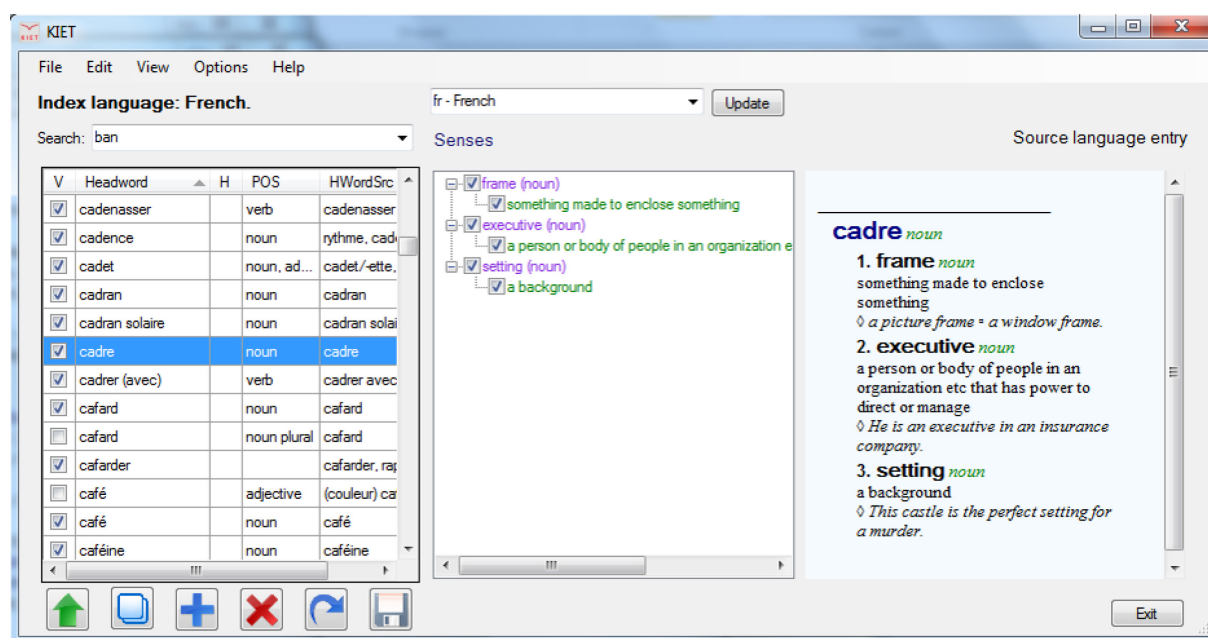


Figure 3: Editing the French Headwords in the Index Editorial Tool

⁴ A detailed account of this editorial process is available in Egorova (2015) in this volume.

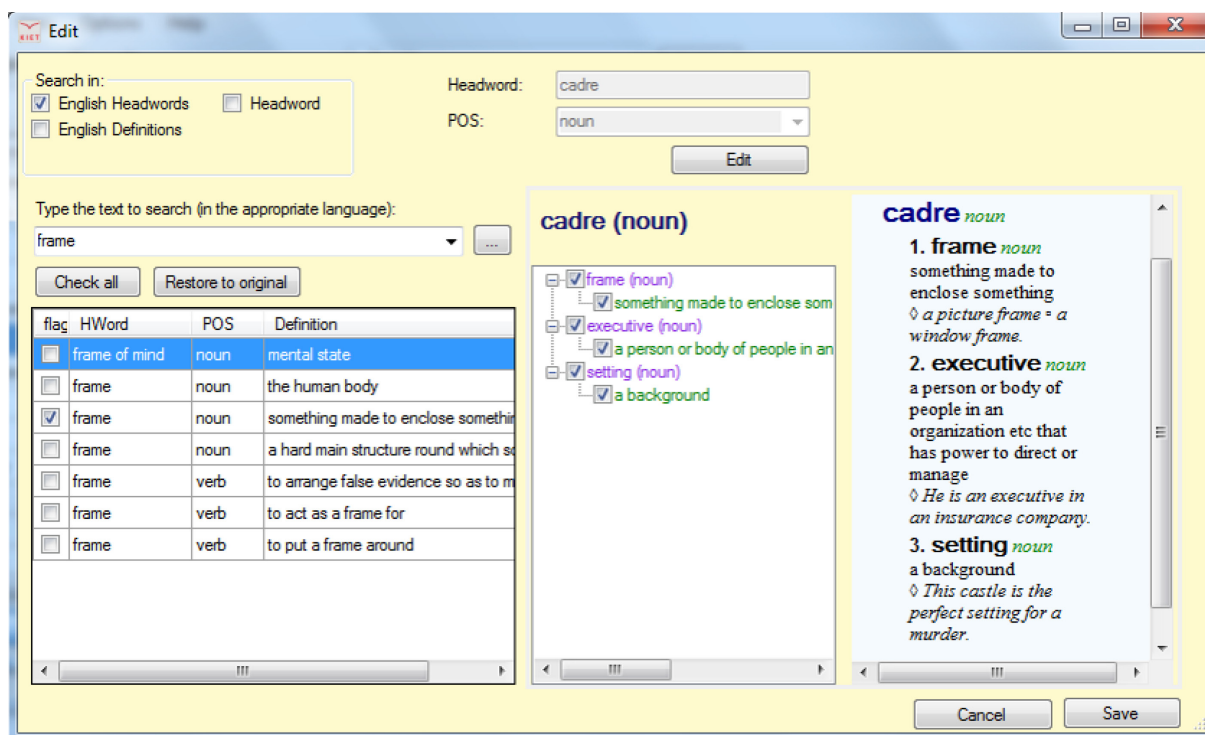


Figure 4: Editing the English Sense equivalents in the Index Editorial Tool

The detailed editing of the English translations according to each specifically matching sense, rather than just suiting the corresponding headword, offers a reasonable base to automatically produce fair-quality multilingual glossaries by adding the translations into all other languages from the EMD. Figures 5 and 6 present two samples of the results, the first featuring the English translation/sense with the other language translations derived from it, and the second integrating the English equivalent together with all other languages without exposing its fundamental linking role.

messen *verb*

1. **gauge** to measure (something) very accurately
2. **measure** to find the size, amount etc of (sth)
3. **measure** to show the size, amount etc of
4. **measure** (with **against**, **besides** etc) to judge in comparison with
5. **measure** to be a certain size
6. **meter** to measure (*especially* electricity etc) by using a meter
7. **take** to make a note, record etc

Figure 5: German multilingual entry exposing its primary English equivalent link

messen verb

1. to measure (something) very accurately

af meet | ar يقيس | bg измервам точно | br medir | ca mesurar, calibrar | cs (z)měřit | dk måle | el (κατα)μετρώ με ακρίβεια | en gauge | es medir, calibrar | et mõõtmä | fa یا دفت اندازه گیری کردن | fi mitata | fr mesurer, jauger | he תימדל | hi प्रमाप, आयाम | hr mjeriti | hu megmér | id mengukur | is mæla | it calcolare | ja 測る | ko 정확히 측정하다 | lt matuoti | lv mērīt | ml mengukur | nl meten | no måle (opp) | pl wymierzyć | pt medir | ro a măsură | ru измерять | sk odmerať | sl izmeriti | sr izmeriti | sv mäta | th วัดด้วยมาตรวัด; เครื่องวัด | tr ölçmek | tw 精確測量 | uk виміряти | ur کسی چیز کو ناپنا | vi đo | zh 精确测量

2. to find the size, amount etc of (something)

af meet | ar يقيس | bg измервам | br medir | ca mesurar | cs (z)měřit | dk måle | el μετρώ | en measure | es medir | et mõõtmä | fa اندازه گیری کردن | fi mitata | fr mesurer | he תימדל | hi नापना | hr mjeriti | hu (meg)mér | id mengukur | is mæla | it misurare | ja 測る | ko 치수를 재다 | lt (iš)matuoti | lv no | ml mengukur | nl meten | no måle, ta mål av | pl (wy)mierzyć | pt medir | ro a măsură | ru измерять | sk odmerať | sl izmeriti | sr izmeriti | sv mäta | th วัดขนาด (ความยาว, ความสูง, ความเร็ว ฯลฯ) | tr ölçmek | tw 測量 | uk міряти, вимірювати | ur حجم و غیرہ معلوم کرنا | vi đo lường | zh 測量

Figure 6: German multilingual entry combining its primary English equivalent link with all the other language equivalents

Unfortunately, these automatically-generated multilingual glossaries are bound to contain inaccuracies due to the indirect nature of juxtaposing different languages via the English common ground. Nevertheless, they offer some merit for basic translation purposes and serve as an advanced base for amending higher quality matching, useful in particular for less-common language pairs. At this stage, there is no information about the precise rates of the “inaccuracies” in the L2–L3 automatic matching, and this remains to be further investigated.

4. Fully multilingual dictionaries

In 2005 KD began to create the Global series, with the first multilingual combinations becoming available since 2009⁵. The Global series has its foundation in monolingual

⁵ KD's BLDS: A brief introduction. *Kernerman Dictionary News*, 17, pp. 1–2 (2009).

lexicographic datasets for different languages (Kernerman, I., 2011)⁶, each serving as a base for adding translations and developing bilingual dictionaries. Thus, whenever one of the core languages has several bilingual versions, putting their data together produces a multilingual dictionary. This process is similar in principle to that of composing the EMD. However, the Global entry microstructure is much more elaborate and allows for more than one translation equivalent per sense, as compared to usually just a single translation per language in the EMD. In addition, the examples of usage are translated as well, unlike the EMD's semi-bilingual base that has translations only for the meanings of the word or phrase. These differences lead to significantly richer results. Moreover, since the languages that consist of translations exist also as L1 cores in the Global series, many of the translations can be associated to their full entries and the information provided can be (re-)expanded again and again. Figures 7, 8 and 9 display French monolingual, bilingual and multilingual entries, respectively.

cité [site] *nf* 1 <ville> grande ville
 ◇ *une cité industrielle*
 ◇ *créer de toute pièce une nouvelle cité*
 2 <quartier> ensemble d'immeubles
 ◇ *vivre dans une cité*
 ◇ *La cité ouvrière est très animée.*
 3 ♦ **cité universitaire** ensemble de logements pour étudiants

Figure 7: Global French monolingual entry

cité [site] *nf* 1 <ville> grande ville
 {br} - **cidade** [si'dadʒi] *f*
 ◇ *une cité industrielle*
 {br} - **uma cidade industrial**
 2 ensemble d'immeubles
 {br} - **bairro** ['barxu] *m*
 ◇ *vivre dans une cité*
 {br} - **viver num bairro**
 3 ♦ **cité universitaire** ensemble de logements pour étudiants
 {br} - **cidade universitária**

Figure 8: Global French bilingual entry (French–Portuguese)

⁶ Global series language cores available so far include Arabic, Chinese Simplified, Chinese Traditional, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Italian, Japanese, Korean, Latin, Norwegian, Polish, Portuguese Brazil, Portuguese Portugal, Russian, Spanish, Swedish, Thai and Turkish.

cité [site] <i>nf</i> 1 <ville> grande ville	
{ja} - 都市 (とし) <i>toshi</i>	{es} - <i>ciudad</i> <i>f</i> [θjuˈðað]
{ru} - го́род [ˈgorət] <i>m</i>	{it} - <i>metropoli</i> [meˈtrɔpɔli] <i>f</i>
{ar} - مدينة [maˈdiːna] <i>f</i>	{nl} - <i>stad</i> <i>m/f</i>
{el} - πόλη [ˈpɔli] <i>f</i>	{de} - (Groß)Stadt <i>f</i>
{pl} - <i>miasto</i> [mjastɔ] <i>nt</i>	{no} - <i>by</i> <i>m</i>
{pt} - <i>cidade</i> [siˈdadə] <i>f</i>	{br} - <i>cidade</i> [siˈdadɔ] <i>f</i>
{tr} - <i>kent</i> [cent]	{he} - עיר [ʔir] <i>m</i> (<i>pl</i> ערים ערים)
{zh} - 城市, 都市, 城邦 <i>chéngshì</i> , <i>dūshi</i> , <i>chéngbāng</i>	{a'rim} - קרייה קרייה [kirˈja] <i>f</i>
◇ <i>une cité industrielle</i>	{sv} - <i>stad</i> <i>common</i>
{ja} - 工業 (こうぎょう) 都市 <i>koogyoo toshi</i>	{zh} - 一个工业城市 <i>yí gè gōngyè</i> <i>chéngshì</i>
{ru} - промышле́нный го́род	{nl} - <i>een industriestad</i>
{ar} - مدينة صناعية	{de} - <i>eine Industriestadt</i>
{el} - βιομηχανική πόλη	{no} - <i>en industriby</i>
{pl} - <i>miasto przemysłowe</i>	{br} - <i>uma cidade industrial</i>
{pt} - <i>uma cidade industrial</i>	{he} - עיר תעשייתית
{tr} - <i>sanayi kenti</i>	{sv} - <i>en industristad</i>

Figure 9: Global French multilingual entry

5. Further developments

In 2014 KD had a first taste of converting its data from XML (Extended Markup Language) to RDF (Resource Description Framework)⁷ format, based on the Lexicon Model for Ontologies (*lemon*)⁸, through academic cooperation at Madrid Polytechnic University and Leipzig University (Klimek & Brümmer, 2015). RDF is a data model developed by the World Wide Web Consortium (W3C), serving as the basic mechanism to formally describe any type of *resource* – whether words, documents, people, physical objects or abstract concepts – along a *subject-object-predicate* pattern and thus making it more easily sharable and interconnectable (Gracia, 2015). The RDF transformation is a vital step in uniformizing our lexicographic datasets into a common structure in order to facilitate cross-linking content from different dictionaries, enriching it by exterior multi-language lexical and other resources, and having it published as Linked Data on the Web.

The processes described in sections 2 and 3 already constitute attempts to link our internal resources to each other, and thereby expand them exponentially, and the same can be said about the fairly simple and straightforward process described in section 4. Next challenges consist of linking the various Global language core resources to each other – such as by linking an L2 translation to the information it has as an (L1) entry in its own monolingual set and on to translations in L3, L4, etc. – and to other internal resources such as the EMD and multilingual glossaries. For example, the Portuguese translation in Figure 7 could be linked to that lemma which exists as a headword in

⁷ Resource Description Framework, cf. <http://www.w3.org/rdf>

⁸ <http://www.lemon-model.net>

the Portuguese core with its translations to another language, and so on and so forth. Likewise, the same item could be linked (also) to the Portuguese translation in the EMD and to the multilingual information it has as part of the Portuguese glossary. This development can be defined as moving on from *multilingual* to *multilayer*, in the sense that each language part in any of the lexicographic datasets constitutes one layer of information and that these different layers are interconnected, as part of further expansion of these multi-language opportunities.

Whereas the internal process described above could suffice with keeping the data in XML format and is just enhanced by its RDFication, linking with other resources on the Web relies exclusively on the RDF format. For example, the data could then be enriched by open resources such as WordNet, Wiktionary and Babelnet, to name just a few well-known open source lexical websites. KD is starting to develop a new API that will enable such exterior linking, both for extracting new data from other resources and for disseminating its own data more efficiently to others. The data manipulation described in this paper may seem in parts as a revolution with respect to traditional lexicography, but it still only scratches the surface of a new threshold to future prospects.

6. Acknowledgements

I would like to thank the reviewers of this paper for their comments, in particular Reviewer 3 for the detailed remarks.

An earlier version of this paper was presented at the 20th Biennial Dictionary Society of North America Meeting (DSNA-20), held at the University of British Columbia in Vancouver (BC), on 6 June 2015.

7. References

- Egorova, K. (2015). Editing an automatically-generated Russian-English index. In I. Kosem et al. (eds.) *Electronic Lexicography in the 21st Century: Linking lexical data in the digital age. Proceedings of eLex 2015, Herstmonceux Castle, 11-13 August 2015. Herstmonceux Castle, UK.* Available at: <https://elex.link/elex2015/>.
- Gracia, J. (2015). Multilingual dictionaries and the Web of Data. *Kernerman Dictionary News*, 23, pp. 1–4.
- Herpiö, M. (2001). *GlobalDix*: A unique multilingual dictionary for the worldwide market. *Kernerman Dictionary News*, 9, p. 12.
- Kernerman, I. (2011). From dictionaries to databases: Developing a global series for language learners. In I. Kosem & K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011. Bled, Slovenia.* Available at: <http://elex2011.trojina.si/Vsebine/proceedings/eLex2011-0.pdf>.

- Kernerman, L. (1994). The advent of the semi-bilingual dictionary. *Password News*, 1, p. 1.
- Klimek, B. & Brümmer, M. (2015). Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23, pp. 5–10.
- Nakamoto, K. (1994). Monolingual or bilingual, that is *not* the question: The ,bilingualised‘ dictionary. *Lexicon*, 24. Tokyo: Iwasaki Linguistic Circle (Kenkyusha).
- Reif, J.A. (1987). The development of a dictionary concept: An English learner’s dictionary and an exotic alphabet. In A. Cowie (ed.) *The Dictionary and the Language Learner: Papers from the Euralex Seminar at the University of Leeds, 1-3 April 1986*. Lexicographica Series Maior, 17. Tübingen: Max Niemeier Verlag, pp. 140–158.

Websites:

- Babelnet. Accessed at <http://www.babelnet.org>. (12 June 2015)
- Dictionary.com*. Accessed at: <http://www.dictionary.com>. (1 August 2007)
- TheFreeDictionary*. Accessed at: <http://www.thefreedictionary.com>. (12 March 2009)
- W3C. World Wide Web Consortium. Accessed at: <http://www.w3.org>. (12 June 2015)
- Wiktionary. Accessed at: <http://www.wiktionary.org>. (12 June 2015)
- Wordnet. Accessed at: <http://www.wordnet.princeton.edu>. (12 June 2015)

Dictionaries:

- Daol. *Kernerman Semi-Trilingual English Korean L3 Dictionaries*. (2013). Seoul: DaolSoft.
- CCUD. *Chambers Concise Usage Dictionary*. (1986). Edinburgh: W&R Chambers.
- GlobalDix. *GlobalDix*. (2001). Helsinki: Kielikone.
- Harrap’s English Dictionary for Speakers of Arabic*. (1987). Toronto: Kernerman Publishing.
- HEED. *Harrap’s Easy English Dictionary*. (1980). London: Harrap.
- Oxford Student’s Dictionary for Hebrew Speakers*. (1986). Tel Aviv: Kernerman Publishing.
- Oxford Student’s Dictionary of Current English*. (1978). Oxford: Oxford University Press.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Multiple Access Paths for Digital Collections of Lexicographic Paper Slips

Toma Tasovac¹, Snežana Petrović²

¹ Belgrade Center for Digital Humanities

² Institute of Serbian Language of the Serbian Academy of Arts and Sciences

E-mail: ttasovac@humanistika.org, snezzanaa@gmail.com

Abstract

The paper describes the process of digitizing and annotating some 23,000 lexicographic paper slips compiled by the amateur lexicographer Dimitrije Čemerikić (1882-1960) to document the Serbian dialect from the historic city of Prizren. This previously unpublished dictionary of the Prizren dialect is an important resource not only for dialectologists and linguists, but also for ethnolinguists and ethnologists who are interested in various aspects of popular culture and urban life in the city of Prizren. The alphabetic arrangement of the macrostructure, however, is not conducive to exploratory searches: if users want to find out which dialect word corresponds to a standard Serbian word, or explore a certain type of vocabulary, they need access paths to the dictionary content that go beyond the indexing of the macrostructure. The paper describes an elaborate annotation strategy based on marking up headwords with standardized orthographic alternatives, providing lexical equivalents and assigning semantic fields to entries in order to achieve robust navigability and searchability of the collection without full-text transcription and/or structural data modeling.

Keywords: digitization; dialect dictionaries; navigation; searchability; access paths

1. Introduction

Despite the dramatic impact which corpus linguistics has had on contemporary lexicographic practice (Sinclair, 1991; Fellbaum, 2009), the history of lexicography cannot be understood without considering the tradition of lexicographic citation slips — the hand-picked excerpts from literary and other sources that are an essential component of the lexicographer's toolkit (Landau, 1984; Wandl-Vogt, 2005; Bakken, 2006). Collections of lexicographic paper slips are not only an important part of European lexicographic heritage (Considine, 2008), but are research objects in their own right. In this paper, we discuss the process of digitizing and annotating one such collection created by the Serbian amateur lexicographer Dimitrije Čemerikić (1882-1960). Čemerikić's manuscript, compiled in the middle of the twentieth century using some 23,000 paper slips, contains approximately 16,000 lemmas with definitions and examples that illustrate the variant of Serbian from the historic city of Prizren that is today an endangered dialect (Петровић, 2012; Петровић & Тасовац, 2013).

The main goal we set ourselves for the digital edition of the Čemerikić paper slips was to provide users with improved retrieval possibilities based on multiple access points.

We will show how our decision to implement an elaborate annotation strategy based on marking up headwords, standardizing orthography, providing lexical equivalents and indicating the entry's semantic fields enabled robust navigability and searchability without full-text transcription and/or structural data modeling.

The paper is structured as follows: Section 2 describes Čemerikić's manuscript itself in greater detail. Section 3 explains how different methods of digitization (image capture, text capture, data modeling and data enrichment) influence the kinds of access paths that an electronic resource can offer. Section 4 analyzes the need for access paths beyond the dictionary macrostructure, while Section 5 presents in detail how the annotation of the Čemerikić collection has helped us achieve the goal of providing multiple access paths to the collection.

2. The Manuscript

The Čemerikić manuscript is part of the inventory of paper slips collected over a period of almost 100 years for the compilation of the *Речник српскохрватског књижевног и народног говора* (Dictionary of Serbo-Croatian Literary and Vernacular Language) of the Serbian Academy of Arts and Sciences (Ристић et al., 2011). It is an accident of history that this collection has not been merged with the rest of the Academy's inventory, but has instead remained physically separate. While a small portion of its valuable content has trickled through to the first 19 volumes of the Academy dictionary that have been published so far, the manuscript contains sufficient interesting material to deserve a publication on its own.

The original of the Čemerikić manuscript is archived at the Institute for the Serbian Language of the Serbian Academy of Arts and Sciences. The digital version has been publicly available since 2013 via *Prepis.Org: The Platform for the Transcription and Digital Editions of the Serbian Manuscript Heritage* (Тасовац & Петровић, 2013). One small part of the manuscript, dealing with 3,848 entries for words starting with letters а, б and в, has survived in typewritten form on sheets of A4 paper. The bulk of the collection, however, consists of entries written in ink and pencil on paper slips of different sizes and quality, torn-out notebook papers and, in some cases, even cigarette paper¹.

Formally, we can distinguish three types of paper slips: those containing only records of individual word forms (cf. цар, ценом, ептен); those containing only citations (cf. басма шиљте), and those, in the majority, which are already formatted as prototypical dictionary entries with highlighted headwords, grammatical information, definitions, citations etc. Čemerikić used various sources for his work: he excerpted words from various trade records and guild protocols (written in the pre-reform Cyrillic alphabet); ethnographic and historical literature, newspapers, travel literature etc. Most

¹ See, for instance, <http://www.prepis.org/items/show/19315>

importantly, however, the manuscript contains an abundance of examples from colloquial, everyday communication as well as numerous descriptions of local cultural traditions. This previously unpublished dictionary of the Prizren dialect is therefore an important resource not only for dialectologists and linguists, but also for ethnolinguists and ethnologists who are interested, for instance, in various aspects of popular culture (customs, superstitions, witchcraft) and urban life (guilds, social and ethnic relations, etc.) in the city of Prizren (Петровић & Тасовац, 2014). We based our approach to digitizing Čemerikić on the premise that electronic access will benefit both scholars (dialectologists, lexicographers and linguists) and the general public interested in the language and culture of the city of Prizren.²

3. Lexicographic Data: From Paper to Screen

Not all digital objects are created equal. We can distinguish four types of methods and activities for creating digital representations of lexical resources: 1) image capture; 2) text capture; 3) (lexicographic) data modeling and 4) (lexicographic) data enrichment. In this section, we will briefly look at these four aspects and their roles in our digitization of the Čemerikić manuscript.

Image capture refers to the process of recording the visual representation of the text by means of digital cameras and scanners and its subsequent delivery to the user as a digital image. Digital images are nowadays quite easy to produce and deliver over the internet but their usability, especially when it comes to lexicographic material, is limited due to a lack of search capabilities. The process of digitizing the Čemerikić manuscript started with the scanning of some 23,000 paper slips. The digital images were made available via the online platform <http://prepis.org> from the very beginning of the project. Initially, however, the scanned paper slips suffered from some of the same shortcomings as their physical counterparts: identifying and retrieving information about particular words would require browsing hundreds if not thousands of digital images.

Text capture refers to the transposition of textual content into a sequence of alphanumeric characters, which can be accomplished either by human operators who retype the original text; or, automatically, by using an optical character recognition (OCR) software to convert images into searchable strings. Optical Character Recognition (OCR) is widely used in mass digitization efforts, but its application in the realm of recognizing unconstrained hand-written texts is not as successful as it is in cases of printed documents or constrained hand-written domains such as numbers

² We have not conducted specific user surveys with the general public, but our own experience with organizing an exhibition about the Čemerikić manuscript at the Science and Technology Gallery of the Serbian Academy of Arts and Sciences, as well as a previous social media project related to the Serbian Dictionary by Vuk Stefanović Karadžić (1787-1864), which had more than 24,000 followers on Facebook alone, makes us confident that there is a broad interest among the Serbian public for topics related to language history and language diversity.

or postal addresses (Vinciarelli, 2002; Bunke, 2003; Plötz and Fink, 2009). Challenges include low paper quality, ink bleed-thru, line positioning variations (skews), overlapping characters, wide personal variations in glyph formation, and, often, a circular dependency between character segmentation and recognition, sometimes referred to as Sayre's paradox (Sayre, 1973).

Manually transcribing the full-text of Čemerikić's paper slips would be a time-consuming and costly process, not just because of the physical qualities of the slips which have not been preserved under ideal archival conditions, but also because of the nature of the material – a dialect with a large number of nonstandard vocabulary items, multilingual content and even nonstandard Cyrillic graphemes. Even if a team of highly-skilled, linguistically-trained transcribers could perform the job, the full-text transcription would not necessarily be sufficient for the creation of robust search and retrieval possibilities.

Lexicographic data modeling refers to the process of explicitly encoding the structural hierarchies and the scope of particular textual components: in the case of lexicographic data, this usually involves marking up both the macrostructure of the dictionary and the microstructure of individual entries (lemmas, grammatical information, senses etc.) A marked-up text increases the information density of the digital surrogate and paves the way for the implementation of more advanced faceted navigation and targeted search capabilities (for instance, retrieving all nouns whose etymology indicates particular linguistic origins; or retrieving all instances of a particular lexeme when it appears in dictionary examples stemming from a particular author). While it would have been ideal to create, for instance, a TEI-encoded ISO-LMF-compatible edition of the Čemerikić manuscript from the outset of the project, this was not a practical choice. With full-text transcription of the entire manuscript remaining beyond our reach due to financial constraints, the structural modeling was also not an option.

Lexicographic data enrichment, on the other hand, does not necessarily depend on the availability of the full text. By *data enrichment* or *annotation*, we refer to the process of encoding additional information that specifies, extends or improves upon the information already present in the lexicographic resource. As will be seen in Section 5, entry-level lexical and semantic annotations of the digitized paper slips can increase their use value even without transcription and/or structural modeling of the content.

Before we turn to the analysis of the data enrichment of the Čemerikić collection, one other question remains to be addressed: why do we need multiple access paths in the first place?

4. Access paths

The alphabetical arrangement of entries in a print dictionary functions as a type of *index* — a retrieval mechanism connecting a known order of symbols to an unknown

order of information (Hass Weinberg, 2010). The user can access dictionary content by consulting the dictionary macrostructure, i.e. the arrangement of lemmas in a given order (see Hausmann & Wiegand, 1989). While alphabetic dictionaries are relatively easy to consult, they are also efficient randomizers of meaning. By grouping lexemes according to their orthography, rather than their sense, standard dictionaries adhere to the abstract convention of alphabetical order, scattering words with similar or related meaning across unpredictable distances. The “psychologically quite unmotivated tyranny of the alphabet” (Makkai, 1980: 127) is both a blessing and a curse. Looking up entries is easy, if one knows precisely what word one is looking for. Discovering unfamiliar words and exploring semantic concepts, however, is considerably more difficult (Tasovac, 2012).

In electronic dictionaries users access lexicographic content not based on a single wordlist but through a search engine: “it may be more appropriate to say that the macrostructure has been replaced by what may be called a data presentation structure.” (Nielsen, 2011: 201; see also Nielsen & Almind, 2011). The lexicographic concept of accessibility needs to be “narrowed down to cover *quick and easy access* to the specific types of data that can cover a specific type of user’s specific types of need in a specific type of extra-lexicographical situation” (Tarp, 2008: 101). What constitutes *quick and easy access*, however, depends as much on a particular situation of use as it does on the type of the dictionary being accessed.

Users resort to historical dictionaries, for instance, in roughly three types of situations: (1) when they have difficulties in the reception of historical texts, (2) when they have difficulties in the production of modern translations; and (3) when they have general questions about linguistic and cultural tradition (see Reichmann, 2012: 54). The first two types of situations are text-related: they arise out of the user’s engagement with a particular text. The user can, when reading texts, experience all sorts of semantic difficulties (encounter unknown lexical units; discover gaps in word meaning; raise questions of morphological, syntactic or pragmatic nature). In these cases, the user will use the macrostructure (or the search engine, in the case of an e-dictionary) to locate a specific entry containing the information that he or she needs.

Reichmann’s third situation of use is *texttranszendierend* [text transcending] (2012: 64). What this means is that lexicographic texts can also be used to study the lexical materialization of cultural and historical relations, processes and transformations. Dictionaries, after all, are not only information-extraction tools: they also serve as texts, models of language and cultural objects deeply embedded in the historical and ideological matrices of their time (Tasovac, 2010). The main difference between the use of dictionaries in specific text reception and text production situations, on the one hand, and more general research situations on the other hand, is the question of initial focus and ultimate scope. In specific, text-related situations of use, the initial focus and ultimate scope are usually the same: extracting the definition of a particular sense of a particular word is usually accomplished by consulting one dictionary entry. In

text-specific situations, the dictionary is used as a look-up tool. In text-transcending situations, it is used as an exploratory tool.

To make the digital edition of the Čemerikić manuscript available in text-specific situations, the images were first digitized and uploaded to *Prepis.Org: The Platform for the Transcription and Digital Editions of the Serbian Manuscript Heritage*, which uses Omeka, an open-source digital collection management system in its backend (Kucsma et al., 2010; Tomás, 2011). After merging entries that are written on both sides of individual slips or across several paper slips, we arrived at 16,626 entries. The headwords for all entries were then transcribed and a search plugin implemented with an autocomplete dropdown menu, allowing users to gain a view of the scope of the entire entry list.

The screenshot shows the Prepis.Org website interface. At the top, there is a search bar with the text 'звигаревић' and a dropdown menu listing various search results. Below the search bar, the main content area displays the entry for 'ЗЪНДАН'. The entry includes a title, a description, a photograph of a handwritten manuscript slip, a quote, and a list of metadata.

Метаподаци

Наслов:	зъндан
Аутор:	Чемерикић (1882—1960)
Извор:	Збирка реч
Опис:	Листић је д збирке реч Димитрија коју је аутор и 1960. год Институ з српскохрв САНУ као г српскохрв књижевног језика.
Датум:	Прва поло
Издавач:	Центар за хуманисти Институ з САНУ
Носилац ауторских права:	Српска ака уметности
Лиценца:	Creative Commons Ауторство - Некомерцијално 3.0 Србија
Медиј:	image/jpeg
Опсер:	1807 x 789 px

Figure 1: Autocomplete search

The entries are marked in terms of priority for subsequent full-text transcription: priority 1 is given to entries that contain Čemerikić's citations of spoken sources. These are given the highest priority because of the scarcity of spoken dialectological data for the Prizren dialect, especially from the middle of the century. Editors are also given the freedom to mark with priority 1 entries that are particularly interesting from the point of view of cultural history. Priority 2 is given to entries that contain citations from previously published written sources, more often than not from historical literature; and priority 3 to all other entries. By default, all entries are marked with priority 3 and then manually upgraded to levels 1 or 2 where required. As of this

writing, of the 6820 manually prioritized entries, 3261 were given priority 1; 1826 were assigned priority 2; and 1724 remained priority 3. Priority 4 is given to transcribed items, and priority 5 to transcribed entries that have been proofread and approved by the senior editor. Due to financial constraints, only entries with priority 1 are currently being transcribed in full.

Direct access to the macrostructure of the Čemerikić collection, while being a *sine qua non*, would not have been sufficient for a text-transcending, exploratory use. If a user wants to find out which dialect word corresponds to a standard Serbian word, or explore a certain type of vocabulary, or certain ethnolinguistic or historical topics, the alphabetic arrangement of the macrostructure will not be able to provide the answers. In these types of situation, the user needs access paths to the dictionary content that go beyond the indexing of the macrostructure.

5. Annotating for multiple access paths

5.1 Standardized Lemmas

The main access structure for the entries in Čemerikić's manuscript is the headword, which is usually underlined on the paper slip. In creating our lemma index, we use the headword, preserving Čemerikić's original spelling. For each graphemically non-standard lemma, however, we provide a standardized spelling alternative. For instance: зъндан > зиндан (semivowel њ > и); тѣмън > таман (semivowel њ > а); зѣмба > зумба (semivowel њ > у); чадър > чадор (semivowel њ > о); диѣбек > дибек (non-standard Cyrillic i-umlaut representing the Turkish vowel ü). The standardized spelling variants are displayed on the page, below the lemma (see Picture 1), and automatically added to the search index so that they appear in the search autocomplete dropdown menu and point to the original entries.

5.2 Near-Synonyms

The entries are furthermore annotated with standard Serbian lexical equivalents. The addition of standard synonyms greatly improves the searchability of the collection because synonyms are also automatically added to the index list. The user can access the entry зъндан, as aforementioned, by searching for the original spelling, the standard orthographic representation of the dialect lexeme (зиндан) as well as its modern standard equivalents затвор or тамница (jail, dungeon).

5.3 Semantic Fields

The collection is furthermore enriched by the application of semantic fields adapted from Buck (1949) in consultation with the questionnaire of the Serbian Dialect Atlas

(Милорадовић, 2012). These top-level semantic fields were chosen specifically to reflect the semantic categories most prevalent in Serbian dialect dictionaries. They have been tested on a wide range of dialect dictionaries to ensure wide coverage and cross-dictionary applicability.

Физички свет (рељеф и метеорологија)	Physical World
Човек (делови тела, физичке и психичке особине)	Man (body parts, physical and psychological features)
Родбина (крвно, бескрвно и духовно сродство, називи за обраћање)	Kinship (consanguine, affinal and spiritual; terms of address)
Медицина (болести, телесни и душевни недостаци, лекови, ветеринарска медицина)	Medicine (illnesses, physical and mental impairments, medicines, veterinary medicine)
Животиње (и сточарство)	Animals (and animal husbandry)
Исхрана (храна и пиће)	Food (and drink)
Одевање (одећа, обућа, накит, нега, дотеривање)	Clothing & Adornment
Кућа (покућство, окућница)	Dwellings & Furniture
Биљке и земљорадња	Vegetation & Agriculture
Кретање (и превоз)	Motion (& Transportation)
Глас (говорење, оглашавање, оноματοпеје)	Voice (speech, including onomatopoeic sounds)
Занимања (занати, алати, предмети везани за занимања, материјали, оружје)	Professions (crafts, tools, objects related to professions, materials, weapons)
Поседовање (имање, трговина)	Possession & Trade
Простор (односи у простору, положај нечега, место, облик, величина)	Spatial Relations
Мере (укључујући новац и бројеве)	Quantity & Number (including money)
Календар (од секунде до века; доба дана, године, месеци, дани у недељи)	Calendar (from second to century; time of the day, seasons, months, days of the week)
Чулна перцепција	Sense Perception
Осећања (све везано за субјективни, морални	Emotion (everything related to the subjective,

или естетски осећај)	moral or esthetic sense)
Ум (интелект, читање и писање; народне умотворине)	Mind & Thought (including reading and writing, folkloric literary expression)
Друштвена организација (територија, институције, право)	Social Organization (territory, institutions, law)
Друштвени живот (све врсте међуљудских односа, игре)	Social Relations (all kinds of interpersonal relations, games)
Веровања (религија, сујеверје, обреди, обичаји)	Beliefs (religion, superstition, rituals, customs)
Ономастика (топоними, антропоними, хидроними, етници, ктетици...)	Onomastics (toponyms, anthroponyms, hydronyms, ethnonyms etc.)
Тајни језици (нпр. бошкачки, гегавачки, слепачки...)	Cant (secret languages meant to exclude or mislead people outside the group that speaks them)

Table 1: Semantic fields

The labels for the semantic fields in each entry can be used as a navigational tool to display a list of all entries from the given field, enabling thus a kind of thematic browsing through the collection.

6. Conclusion and Further Work

The agile approach to digitization of the Čemerikić manuscript allows us to deliver rapidly and annotate incrementally, continuously increasing the use value of the collection by providing new access paths for searching and navigation (lemmas, standardized lemmas, synonyms, semantic fields). Since the work on the collection is ongoing, it would be difficult to provide a reliable quantitative overview of the elements added at this point. Once the current process of annotation is complete, however, we will be able not only to assess our own annotations statistically, but also to quantify the distribution of semantic fields across Čemerikić's collection as a whole.

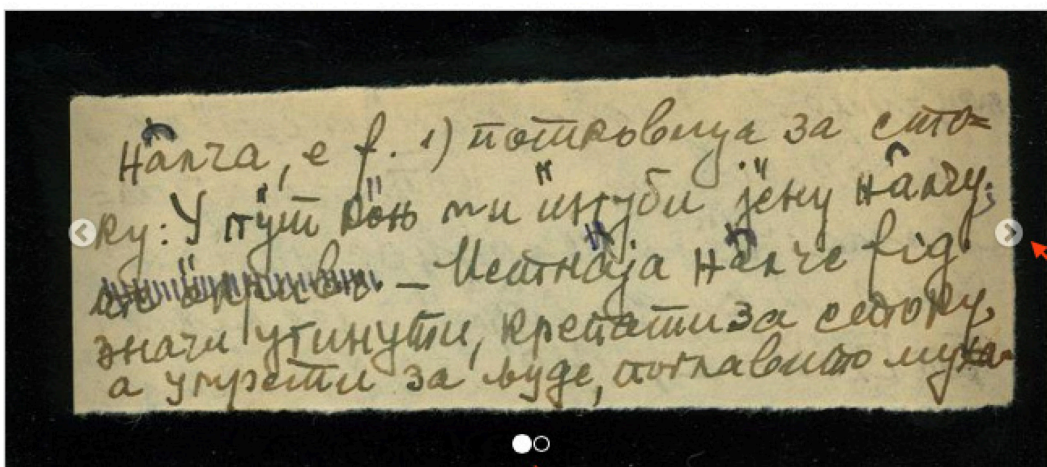
In addition to the semantic fields, which offer a closed set of choices for tagging entries in the Čemerikić collection, we are planning to implement a free-text tagging option as well, to allow for even more flexibility in the tagging process. The multiple access paths will be especially useful in a future iteration of the project, in which we will also open API access to the collection in order to facilitate the integration of the digitized paper slips with other electronic dictionaries and/or multi-dictionary portals.

← Претходни запис Смути па проспи Следећи запис → **admin commands for data enrichment**

налча lemma synonyms semantic fields admin commands for data enrichment

⇒ потковица ⇒ плоча zanimanja zivotinje čovек **5** priority level

налча, е f. 1) потковица за стоку: У пўт коњ ми йзгуби јёну нълчу; - Метнълја нълче fig. значи угинути, крепати за стоку, а умрети за људе, поглавито мухамеданце. - Исп. метнълти, мётнем. 2) потковица на ципелама, чизмама: Нёси кондўре у Рйсте Кикмйра да тўри нове нълче. Ет. Ел. Реч. I, 440. У Арб. potkua = поткова, потковица. transcription



Цитат

Димитрије Чемериќић, "налча," *prepis.org*, приступљено 05.06.2015., <http://www.prepis.org/items/show/20304>.

Транскрибуј овај запис

1. [DC.ZRP.Nn10289.jpg](#)
2. [DC.ZRP.Nn10290.jpg](#) Links toward the transcription interface

Подели преко друштвених мрежа

f t g p Share via social networks

Figure 2: Entry for нълча

7. Acknowledgements

This article is the result of research on the project Nr. 47016 “Interdisciplinary Research of the Cultural and Linguistic Heritage of the Republic of Serbia and the Development of the Web Lexicon of Serbian Culture” which is fully financed by the Ministry of Education and Science of the Republic of Serbia. Further financing for the advanced annotation of the manuscript has been provided by the Ministry of Culture and Information of the Republic of Serbia.

8. References

- Bakken, K. (2006). The Dictionary and Its Sources: The Ideal of Integration and the Example Norsk Ordbok. *Atti del XII Congresso Internazionale di Lessicografia*: Torino, 6-9 settembre 2006, pp. 117-22.
- Buck, C. D. (1949). *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A contribution to the History of Ideas*. Chicago: University of Chicago Press.
- Bunke, H. (2003). Recognition of Cursive Roman Handwriting: Past, Present and Future. In *Document Analysis and Recognition: Proceedings of the Seventh International Conference*, pp. 448-59.
- Considine, J. (2008). *Dictionaries in Early Modern Europe: Lexicography and the Making of Heritage*. Cambridge: Cambridge University Press.
- Fellbaum, C. (2009). *Idioms and Collocations: Corpus-Based Linguistic and Lexicographic Studies*. London: Continuum.
- Hass Weinberg, B. (2010). Indexing: History and Theory. In Bates, Marcia J. and Mary Niles Maack (eds.), *Encyclopedia of Library and Information Sciences*, Boca Raton, FL: CRC Press.
- Hausmann, F. J. & Wiegand, H. E. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey. In F. J. Hausmann, O. Reichmann, & H. E. Wiegand (eds.) *Wörterbücher: ein internationales Handbuch zur Lexikographie*. Berlin/New York: W. de Gruyter.
- Kucsma, J., Reiss, K. & Sidman, A. (2010). Using Omeka to Build Digital Collections: The METRO Case Study. *D-Lib Magazine*, 16(3): np.
- Landau, S. I. (1984). *Dictionaries: The Art and Craft of Lexicography*. New York: The Scribner Press.
- Makkai, A. (1980). Theoretical and Practical Aspects of an Associative Lexicon for 20th-Century English. In L. Zgusta (ed.) *Theory and Method in Lexicography: Western and Non-Western Perspectives*, Columbia, S. Carolina: Hornbeam Press.
- Nielsen, S. & Almind, R. (2011). From Data to Dictionary. In P. Fuertes Olivera & H. Bergenholtz (eds.), *E-Lxicography: The Internet, Digital Initiatives and Lexicography*. London and New York: Continuum, pp. 141-167.

- Nielsen, S. (2011). Function- and User-Related Definitions in Online Dictionaries. In Карташкова, Ф. И. (ed.), *Ивановская лексикографическая школа: традиции и инновации: сб. науч. ст, посвященный юбилею научного руководителя школы, заслуженного работника Высшей школы РФ, доктора филологических наук, профессора Ольги Михайловны Карповой*. Иваново: Ивановский Государственный Университет, pp. 197-219.
- Plötz, T. & Fink, G. A. (2009). Markov Models for Offline Handwriting Recognition: A Survey. *International Journal on Document Analysis and Recognition (IJ DAR)*, 12(4), pp. 269-298.
- Reichmann, O. (2012). *Historische Lexikographie: Ideen, Verwirklichungen, Reflexionen an Beispielen des Deutschen, Niederländischen und Englischen*. Berlin; Boston: De Gruyter.
- Sayre, K. M. (1973). Machine Recognition of Handwritten Words: A Project Report. *Pattern Recognition*, 5(3), pp. 213-228.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Tarp, S. (2008). *Lexicography in the Borderland Between Knowledge and Non-Knowledge: General Lexicographical Theory With Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Tasovac, T. (2010). Reimagining the Dictionary, or Why Lexicography Needs Digital Humanities. *Digital Humanities 2010*, <http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab>.
- Tasovac, T. (2012). Potentials and challenges of WordNet-based pedagogical lexicography: The Transpoetika Dictionary. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford University Press, pp. 237-258.
- Tomás, S. (2011). Exposiciones digitales y reutilización: aplicación del software libre Omeka para la publicación estructurada. *Métodos de información*, 2(2), pp. 29-46.
- Vinciarelli, A. (2002). A survey on off-line cursive word recognition. *Pattern Recognition*, 35(7), pp. 1433-1446.
- Wandl-Vogt, E. (2005). *From Paper Slips to the Electronic Archive. Cross-linking Potential in 90 years of Lexicographic Work at the Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*. Budapest: Linguistic Institute, Hungarian Academy of Sciences.
- Милорадовић, С. (2012). Лингвистички атласи – „централни инструмент“ савремене дијалектологије. *Зборник радова Етнографског института САНУ: Теренска истраживања – поетика сусрета*, 27, pp. 141-51.
- Петровић, С. (2012). *Турцизми у српском призренском говору: на материјалу из рукописне збирке речи Димитрија Чемериџића*. Београд: Институт за српски језик САНУ.
- Петровић, С. & Тасовац, Т. (2013). *Призрен - живот у речима*. Београд: Институт за српски језик САНУ.
- Петровић, С. & Тасовац, Т. (2014). *Збирка речи Димитрија Чемериџића као извор за*

етнолингвистичка и етнологска истраживања. *Гласник Етнографског института*, LXII(2), pp. 171-179.

- Ристић, С., Самарџић, Т., Јакић, М., Марковић, А. & Ивановић, Н. (2011). Значај дигитализације језичких ресурса Речника српскохрватског књижевног и народног језика САНУ за развој науке и очуване културне баштине. In *Дигитализација културне и научне баштине*, 4, pp. 79-108.
- Тасовац, Т. & Петровић, С. (eds.) (2013). *Препис.орг: платформа за дигитална издања и транскрипцију српског рукописног наслеђа*. Београд: Центар за дигиталне хуманистичке науке.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Longest–commonest Match

Adam Kilgarriff¹, Vít Baisa^{1,2}, Pavel Rychlý^{1,2}, Miloš Jakubíček^{1,2}

¹Lexical Computing Ltd., Brighton, United Kingdom

²Natural Language Processing Centre, Masaryk University, Faculty of Informatics, Brno, Czech Republic
{vit.baisa,pavel.rychly,milos.jakubicek}@sketchengine.co.uk

Abstract

Finding two-word collocations is a well-studied task within natural language processing. The result of this task for a given headword is usually a list of collocations sorted by a salience score. In corpus manager Sketch Engine, these pairs are extracted from data using a word sketch grammar relation rules and log-dice statistics resulting in a sorted list of triples <headword, grammar-relation, collocate>. The longest–commonest match is a straightforward extension of these two-word collocations into multiword expressions. The resulting expressions are also very useful for representing the most common realisation of the collocational pair and to facilitate the interpretation of the raw triplet because sometimes, for such a triple, it is not clear from what texts it comes. We present here an algorithm behind the longest–commonest match together with a simple evaluation. The longest–commonest match is already implemented in Sketch Engine.

Keywords: multiword expression; collocation; word sketch; Sketch Engine

1. Introduction

The prospects for automatically identifying two-word multiwords¹ in corpora have been explored in depth, and there are now well-established methods in widespread use². But many multiwords are of more than two words and research into methods for finding items of three and more words has been less successful (Kilgarriff et al., 2012). Here we introduce a method for finding salient multiword expressions based on collocations—word sketches (Kilgarriff et al., 2004). The resulting multiword expressions are also very useful when it is not clear from what texts a collocation pair comes, e.g. <flame_n, object-of, put_v>, <love_v, object, neighbor_n>, etc. The longest–commonest match is therefore also a representative expression for collocational pairs. In the next section we describe the longest–commonest match, the algorithm and a rationale behind it. Then we present a small scale evaluation of the algorithm which was done on an English corpus and a set of collocation pairs. In the fourth section we discuss some issues with finding the longest–commonest matches and in the fifth section we propose some possible improvements of the algorithm.

¹ We use ‘multiwords’ as a cover-all term to include collocations, colligations, idioms, set phrases etc.

² (Church and Hanks, 1990; Pearce, 2002) and others.

2. Longest–commonest match

In this section we describe an algorithm for identifying candidate multiwords of more than two words called the *longest–commonest match* (LC match; in the previous works we have used the terms commonest match or commonest string). It starts from a two-word collocation, as identified using well-established techniques (dependency-parsing, followed by finding high-salience pairs of lexical arguments to a dependency relation) (Kilgarriff et al., 2004). We then explore whether a sufficient proportion of all collocation examples is accounted for by a particular string—the longest–commonest match.

The two-word collocations from which we start are triples: $\langle \text{lemma1}, \text{grammar-relation}, \text{lemma2} \rangle$ for example $\langle \text{drink}_v, \text{object}, \text{tea}_n \rangle$. The lexical arguments are lemmas, not word forms, and are associated with word class, here n for noun, v for verb. The corpus instances that will have contributed to giving a high score include “They were drinking tea.” and “The tea had been drunk half an hour earlier.” The first argument may be to the right, or to the left, of the second. It depends on a particular grammar relation which is described in word sketch grammar rules.

If a particular string (consisting of word forms, not lemmas) accounts for a high proportion of the corpus instances, it becomes a candidate multiword-of-more-than-two-words. We want the string to be common and we want it to be long. Hence the name. We find the longest–commonest match as follows:

Input: two lemmas forming a collocation pair, and N hits for the pair in a given corpus; parameters: proportion p (1/4), minimum frequency minf (5) and minimum number of hits minhits (10).

Initialization: initialize the match as, for each hit, the string that starts with the beginning of the first of the two lemmas and ends with the end of the second. If the initial number of hits is less than minhits then return empty string, i.e. there is no LC match for a given lemmas.

For each hit, gather the contexts comprising the match, the preceding three tokens (the left context) and the following three tokens (the right context).

1. Count the instances of each unique string. Do any of them occur more than $p \times N$?
2. If no, return empty string.
3. If yes
 - (a) Call the most frequent string **LC match**
 - (b) Look at the first tokens in its right and left contexts (max 3 positions), if we cannot expand farther, return **LC match**
 - (c) Do any of the expanded strings occur more than $p \times N$ times?
 - (d) If no, return the current **LC match**.

- (e) If yes:
- i. Assign the most frequent expanded string to **LC match**.
 - ii. Go to 3.b.

If there are no strings meeting the thresholds, there is no **LC match** (it is empty). Since LC match is extracted from corpus examples it consists from word forms not from lemmas.

An earlier version of this work was presented at EURALEX 2012 (Kilgarriff et al., 2012). We present it here again because it was only covered very briefly, and in the meantime we have developed a version of the algorithm that works very fast even for multi-billion word corpora, and is fully integrated into our corpus query system Sketch Engine, see Figure 1. It is a word sketch table for the headword *put* (verb). The first column contains collocates, the second column contains grammar relations, the third and fourth columns contain frequency and salience score and the last column contains LC matches.

put <small>(verb)</small> British National Corpus + Commonest Match freq = 67,367				
down	<i>part_trans</i>	2,534	9.94	put down
forward	<i>modifier</i>	1,720	11.56	put forward
up	<i>part_intrans</i>	1,176	7.83	to put up with
just	<i>modifier</i>	832	8.77	just put
in	<i>part_intrans</i>	796	8.88	put in
pressure	<i>object</i>	519	8.20	put pressure on
then	<i>modifier</i>	438	8.05	and then put
head	<i>object</i>	387	6.92	put his head
thing	<i>object</i>	382	6.57	put things
end	<i>object</i>	327	6.80	to put an end to
off	<i>part_intrans</i>	320	7.76	be put off
place	<i>pp_in-p</i>	241	6.61	put in place
right	<i>np_adj_comp</i>	217	7.96	put it right
phone	<i>object</i>	213	7.47	put the phone down
hand	<i>part_out-a_obj</i>	167	5.92	put out a hand
lot	<i>object</i>	147	6.30	put a lot of
use	<i>pp_to-p</i>	145	6.33	put to good use
kettle	<i>object</i>	142	7.15	put the kettle on

Figure 1: Integration of the longest–commonest match in Sketch Engine

Comment on Figure 1 In some cases, the LC match is simply a bigram of adjoint collocates: put down, put in, etc. Sometimes the two collocates are separated by a token thus producing a trigram: put his head, put in place. This may occur when a headword is a phrasal verb with an object (put in place). In the example there are also 4-grams, e.g. put the phone down. It again captures phrasal verb and this time the object comes together with the determiner. It

results directly from the LC match algorithm that these “examples” are the most frequent realisations of the collocation pairs.

Implementation We have implemented the longest–commonest match in Python language and integrated it into Bonito/manatee corpus manager (Rychlý, 2007). The script is run only once at the time of corpus compilation and the resulting longest–commonest matches for each collocation pair are saved into word sketch data index files. That is why it is immediately available when showing word sketch data (as in Figure 1). The downside is that we need to set the parameters \mathbf{p} , `minhits`, `minf` before the corpus compilation process. To compute and show LC matches with different settings, we need to process the whole corpus again and store the found matches in separate index files.

3. Evaluation

To overcome the issue of pre-setting the parameters, we designed a simple evaluation of various settings to find out what is the best combination of the parameters. We were most interested in the proportion, parameter (\mathbf{p}). Other parameters (`minhits`, `minf`) are good for controlling coverage of the output and for limiting the time needed for computing LC matches for all collocation pairs in a corpus. The width of the token context (3 to the left and to the right) is not adjustable, but it could be another parameter available for tuning. Nevertheless we have decided to compare results for various settings of the only parameter, \mathbf{p} .

Since this is not a classification task, it is not possible to measure the standard metrics precision and coverage. We have let two annotators decide for a set of 500 LC matches (extracted from SkELL corpus (Baisa and Suchomel, 2014)) which are good (helpful, well-formed, informative) and which are wrong but the definition of what is good and wrong was hard to agree on. Instead, we extracted LC matches for various settings of the proportion parameter \mathbf{p} and let two annotators compare the resulting LC matches. The features were the same as before. Is one LC match a better example for a collocation pair? Is one LC match more informative, explanatory and understandable than other matches? The difference was that annotators were comparing three LC matches instead of telling yes or no for particular LC matches. The agreement was much better for this variant. For the results, see Table 1.

Two annotators (A1, A2) were provided with 102 randomly selected collocation pairs (examples below) together with three LC matches where the proportions (parameter \mathbf{p} in the algorithm) were 0.5, 0.25 and 0.16 (columns LC match 1, 2 and 3, respectively). Their task was to select the most helpful LC match for understanding the collocation pair (first three columns). When two columns were the same, both column numbers were used in the annotation (last two columns labelled with annotator’s indication). The most frequently favoured LC match (61%) was the least restrictive ($\mathbf{p} = 0.16$) which means that in general, the length was preferred against the commonness of the strings. LC match 2 has been selected in 58% of

Headword	Relation	Collocate	LC match 1	LC match 2	LC match 3	A1	A2
love-v	modifier	personally-a	I personally love	. I personally love	. I personally love	1	1
calorie-n	object-of	need-v			calories needed	3	3
flame-n	object-of	put-v		put the flames out	put the flames out	23	23
vision-n	modifier	limited-j	limited vision	limited vision	limited vision .	12	12
meeting-n	modifier	joint-j	joint meeting	a joint meeting	a joint meeting of the	2	3
classroom-n	modifier	virtual-j	virtual classroom	virtual classroom	a virtual classroom	3	12
unofficial-j	modifies	symbol-n	unofficial symbol of	an unofficial symbol of	an unofficial symbol of	23	23
worthwhile-j	adj-comp-of	seem-v		seems worthwhile	seems worthwhile to	3	3
climb-v	modifier	gradually-a		gradually climbing	gradually climbing	23	23
delicate-j	modifies	matter-n	delicate matter	a delicate matter	a delicate matter	23	23

Table 1: Example of lines from evaluation data together with annotators’ choices.

cases and LC match 1 (the most restrictive p) in 33% of cases. Mind that it was not a simple classification but rather the assignment of (multiple) labels to the LC matches (columns). That is why the percentages do not sum up to 100%. There was 67% agreement between the two annotators.

Evaluation data We have used a random sample from the dataset used in (Kilgarriff et al., 2014). The dataset³ contains only verbs, nouns and adjectives as headwords in the English language. Here we include some examples of collocation pairs from the gold standard dataset (headword, collocate): (average_j, age_n), (black_j, hole_n), (circuit_n, short_j), (delicate_j, ecosystem_n), (empty_j, bin_n), (free_j, lunch_n), (global_j, crisis_n), (harp_n, player_n), (inject_v, vaccine_n), (kid_v, entirely_a), (love_v, genuinely_a), (metal_n, galvanized_j), (operational_j, remain_v), (past_j, participle_n), (root_v, firmly_a), (slow_v, abruptly_a), (tempting_j, extremely_a), (unofficial_j, biography_n), (virulent_j, campaign_n), (weed_n, grow_v), (worthwhile_j, highly_j).

4. Discussion

The evaluation helped us to discover some issues which we need to address. The most obvious is the punctuation being part of LC matches which was never preferred by annotators. It would be straightforward to strip it from the LC matches, nevertheless we are not sure if this is desirable. Sometimes it might be helpful to know that some phrases contain a comma or a full stop. It might help users understand that a certain phrase is used usually at the end of sentence (or at the beginning as the first example from Table 1 indicates) or that it is separated from the rest of the sentence by a comma.

Since the algorithm is language-independent (once we have a list of collocation pairs), adding a language-dependent list of punctuation to be removed from LC matches would spoil this desired feature. But a simple approach usable for most European languages would be simply to strip all commas, semicolons, full stops, exclamation and question marks. The punctuation

³ Available for download: <http://www.sketchengine.co.uk/documentation/wiki/CorpEval>

would be removed only from the beginning and the end of a LC match as a punctuation mark within an LC match will have an obvious interpretation.

It is also clear that any match is preferred against an empty LC match. As for finding multiword expressions, empty matches decrease coverage which is not a big issue; but regarding the second goal of a LC match it surely decreases understanding of the original collocation pair. In other words, it is always helpful to have at least the collocation pair in the most common order (see examples in Table 1: limited vision, joint meeting, etc.) than to rely only on the original collocation pair. Thus it is reasonable to use a rather less restrictive parameter \mathbf{p} .

The original combination of parameters proved to be solid. We found that using a somewhat less restrictive parameter \mathbf{p} yields slightly better results but the difference is too small (3%) for us to change the default settings currently used in Sketch Engine.

5. Further work

Based on the evaluation and on a brief error analysis of the algorithm, we want to explore a few possible improvements of the algorithm in the future.

First, in some cases, LC matches were skewed by many occurrences of a string within one specific document. It could be treated by filtering input concordances to contain one (e.g. the first one) hit per document. This filter is already implemented in Sketch Engine.

In general, the algorithm suffers when duplicate documents are present in a corpus. This is addressed by de-duplication phase when building such corpus and has been treated in (Pomikálek, 2011). Sketch Engine uses procedures described in the PhD thesis.

Second, the current algorithm works with parameters which are fixed for all concordances / collocation pairs. It is to be evaluated whether making the parameters relative to concordance size (N input hits) would help.

Another improvement to the algorithm efficiency would be sampling of input concordances. The time complexity of the algorithm is roughly linear to the length of the input (concordance with N lines). For very large concordances (concordance for collocation pair $take_v$, $place_n$ has almost 1 million hits in corpus enTenTen12) it would be reasonable to use a random sample of such concordances. The question is whether the sample should have a fixed size or if the size should be (again) relative to the size of the original concordance. Despite the resulting LC matches being thought to be the same it is necessary to try and evaluate it. The sampling is also already available in Sketch Engine.

It was not mentioned earlier but the algorithm does not depend on collocation pairs. It is simply applicable for any concordance, meaning that for any search in a corpus, we can

compute (on-the-fly) the longest–commonest match or the longest–commonest KWIC as a generalized and expanded representation of the original corpus search query. It could be a handy feature to provide such generalized KWIC for all searches in Sketch Engine but again, we would need to evaluate its contribution based probably on user feedback.

6. Conclusion

We believe that the LC match will improve understanding of sometimes cryptic collocation pairs (triples) as available in Sketch Engine. The resulting strings are also salient multiword expressions despite the fact that it is not straightforward to properly evaluate the quality of these multiwords.

7. Acknowledgement

This paper was published posthumously for Adam Kilgarriff died on Saturday, May 16th, 2015. He has been working on this paper even in his later days while undergoing a palliative chemotherapy. We dedicate this paper to him, as the originator of the longest–commonest match.



Adam Kilgarriff (12 February 1960 – 16 May 2015)

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

8. References

- Kilgarriff, A., Rychlý, P., Kovář, V., & Baisa, V. (2012). Finding multiwords of more than two words. In Fjeld, R. V. & Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, Oslo, Norway. Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 693–700.
- Baisa, V. & Suchomel, V. (2014). SkELL: Web interface for english language learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 63–70.
- Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1), pp. 22–29.
- Kilgarriff, A., Rychlý, P., Jakubíček, M., Kovář, V., Baisa, V., & Kocincová, L. (2014). Extrinsic corpus evaluation with a collocation dictionary task. In N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., & Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 454–552.
- Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In Williams, G. & Vessier, S., editors, *Proceedings of the 11th EURALEX International Congress*, Lorient, France. Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, pp. 105–115.
- Pearce, D. (2002). A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation, LREC'02*, pp. 1530–1536.
- Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. *PhD en informatique, Masarykova univerzita, Fakulta informatiky*.
- Rychlý, P. (2007). Manatee/bonito-a modular corpus manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 65–70.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary

Franck Sajous and Nabil Hathout

CLLE-ERSS (CNRS & Université de Toulouse 2)
franck.sajous@univ-tlse2.fr, nabil.hathout@univ-tlse2.fr

Abstract

This article introduces GLAWI, a large XML-encoded machine-readable dictionary automatically extracted from Wiktionnaire, the French edition of Wiktionary. GLAWI contains 1,341,410 articles and is released under a free license. Besides the size of its headword list, GLAWI inherits from Wiktionnaire its original macrostructure and the richness of its lexicographic descriptions: articles contain etymologies, definitions, usage examples, inflectional paradigms, lexical relations and phonemic transcriptions. The paper first gives some insights on the nature and content of Wiktionnaire, with a particular focus on its encoding format, before presenting our approach, the standardization of its microstructure and the conversion into XML. First intended to meet NLP needs, GLAWI has been used to create a number of customized lexicons dedicated to specific uses including linguistic description and psycholinguistics. The main one is GLÀFF, a large inflectional and phonological lexicon of French. We show that many more specific on demand lexicons can be easily derived from the large body of lexical knowledge encoded in GLAWI.

Keywords: French Machine-Readable Dictionary; Free Lexical Resource; Wiktionary; Wiktionnaire

1. Introduction

Recent papers on electronic lexicography investigate if and how linguistics (computational or not) can contribute to lexicography (Rundell, 2012), how NLP can automate the process of collecting material and analyze it (Rundell and Kilgarriff, 2011) or what are the skills and the needs of specific end-users (Lew, 2013). As linguists and NLP researchers, we are reciprocally interested in the exploitation of dictionaries for linguistic description (phonology, morphology, lexicology, semantics, etc.) and NLP use. Leveraging machine-readable dictionaries (MRDs) for the acquisition of lexical and semantic relations, for the development of derived lexical resources, or for various linguistic studies, was common practice in 1980's (Calzolari, 1988; Chodorow et al., 1985; Markowitz et al., 1986). The availability of large corpora and the subsequent rise of corpus linguistics highlighted MRDs' restricted coverage and their potential out-of-dateness. However, new online dictionaries with no size restriction and a steadily ongoing development such as Wiktionary may renew the interest for electronic lexicons. Besides its wide coverage and its potential for constant updates, Wiktionary has an interesting

macrostructure and features a rich lexical knowledge: articles include etymologies, definitions, lemmas and inflected forms, lexical semantic and morphological relations, translations and phonemic transcriptions.

For six years, we have exploited Wiktionary and more specifically its French language edition called Wiktionnaire, assessed its quality and investigated to what extent it can meet linguistics and NLP’s needs in terms of lexical resources. Each experiment led us to extract various information from the collaborative dictionary and develop specific resources targeting different uses. In order to experiment algorithms based on random walks to enrich lexical networks (Sajous et al., 2010), we produced partial XML versions of the French and the English editions of Wiktionary, called WiktionaryX.¹ This resource contains a selection of fields extracted from the English and French wiktionaries: definitions, lexical semantic relations and translations. We then produced an inflectional lexicon called GLÀFF (Hathout et al., 2014b; Sajous et al., 2013a) that contains inflected forms, lemmas, morphosyntactic features and phonemic transcriptions.² This lexicon was intended to be used by syntactic parsers like Talismane (Urieli, 2013) or for research in computational morphology (Hathout, 2011; Hathout and Namer, 2014). A conclusion we drew is that Wiktionnaire’s rich content is a valuable resource whose main drawback is its heterogeneous and volatile format, which impedes an easy and direct exploitation. A significant contribution of GLAWI is the standardization of Wiktionnaire’s microstructure. Standing for “*GLÀFF and WiktionaryX*”, GLAWI also results from our will to unify parallel efforts and produce a single resource that includes all information contained in Wiktionnaire in a workable format (XML). It is however not a simple merge of GLÀFF and WiktionaryX: new information is also extracted, like the morphological relations omitted from the two previous resources. We also went one step further in the homogenizing process. Our aim is to finely parse Wiktionnaire so that we can make accessible in a standard and coherent format as much information as available. To that extent, our approach differs from that of Sérasset (2012), whose aim is to build a multilingual network containing “easily extractable” (i.e. *regular*) entries, which results in a restricted coverage. Conversely, we made a particular effort to detect information, whatever format it is encoded into and wherever it occurs.

GLAWI is conceived as a general-purpose MRD intended to be easy to use, like such or as a starting point to tailor specific lexicons. GLÀFF, as well as other resources that we extracted so far from Wiktionnaire, will now be derived easily from GLAWI.

This article is organized as follow: in section 2, we give some insights into the Wiktionnaire’s nature; we describe GLAWI in section 3 and explain how we developed it by converting Wiktionnaire into a structured format. We illustrate in section 4 how we derived specific lexicons for various purposes directly from GLAWI, before contemplating some perspectives in section 5.

1 WiktionaryX is available at http://redac.univ-tlse2.fr/lexicons/wiktionaryx_en.html

2 GLÀFF is available at http://redac.univ-tlse2.fr/lexicons/glaff_en.html

2. Wiktionary and Wiktionnaire

Wiktionary, presented as “*the lexical companion to Wikipedia*”,³ is, like Wikipedia and other related wikis, a public collaborative project. Any internet user can contribute, whatever their skills. Editorial policies exist, however modifications are published immediately. “Wiktionary” is used to refer both to the English edition and to the whole project (the 171 language editions). We hereafter give some details about the nature of Wiktionary and its French edition called Wiktionnaire.⁴

General description. The basic unit of Wiktionnaire’s articles is the word form. A given article (described in a web page, at a URL) may contain several entries having distinct or identical parts of speech (POSS). A POS section may correspond to a canonical form (lemma) or an inflected form. Figure 1a depicts an excerpt of the page of *affluent*.

This page shows that the word form is the lemma of an adjective ‘tributary’, a noun ‘tributary’, and is an inflected form of the verb *affluer* ‘to flow’. The adjective POS-section gives the four inflected forms of its paradigm, each form linking to a dedicated page of the dictionary. Figure 1c shows the page corresponding to the feminine singular form *affluente*, which links back to the lemmatized form *affluent*. The inflected verbal forms of Figure 1a link to the page of the infinitive form, depicted in Figure 2. Unlike the pages of noun and adjective lemmas, the ones corresponding to verb infinitive forms do not contain their paradigms (a verb’s paradigm amounts to 48 forms in French which would cause a display overload). Instead, a link to a conjugation table is inserted. A shortened example of such a table is given for *affluer* in Figure 3. Each inflected form links to a dedicated page, when this page exists. This hypertextual macrostructure shows that the relations between the different forms of a given paradigm are located in different parts of the dictionary. We discuss the incidence of this feature in section 3.2.


The microstructure of an article contains an etymology section and one or more POS sections which provide a sense inventory including glosses and examples. POS sections may also include translations, lexical semantic relations (synonymy/antonymy, hypernymy/hyponymy, holonymy/meronymy), morphological relations (derivation, compounds) or more fuzzy relations such as *apparentés* ‘related’. Phonemic transcriptions may appear at the article level (when all entries share a common pronunciation), in the first line of the POS level and/or in the paradigms. It is worth noting that each language edition has its own microstructure. For example, the semantic relations are indexed to the word senses in the German Wiktionary. They are listed in POS sections in Wiktionnaire but appear at the article top level in the Italian Wiktionary.

³ <http://en.wiktionary.org>

⁴ Additional descriptions can be found in (Meyer, 2013; Navarro et al., 2009; Sajous et al., 2013b)

<http://fr.wiktionary.org/wiki/affluent>


affluent

 **Adjectif**

affluent

1. (*Géographie*) Qui se **jette dans** un **autre** en **parlant** d'un **cours** d'eau.
2. (*Médecine*) Qui **affluent**, qui se **portent** en **abondance vers** quelque **partie** du **corps**.


	Singulier	Pluriel
Masculin	affluent <i>/a.fly.ɑ̃/</i>	affluents <i>/a.fly.ɑ̃/</i>
Féminin	affluente <i>/a.fly.ɑ̃t/</i>	affluentes <i>/a.fly.ɑ̃t/</i>

 **Nom commun**

affluent */a.fly.ɑ̃/* masculin

1. (*Géographie*) **Cours d'eau** qui se **jette dans** un **autre**.

Singulier	Pluriel
affluent	affluents
<i>/a.fly.ɑ̃/</i>	

 **Forme de verbe**

affluent */a.fly/*

1. *Troisième personne du pluriel de l'indicatif présent de affluer.*
2. *Troisième personne du pluriel du subjonctif présent de affluer.*

Conjugaison du verbe <i>affluer</i>		
INDICATIF	Présent	ils/elles affluent
SUBJONCTIF	Présent	qu'ils/elles affluent

(a) POS sections of the article *affluent*

```

{{-adj-|fr}}
{{fr-accord-cons|a.fly.ɑ̃|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]] d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ɑ̃}}


{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui|}}
'''affluent''' {{pron|a.fly|fr}}
# ''Troisième personne du pluriel de l'indicatif présent de'' [[affluer]].
# ''Troisième personne du pluriel du subjonctif présent de'' [[affluer]].

```

(b) Wikicode of the article *affluent*

<http://fr.wiktionary.org/wiki/affluente>

affluente

 **Forme d'adjectif**

affluente *féminin /a.fly.ɑ̃t/*

1. *Féminin singulier de affluent.*

```

{{-flex-adj-|fr}}
'''affluente''' {{f}} {{pron|a.fly.ɑ̃t|lang=fr}}
# ''Féminin singulier de'' [[affluent#fr-adj|affluent]].

```


(c) Article *affluente* and corresponding wikicode

Figure 1: Excerpts of Wiktionnaire's articles *affluent* and *affluente*

An inappropriate software infrastructure (and its consequences). Launched in 2003, one year after the English edition, Wiktionnaire's underlying infrastructure is the MediaWiki engine, used by all the Wikimedia projects. Examples of the encoding format, called *wikicode*, are given in Figures 1b and 1c.

Rundell and Kilgarriff (2011) attribute to Laurence Urdang the first vision, in mid 1960's, of the dictionary as a database “*facilitating and rationalizing the capture, storage and manipulation of dictionary text*”. Systematic check of cross-references was seen as an early benefit of this approach. Four decades later, Wiktionary, a dictionary born online, was encoded into unstructured text, ignoring the necessity of a database oriented design. Evan Jones, the author of the tool *wikipedia2text*,⁵ states that “*one of the biggest problems is that there is*

⁵ <http://www.evanjones.ca/software/wikipedia2text.html>

 **Verbe**

affluer /a.fly.e/ *intransitif* 1^{er} groupe (conjugaison)

- Couler vers, en parlant des eaux qui se rendent et se réunissent dans un même lit.
 - Plusieurs ruisseaux et plusieurs rivières **affluent** dans la Seine, dans le Rhône.
- (Par analogie) Couler vers, en parlant des humeur du corps
 - Il faut empêcher le sang d'**affluer** vers telle partie.
- (Figuré) Abonder, arriver en abondance.

Figure 2: Excerpt of Wiktionnaire's article *affluer*

http://fr.wiktionary.org/wiki/Annexe:Conjugaison_en_français/affluer

Modes impersonnels

Mode	Présent		Passé	
Infinitif	affluer	/a.flɥe/	avoir afflué	/a.vwaʁ_a.flɥe/
Gérondif	en affluant	/ɑ̃.n_a.flɥɑ̃/	en ayant afflué	/ɑ̃.n_ɛ.jɑ̃.t_a.flɥe/
Participe	affluent	/a.flɥɑ̃/	afflué	/a.flɥe/

Indicatif

Présent		Passé composé	
j'afflue	/ʒ_a.flɥ/	j'ai afflué	/ʒ_e a.flɥe/
tu afflues	/ty a.flɥ/	tu as afflué	/ty a.z_a.flɥe/
il/elle/on afflue	/[il/ɛl/ɔ̃] a.flɥ/	il/elle/on a afflué	/[i.l/ɛ.l/ɔ̃.n]_a.a.flɥe/
nous affluons	/nu.z_a.flɥɔ̃/	nous avons afflué	/nu.z_a.vɔ̃.z_a.flɥe/
vous affluez	/vu.z_a.flɥe/	vous avez afflué	/vu.z_a.ve.z_a.flɥe/
ils/elles affluent	/[il/ɛl].z_a.flɥ/	ils/elles ont afflué	/[i/ɛ].z_ɔ̃.t_a.flɥe/
Imparfait		Plus-que-parfait	
j'affluais	/ʒ_a.flɥɛ/	j'avais afflué	/ʒ_a.vɛ.z_a.flɥe/
tu affluais	/ty a.flɥɛ/	tu avais afflué	/ty a.vɛ.z_a.flɥe/
il/elle/on affluait	/[il/ɛl/ɔ̃] a.flɥɛ/	il/elle/on avait afflué	/[i.l/ɛ.l]
nous affluions	/nu.z_a.flɥjɔ̃/		/ɔ̃.n]_a.vɛ.t_a.flɥe/
vous affluiez	/vu.z_a.flɥje/	nous avions afflué	/nu.z_a.vjɔ̃.z_a.flɥe/
ils/elles affluaient	/[il/ɛl].z_a.flɥɛ/	vous aviez afflué	/vu.z_a.vje.z_a.flɥe/
		ils/elles avaient afflué	/[i/ɛ].z_a.vɛ.t_a.flɥe/

Figure 3: Excerpt of the inflectional paradigm of the verb *affluer* in Wiktionnaire

no well-defined parser for the wiki text that is used to write the articles. The parser is a mess of regular expressions, and users frequently add fragments of arbitrary HTML". Several consequences arise from this situation:

- as no formal syntax of the wikicode is defined, no compliance-check is performed when a contributor edits an article. Encoding errors add to occasional contributors' amateurism.

2. cross-references and consistency checking is impossible. For example, a possible discrepancy between an inflected form given in its dedicated page and another form given in its lemma’s paradigm cannot be detected. Similarly, Figure 1b shows that the same information, namely the inflectional features of the verbal form, appears in two ways: *affluent* as third person plural indicative of *affluer* is both given by the code `ind.p.3p` and by the plain text definition *Troisième personne du pluriel de l’indicatif présent*. Ideally, the two views of the same fact should be generated from the same data. In other words, the plain text definition should be generated from `ind.p.3p`. Instead, it has been manually typed by a contributor. In this example, the redundant information is consistent. Section 3.2 illustrates situations of inconsistencies.
3. the infrastructure, intended to receive contributions in mass, is in reality restricted to internet users who feel at ease with wikicode editing.

The two first items impact both the quality of Wiktionary itself and the conversion process described in section 3.2. The latter item may lead to an under-participation to the project, and a bias regarding what kind of internet users contribute to Wiktionary. A good initiative, first appeared as an optional *gadget* (in Wiktionary’s jargon), is the input field designed to add translations: once a contributor has typed a translation, the graphical interface carries out the corresponding edition of the wikicode. Thus, users unable to edit the wikicode can contribute, and the interface generates an error-free encoding.

The wikicode is volatile over time and is unstable from a language edition to the other. Thus, a parser written for a given edition has to be maintained and cannot be used without adaptation to parse another language edition. A direct consequence is that no fully-automatic update of GLAWI is desirable: potential changes in the wikicode have to be monitored to adapt a given parser to every release of a new dump.

“*Experts and Crowds*” rather than “*Experts vs. Crowds*”. Like Wikipedia, Wiktionary is a wiki that any internet user willing to contribute can edit, whatever their skills, with immediate effect. Zesch and Gurevych (2010) assessed Wiktionary’s usefulness for semantic relatedness computation. Thus, they illustrated the potential of Wiktionary as a resource for NLP, not its primary quality as a dictionary. Kosem et al. (2013) rely on crowdsourcing in a controlled way to perform specific tasks: identifying false collocations and incorrect examples among automatically selected ones. The case of Wiktionary is different: the resource is entirely crowdsourced, with no strong editorial constraint. The legitimacy of the so-called “*wisdom of crowds*” in a lexicographical perspective is discussed by Penta (2011) and Sajous et al. (2014). Regarding Wiktionnaire, it is worth noting that a binary opposition between experts and crowds is not accurate because it has been primarily bootstrapped by automatic imports from editions of two dictionaries fallen into the public domain. Table 1 shows that more than 16% of the entries corresponding to lemmas originate from the 8th edition (1932-1935) of the *Dictionnaire de l’Académie française* (DAF8) or from the 2nd edition (1872-1877)

of the *Littré*. The table also reports the number of articles that refer to another resource (only resources with more than 100 references are listed).⁶ These resources include public-domain editions of digitized dictionaries (DAF8, Littré, Bescherelle, Rivarol), Latin (Gaffiot) or Provençal (Mistral) dictionaries, institutional normative websites such as FranceTerme (France) and GDT (Quebec) and specialized websites (Meyer, an online dictionary of animal sciences).

# Imports	# Articles	%	Reference sources	# Articles	%
0	242499	83.42%	Littré	6497	19.56%
1	48162	16.57%	DAF8	6311	19.00%
2	46	0.02%	TLFi	6256	18.84%
Import sources			Rivarol	4358	13.12%
DAF8	27945	57.91%	Meyer	3523	10.61%
Littré	20278	42.02%	FranceTerme	2922	8.80%
Larousse XIXe	24	0.05%	Mistral	650	1.96%
# References			ODS5	394	1.19%
0	260362	89.56%	GDT	200	0.60%
1	27818	9.57%	DAF9	195	0.59%
2	2268	0.78%	Bescherelle	116	0.35%
3	208	0.07%	Gaffiot	105	0.32%
4	32	0.01%	Reverso	100	0.30%

Table 1: Imports and references in Wiktionnaire’s articles (lemmas)

3. GLAWI

3.1 Resource description

GLAWI is a MRD resulting from the conversion of the Wiktionnaire into an XML-structured format. The resource, released under a free license (CC By-SA),⁷ contains 1,341,410 articles, one for each page of Wiktionnaire. GLAWI’s general structure is similar to that of Wiktionnaire as exemplified by the article of *mousse* given in Figure 4.

The meta section. The **meta** markup is used to indicate that an article has been imported from, or refers to, another dictionary (cf. section 2): the article *nénu Phar* (Figure 5) has been primarily imported from DAF8, while the article *mousse* (Figure 4) refers to the TLFi. This same section is also used to indicate that an article corresponds to a spelling variant such as *nénu Phar*, an alternative form of *nénu far*. Just as in Wikipedia, categories are assigned to pages in Wiktionary. GLAWI’s **meta** section indicates the categories an article belongs to (if

⁶ A reference means that a contributor manually indicated that she/he consulted a given resource when editing an article.

⁷ GLAWI is available at <http://redac.univ-tlse2.fr/lexicons/glawi.html>

```

<article>
  <title>mousse</title>
  <pageId>7930</pageId>
  <meta>
    <category>Lexique en français de la navigation</category>
    <category>Noms multigenres en français</category>
    <reference>TLFi</reference>
  </meta>
  <text>
    <pronunciations>
      <pron region="France">mus</pron>
    </pronunciations>
    <pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
      <paradigm>
        <wiki>{{fr-rég|mus}}</wiki>
        <inflection form="mousse" gracePOS="Ncfs" pron="mus"/>
        <inflection form="mousses" gracePOS="Ncfp" pron="mus"/>
      </paradigm>
      ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="2" gender="m" number="s">
      ...
    </pos>
    <pos type="nom" lemma="1" locution="0" homoNb="3" gender="m" number="s">
      ...
    </pos>
    <pos type="adjectif" lemma="1" locution="0" gender="epicene" number="s">
      ...
    </pos>
    <pos type="verbe" lemma="0" locution="0">
      <inflectionInfos>
        <inflected gracePOS="Vmip1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmip3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmisp1s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmisp3s-" lemma="mousser" pron="mus"/>
        <inflected gracePOS="Vmmp2s-" lemma="mousser" pron="mus"/>
      </inflectionInfos>
    </pos>
  </text>
</article>

```

Figure 4: General structure of an article in GLAWI: *mousse* entries

```

<article>
  <title>nénuphar</title>
  <pageId>168915</pageId>
  <meta>
    <category>Plantes en français</category>
    <category>Fleurs en français</category>
    <import>DAF8</import>
    <spellingVariation norm="nénufar"/>
  </meta>
  ...
</article>

```

Figure 5: GLAWI's metadata for article *nénuphar*

any): for example, *mousse* belongs to nautical slang and is a multigender noun; *nénuphar* belongs to the *Flowers* and *Plants* categories.

POS sections. Articles may contain several POS sections marked by **pos** tags that include grammatical features such as gender, number, valency, homograph number (when relevant) and specify whether a form is multiword or not. An attribute also indicates the lemma of the inflected forms. For example, in Figure 4, the *verb* **pos**-section specifies that *mousse* corresponds to five inflected forms of the verb *mousser* and gives their morphosyntactic descriptions in GRACE format (Rajman et al., 1997).

POS sections also include translations, lexical semantic (synonyms, antonyms, hypernyms, etc.) and morphological (derivative, compound, etc.) relations. An example of such subsections is given in Figure 6 for the feminine noun *mousse* ‘foam’, ‘moss’.

```

<pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
...
<subsection type="translations">
  <trans lang="de">Moos</trans>
  <trans lang="en">foam</trans>
  <trans lang="en">moss</trans>
  <trans lang="es">espuma</trans>
</subsection>
<subsection type="lexSemRel">
  <item type="synonym">écume</item>
  <item type="synonym">bière</item>
</subsection>
<subsection type="morphoRel">
  <item type="derivative">moussant</item>
  <item type="derivative">mousser</item>
  <item type="derivative">mousseux</item>
</subsection>
...
</pos>

```

Figure 6: GLAWI’s lexical relations: translations, lexical semantic, morphological relations

Definitions. Word senses, marked by **definition** tags, are listed in the POS sections of lemmas. A definition contains a gloss and possibly one or more usage examples. Definitions may include labels that give attitudinal, diatopic, diachronic, diafrequential information or indicate that the word belongs to a specialized language. The example in Figure 7 indicates that *mousse*, when used to refer to a beer, is a familiar metonym. This figure also shows that every textual part (gloss, example) is available in four different versions:

1. the original wikicode;
2. an XML formatted version where markups encode wiki typesetting (boldface, italic, etc.), dates, foreign words, mathematical/chemical formulae and external/inner links;
3. a raw text version;


```

<pos type="nom" lemma="1" locution="0" homoNb="1" gender="f" number="s">
...
<definitions>
  <definition>
    <gloss>
      <labels>
        <label type="sem" value="métonymie"/>
        <label type="attitudinal" value="familier"/>
      </labels>
      <wiki>{{méton|fr}} {{familier|fr}} [[bière|Bière]].</wiki>
      <txt>Bière.</txt>
      <xml><innerLink ref="bière">Bière</innerLink>.</xml>
      <conll>1 Bière  bière  NC   nc   n=s       0   root</conll>
    </gloss>
    <example>
      <wiki>''Une bonne ''mousse'' sans faux-col est un oxymore.''</wiki>
      <xml><i>Une bonne <b>mousse</b> sans faux-col est un oxymore.</i></xml>
      <txt>Une bonne mousse sans faux-col est un oxymore.</txt>
      <conll> 1 Une      une      DET  DET  g=f|n=s      3   det
              2 bonne   bon      ADJ  adj  g=f|n=s      3   mod
              3 mousse  mousse  NC   nc   n=s          6   suj
              4 sans    sans    P     P     -            3   dep
              5 faux-col _       NC   -     -            4   prep
              6 est     être    V     v    n=s|p=3|t=pst 0   root
              7 un     un      DET  DET  g=m|n=s      8   det
              8 oxymore -       NC   -     -            6   ats</conll>
    </example>
  </definition>
  ...
</definitions>
...
</pos>

```

Figure 7: A given sense of *mousse* (fem. noun, homograph #1) as a metonym for *bière* ‘bier’

4. a CoNLL (Nivre et al., 2007) output of the Talismane syntactic parser.

The XML version of the textual parts could be used to generate other customized versions of the definitions or the etymology sections. The relevance of some elements is actually task-dependent: markups can be used for example to remove non-textual content (formulae) or unwanted words (foreign words). Links can be used by a weighting scheme in information retrieval (Cutler et al., 1997) or to build hyperlink graphs for semantic similarity computation (Weale et al., 2009). The original format is intended for developers that need specific extractions or conversions. Hathout et al. (2014a) for example, leveraged them to acquire morphological relations.

Phonemic transcriptions. 94% of GLAWI’s entries contain one or several phonemic transcriptions, potentially including diatopic variations. A given transcription may occur at the article level, and therefore correspond to all the forms described in the article. Transcriptions may also appear in POS sections, especially when homographs have different pronunciations. Figure 8 shows two **pos**-sections of two homographs of *plus*, both adverbs (other POSs omitted). The first one, used in affirmative clauses, is a superlative or a comparative pronounced /ply/ or /plys/. The second homograph, used in negative clauses, is pronounced /ply/. In

Figure 9, the transcriptions for *moins*, given at the entry level, indicate that for all parts of speech, *moins* is pronounced /mw̃ɛ/ both in “standard” French (Paris) and /mw̃ɛs/ in Southern France (Marseille, Haut Languedoc).

```

<text>
...
<pos type="adverbe" lemma="1" locution="0" homoNb="1" gracePOS="Rgp">
  <pronunciations>
    <pron>ply</pron>
    <pron>plys</pron>
  </pronunciations>
...
</pos>
<pos type="adverbe" lemma="1" locution="0" homoNb="2" gracePOS="Rgp">
  <pronunciations>
    <pron>ply</pron>
  </pronunciations>
...
</pos>
...
</text>

```

Figure 8: Phonemic transcriptions of *plus*

```

<text>
...
<pronunciations>
  <pron region="Marseille, Haut Languedoc">mw̃ɛs</pron>
  <pron region="Paris, France">mw̃ɛ</pron>
</pronunciations>
...
</text>

```

Figure 9: Phonemic transcriptions of *moins*

3.2 Conversion process: the boundary between standardizing and correcting

As aforementioned, a significant contribution of GLAWI is the standardization of Wiktionnaire’s microstructure⁸ where a given type of information may appear under different forms (predefined templates, aliases, hardcoded text typed by contributors, etc.), and where the same piece of information appearing at different places may lead to inconsistencies. We present two representative examples of consistency checks and standardizing which illustrate the boundary between standardizing and correcting.

⁸ Complementary details on the extraction process required to convert Wiktionnaire’s loosely wiki-encoded data into a structured format can be found in (Hathout et al., 2014a; Navarro et al., 2009; Sajous et al., 2013b).

Linguistic labels. Contributors can use predefined templates to attach linguistic labels to given definitions. Unlike the English Wiktionary where only two templates (`context` and `label`), apparently interchangeable, are used to introduce all the linguistic labels (e.g. `label|dated`), `label|transitive`, `label|oenology`), Wiktionnaire has no generic prefix for these labels: `désuet`, `transitif` and `oenologie`. Detecting linguistic labels in definitions is an important step:

1. to remove them from definitions in order to obtain “clean” text ;
2. to encode the labels into formal markups to ease look-ups (e.g. to target a given label).

Processing the large number of labels used in Wiktionnaire is made even more difficult by their numerous aliases. The diachronic label `vieilli` ‘old’, for instance, also occurs under the forms `vieux` and `vx`. The domain label `oenologie` has three other aliases `œnologie` (ligature), `oenol` and `œnol` (abbreviations). A contributor may also ignore these templates and type the domain name between brackets (*oenologie*) directly in the definition. We inventoried more than 6,000 different labels and aliases used in definitions to normalize the different ways the same information is encoded. As there is no reason to expect that linguistic labels are used in a more relevant (or, at least, coherent) way in Wiktionnaire than in experts-written dictionaries (Baider et al., 2011), we made no attempt to normalize them further. However, we grouped the linguistic labels into categories (*diatopic*, *diachronic*, *attitudinal*, etc.) that are not encoded in Wiktionnaire. A help page⁹ enumerates most of the labels and classifies them into (questionable) categories: *anglicisme*, *germanisme* and *hispanisme* for example, fall into the *registres d’emploi* ‘usage registers’ category, just as *désuet* ‘obsolete’, *rare* ‘rare’ or *enfantin* ‘childish’ do. The label *euphémisme* (euphemism) appears under the category *relations entre les sens* ‘relations between senses’ whereas *dérision* ‘derision’, *mélioratif* ‘meliorative’ and *péjoratif* ‘pejorative’ belong to *registres d’emploi*. This latter category contains the label *informel* ‘informal’ while *soutenu* ‘formal’ belongs to *registres de langue* ‘level of language’. We did not use these categories and decided to manually build coarse-grained ones to which each label can be assigned. Except for the aforementioned normalization of aliases, we did not modify label values and maintained label pairs that look interchangeable. For example, if the difference between *archaïque* ‘archaic’ and *vieilli* ‘old’ is clear, *vieilli* and *désuet* are not clearly distinguished:

- *désuet* = “*pour indiquer que le mot vedette n’est plus employé par la langue moderne*” ‘to indicate that a headword is not used any longer in modern language’
- *vieilli* = “*pour indiquer que le mot vedette est vieilli*” ‘to indicate that a headword is dated’

⁹ http://fr.wiktionary.org/wiki/Wiktionnaire:Liste_de_tous_les_modèles/Précisions_de_sens

Similarly, guidance could be expected to differentiate *littéraire* from *soutenu*, but *littéraire* has no definition and the use of *soutenu* is recommended when the headword belongs to the language level. . . *soutenu*.

Inflectional paradigms. We have described Wiktionnaire’s macrostructure in section 2 and shown the multiple links between the paradigm of a lemma and the corresponding inflected forms. The four inflected forms of the adjective *affluent* (Fig. 1a) are generated by the wiki template `{{fr-accord-cons|a.fly.ã|t}}` (Fig. 1b). Parsing the article dedicated to the form *affluente* (Fig. 1c) confirms that it is the feminine singular form of the adjective *affluent*. However, scattered information is not always redundant: for instance, the gender of the noun *arrivages* ‘arrivals’ is missing in the corresponding page;¹⁰ but the definition indicates that this entry is the plural of *arrivage* ‘arrival’. The masculine gender of *arrivage* being mentioned in its page, we can infer that *arrivages* is masculine too. Unfortunately, contradictory information occurs as well. For example, in the page *clavardeuses*¹¹ (chatters, feminine plural noun in French from Quebec), the gender of the entry is specified as *masculine* whereas the definition states “*Féminin pluriel de clavardeur*”. In such cases, information is left as is and an “*inconsistent*” attribute is added to the GLAWI’s entry (only 65 entries are concerned).

All the inflectional information is propagated in this way and if some features are still missing, we lookup in Leff (Sagot et al., 2006) and Morphalou (Romary et al., 2004) to fill some of the lacks. We used these lexicons to complete GLAWI by adding:

- 366 missing lemmas of inflected forms having full morphosyntactic description in Wiktionnaire;
- 17,446 incomplete morphosyntactic description of inflected forms whose lemma is known;
- 444 genders of nouns or adjectives.

After this last completion, 1.4% of the inflected adjectival forms and 3.7% of the inflected nominal forms still have a missing number or gender (when considering monolexical forms only).

Verb paradigms may be problematic as well: missing inflected forms may be lacking or denote verb defectiveness. Several forms for a given inflection may originate from a superabundant verb, or results from inconsistencies. For example, the conjugation page of *payer*¹² ‘to pay’ gives the two paradigms of this verb. An apparently similar case could explain the two forms *contredisez* and *contredites* of the second person plural of the verb *contredire* ‘to contradict’, imperative mood. The former is the correct form, found in the corresponding page. The latter,

¹⁰ <http://fr.wiktionary.org/w/index.php?title=arrivages&oldid=19099721>

¹¹ <http://fr.wiktionary.org/w/index.php?title=clavardeuses&oldid=19129490>

¹² http://fr.wiktionary.org/wiki/Annexe:Conjugaison_en_français/payer

given in the conjugation table¹³, is erroneous. Another example is given by the two forms *végèterai/végéterai* of the verb *végéter* ‘to vegetate’, first person singular of future indicative, which are neither erroneous nor superabundant. The former is the modern spelling while the latter corresponds to the spelling in use before the 1976 orthographic reform. This latter case is easy to deal with as a specific template identifies the *é/è* alternations due to this reform. In such case, the detected phenomenon is reported into GLAWI by a specific markup. When there is no element to decide whether forms are legitimate or erroneous, we include them all, leaving the opportunity to the users exploiting GLAWI to perform subsequent processing. Handling such cases can also constitute a possible improvement for future versions of GLAWI.

3.3 Next steps

From GLAWI back to Wiktionnaire? GLAWI’s existence is only possible thanks to the contributions of the wiktionarians. Reciprocally, the efforts we made in the standardization and consistency checking process could benefit Wiktionnaire, even if the collaboration between academics and wiktionarians may not be self-evident. Wikis are sometimes presented as knowledge democracy. Hanks (2012) presents Wiktionary as an “*anarcho-syndicalist approach to lexicography*”; Meyer and Gurevych (2012) write that Wiktionary is constructed by a large community of ordinary web users and that the community has a lively discussion culture. In reality, the community only has a small number of *active* contributors who perform most of the contributions: only 117 contributors to Wiktionnaire performed at least five edits in March 2015; 35 of them performed at least 100 edits.¹⁴ These contributors often have responsibility in the management of the dictionary: each wiki project functions as an ecosystem with its administrators, patrollers, functionaries, clerks, bots, etc. There is no denying that discussions may be lively, but they essentially take place among the small world of active contributors. The observation of Wiktionnaire’s discussion pages shows that hours of voluntary work make the contributors quite reluctant to be “dispossessed” from the fruits of their labour. In this context, a newcomer, whether or not a language professional, has to become part of the community before getting credit and fruitfully proposing changes. Anyway, we will not seek to impose standardization or corrections. We take Wiktionary as it is: Wiktionnaire would certainly have attracted fewer contributors if it was more constrained. GLAWI is at the wiktionarians’ disposal, who can use it to reinject information in Wiktionnaire if the community judge it relevant.

Forward synchronization. We previously mentioned Wiktionary’s potential for constant update. We also highlighted that its volatile format makes regular fully-automatic conversions impossible. In order to reflect Wiktionnaire’s up-to-dateness, new versions of GLAWI will be

¹³ http://fr.wiktionary.org/w/index.php?title=Annexe:Conjugaison_en_français/contredire&oldid=8789428

¹⁴ <http://stats.wikimedia.org/wiktionary/EN/TablesRecentTrends.htm>

released in the future. GLAWI update frequency will however not follow the periodicity of XML dumps releases: manual checks have to be performed to ensure that a given parser is still compliant with a new dump. If not, maintenance is required to adapt to format changes.

Other languages. Similarly, due to the format heterogeneity between all language editions, adapting a parser designed for a given language to another one may require heavy changes. Hence, the benefits that can be expected from such work have to be balanced with the size of the targeted language edition and its estimated quality/density. Regarding the size, the number of articles per edition ranges from 45 to more than 4 million¹⁵ and is not necessarily correlated with the number of native speakers: for instance, the second most represented language in Wiktionary is Malagasy while (Mandarin) Chinese ranks sixth.

4. From GLAWI to on demand tailored lexicons

GLAWI has been used to create a number of customized lexicons dedicated to specific uses including NLP, linguistic description and psycholinguistics. The main one is GLÀFF, a large inflectional and phonological lexicon of French. We also derived from GLAWI a morphological derivational resource and a list of people's names.

GLÀFF, a large inflectional and phonological lexicon of French. Collecting the inflectional and phonological information described in GLAWI is quite easy. We just need to traverse the XML file and fill them into the lexicon slots. Since GLAWI provides morphosyntactic tags, we do not even have to parse the inflected words definitions nor the inflectional paradigms of the lemmas. Similarly, GLAWI makes the phonological information available in API with the syllables boundaries. No further processing is needed to fill in the phonological fields in the lexicon.

The extracted lexicon called GLÀFF includes more than 1.4 million entries, each one containing a wordform, a tag in GRACE format, a lemma and, when present in Wiktionnaire, phonemic transcriptions (cf. Fig. 10). Entries also contain word frequencies computed over different corpora.

GLÀFF is by far larger than any other inflectional and/or phonological lexicon of French we know of. Sajous et al. (2013a), Hathout et al. (2014b) and Sajous et al. (2014) compare GLÀFF with four of them¹⁶ and show that it contains three to four times more lemmas and

¹⁵ The number of articles per language edition is given at: https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries

¹⁶ The aforementioned morphological lexicons Leff and Morphalou; Lexique (New, 2006), a free lexicon popular in psycholinguistics, which contains phonemic transcriptions but has a restricted coverage; BDLex (Pérennou and de Calmès, 1987) a non-free lexicon with both an exploitable coverage and phonemic transcriptions.

```

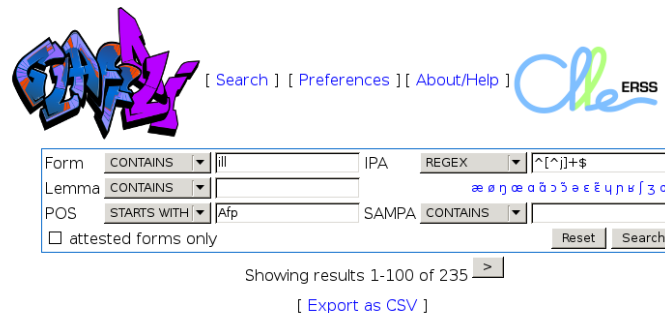
affluent|Ncms|affluent|a.fly.ã|a.fly.A~|22|0.76|38|1.31|232|1.05|444|2.02|1234|0.98|3655|2.91
affluents|Ncmp|affluent|a.fly.ã|a.fly.A~|16|0.55|38|1.31|212|0.96|444|2.02|2421|1.93|3655|2.91
affluent|Vmp3p-|affluer|a.fly|a.fly|9|0.31|187|6.48|369|1.67|1207|5.49|500|0.39|1929|1.53
affluent|Vmsp3p-|affluer|a.fly|a.fly|9|0.31|187|6.48|369|1.67|1207|5.49|500|0.39|1929|1.53

```

Figure 10: Extract of GLÀFF

three to nine times more inflected forms. This size is an important asset when the lexicon is used for research in derivational or inflectional morphology. It is also an advantage for the development of NLP tools such as morphosyntactic taggers and parsers. The comparison also reveals that GLÀFF has a better coverage of the vocabulary of corpora of various types and that it includes many usual words such as: *attractivité* ‘attractivity’, *diabolisation* ‘demonetization’, *homophobie* ‘homophobia’ or *hébergeur* ‘host’, etc. missing from the other lexicons. In addition, GLÀFF’s phonemic transcriptions are highly consistent with those of BDLex and Lexique.

Another interesting feature of GLÀFF is its online browsing interface, called GLÀFFOLI.¹⁷ This interface, illustrated in Figure 11, enables any user to build a multicriteria query. Request fields may include wordform, lemma, part of speech and/or pronunciation. When the user chooses to display corpora frequencies, the wordforms attested in FrWaC are linked to the NoSkecthEngine concordancer (Rychlý, 2007).



Form	POS	Lemma	IPA	SAMPA	Frantext 20 ^e		LM10		FrWaC	
					Form ↓ ↑	Lemma ↓ ↑	Form ↓ ↑	Lemma ↓ ↑	Form ↓ ↑	Lemma ↓ ↑
achilletalonesques	Afppf	achilletalonesque	a.fil.ta.lɔ.nesk	a.Sil.ta.IO~.nEsk	0 0	0 0	0 0	0 0	0 0	0 0
capillaires	Afppf	capillaire	ka.pi.lɛʁ	ka.pi.IER	12 0.415	20 0.693	87 0.395	144 0.655	1123 0.895	3019 2.407
capillotractées	Afppf	capillotracté	ka.pi.lɔ.tʁak.te	ka.pi.IO.tRak.te	0 0	0 0	0 0	0 0	2 0.001	11 0.008
bacillaire	Afpps	bacillaire	ba.si.lɛʁ	ba.si.IER	1 0.034	2 0.069	2 0.009	2 0.009	36 0.028	44 0.035
ancillaire	Afpps	ancillaire	ã.si.lɛʁ	A~.si.IER	10 0.346	25 0.866	10 0.045	25 0.113	66 0.052	128 0.102

Figure 11: GLÀFFOLI, the GLÀFF OnLine Interface

PsychoGLÀFF. GLÀFF has in turn been used to create an even more specific lexicon designed to meet the psycholinguistic needs. Calderone et al. (2014) present PsychoGLÀFF, a version of GLÀFF especially dedicated to the creation and calibration of experimental ma-

¹⁷ <http://redac.univ-tlse2.fr/glaffoli/>

terial that provides a range of additional features of the phonological and written forms such as frequency, lexical neighborhoods, syllabic complexity and phonotactic likelihood.

Extracting derivational relations from GLAWI. GLAWI actually provides information on all aspects of morphology including derivational morphology. Hathout et al. (2014a) present several methods to acquire derivational relations and morpho-semantic knowledge. The first is simply to extract the derivational relations listed in GLAWI's `morphoRel` tags. A second, and more sophisticated method, acquires the relations from the morphological definitions, that is, definitions where the *definiens* contains a word from the morphological family of the *definiendum*. These relations were then further filtered out so that only the ones that can form analogies with the relations listed in `morphoRel` tags were kept. Over all, the derivational resource that resulted from this acquisition contains more than 170,000 relations and is the largest one available for French at the moment.

Human names extraction. Flaux et al. (2014) study the human names that denote a creative activity, such as *symphoniste* (symphonist), *sculpteur* (sculptor) or *romancier* (novelist). Such names have been collected into the NHUMA database¹⁸ from different sources such as a language dictionary (TLFi), a dictionary of synonyms (DicoSyn) and WaliM (Namer, 2003), a tool for harvesting the web. After these resources have been exploited, a simple lookup in GLAWI's glosses, based on lexical cues only, enabled a 15% increase of the database.

Other possibilities. Filtering GLAWI's linguistic labels or other markups instantly permits on demand tailoring of lexicons such as loanwords used in French, masculine/feminine noun equivalents, dated words, domain-specific sublexicons, etc. Regarding lexicography, an immediate application could be the use of GLAWI for neology monitoring. Automatic detection of neologisms in corpora produces a lot of noise. GLAWI can be used to detect true positives among the candidates. When a form extracted from a corpus is absent from the reference lexicon, its occurrence in GLAWI is a serious hint of actual neology.

5. Conclusion and perspectives

This paper introduces GLAWI, an XML-encoded MRD automatically extracted from Wiktionnaire. Therefore, GLAWI inherits most of Wiktionnaire's strong points, including the exceptional number of its headwords and an original macrostructure. This has been assessed through detailed comparisons with well-known inflectional and phonological lexicons.

Wiktionnaire's editorial success is linked to its use of MediaWiki which imposes no constraint on how information is represented. The flip side is the great heterogeneity of its

¹⁸ <http://nomsdhumains.weebly.com>

microstructure which makes it difficult to use in NLP and prevents the selection of articles with targeted queries such as “I am looking for particle nouns ending in *-on*” like *neutron*, *gluon* or *boson*. GLAWI specifically addresses these needs: the XML markups encode the microstructure explicitly; it standardizes the Wiktionnaire’s content and enhances its coherence, standardization being clearly a prerequisite to any automated exploitation.

GLAWI is also an answer to other needs, like the creation of specific lexical resources. Indeed, it is likely that the development of the mobile web is changing the way users access MRDs. Complex interfaces like the one of the *Trésor de la Langue Française informatisé* (TLFi), a large French MRD (Dendien, 1994), are losing ground in favor of applications built around specific information subsets such as thesauri, quotation, slang, rhyming, etymological or bilingual dictionaries, but also less traditional derivative works like dictionaries of Latin loanwords, morphological dictionaries or dictionaries of epicene nouns. However, the need to access dictionaries through targeted queries remains, particularly for skilled users (Lew, 2013) and for language specialists, especially linguists and lexicographers. To this end, we plan to design a user-friendly interface for GLAWI, similar to GLÀFFOLI (see Figure 11).

Another remarkable feature GLAWI inherits from Wiktionnaire is its free license which makes it a resource adapted to current research practice in NLP. NLP is indeed becoming a discipline where experimentation occupies an increasingly important place and where experiment replication is becoming common. One consequence of this development is the requirement to use freely available resources and data sets. GLAWI fulfills this condition but similar resources for French are in short supply as traditionally, researchers and labs greatly restrict the access to the data they produce. Notable exceptions are Lefff, an inflectional lexicon used by several taggers, Lexique, until recently the only free resource including phonemic transcriptions and Flexique (Bonami et al., 2014), produced by semi-automatically filling the paradigms of Lexique’s entries. Notice however that there is no satisfactory resource providing definitions. TLFi is not available for download and, according to Eckard et al. (2012), WOLF (Sagot and Fišer, 2008), a free French WordNet built automatically by aggregating and translating other resources, is sparse and not completely translated. The lack of free satisfactory lexical resources does not only impact research. It is also an impediment to the development of language processing applications. The long-term survival of dictionaries is questioned by Rundell (2012), who envisages that their heterogeneous functions might be better performed by separate specialized tools. If this happens, such tools, while contributing to the disappearance of dictionaries in their current forms, will still necessitate lexical knowledge embedded in electronic dictionaries. GLAWI could meet such needs.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments. Syntactic parsing has been performed using the OSIRIM platform that is administered by IRIT and supported by CNRS, the Region Midi-Pyrénées, the French Government and ERDF.

7. References

- Baider, F., Lamprou, E., & Monville-Burston, M. (2011). *La marque en lexicographie: états présents, voies d'avenir*. La lexicothèque. Lambert-Lucas.
- Bonami, O., Caron, G., & Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *Actes du 4^e Congrès Mondial de Linguistique Française (CMLF 2014)*. Berlin, pp. 2583–2596.
- Calderone, B., Hathout, N., & Sajous, F. (2014). From GLÀFF to PsychoGLÀFF: a large psycholinguistics-oriented French lexical resource. In *Proceedings of the 16th EURALEX International Congress*, Bolzano, pp. 431–446.
- Calzolari, N. (1988). The dictionary and the thesaurus can be combined. In Evens, M., editor, *Relational Models of the Lexicon*. Cambridge University Press, pp. 75–96.
- Chodorow, M. S., Byrd, R. J., & Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting on Association for Computational Linguistics*, ACL'85, Chicago, pp. 299–304.
- Cutler, M., Shih, Y., & Meng, W. (1997). Using the Structure of HTML Documents to Improve Retrieval. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Monterey, pp. 241–252.
- Dendien, J. (1994). Le projet d'informatisation du TLF. In Éveline Martin, editor, *Les textes et l'informatique*, chapter 3. Didier Érudition, Paris, France, pp. 31–63.
- Eckard, E., Barque, L., Nasr, A., & Sagot, B. (2012). Dictionary-Ontology Cross-Enrichment. Using TLFi and WOLF to enrich one another. In *COLING Workshop on Cognitive Aspects of the Lexicon*, Mumbai, pp. 81–93.
- Flaux, N., Lagae, V., & Stosic, D. (2014). Romancier, symphoniste, sculpteur : les noms d'humains créateurs d'objets idéaux. In *Actes du 4^{eme} Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, pp. 3075–3089.
- Hanks, P. (2012). Corpus evidence and electronic lexicography. In Granger, S. & Paquot, M., editors, *Electronic Lexicography*, chapter 4. Oxford University Press, Oxford, pp. 57–82.
- Hathout, N. (2011). Morphonette: a paradigm-based morphological network. *Lingue e linguaggio*, 2011(2), pp. 243–262.
- Hathout, N. & Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5), pp. 125–168.
- Hathout, N., Sajous, F., & Calderone, B. (2014a). Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary. In *Proceedings of the*

- COLING Workshop on Lexical and Grammatical Resources for Language Processing*, Dublin, pp. 65–74.
- Hathout, N., Sajous, F., & Calderone, B. (2014b). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, pp. 1007–1012.
- Kosem, I., Gantar, P., & Krek, S. (2013). Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In *Proceedings of eLex 2013*, Tallinn, pp. 32–48.
- Lew, R. (2013). Online dictionary skills. In *Proceedings of eLex 2013*, Tallinn, pp. 16–31.
- Markowitz, J., Ahlswede, T., & Evens, M. (1986). Semantically significant patterns in dictionary definitions. In *Proceedings of the 24th Annual Meeting on Association for Computational Linguistics*, New York, pp. 112–119.
- Meyer, C. M. (2013). *Wiktionary: The Metalexigraphic and the Natural Language Processing Perspective*. PhD thesis, Technische Universität Darmstadt.
- Meyer, C. M. & Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Granger, S. & Paquot, M., editors, *Electronic Lexicography*, chapter 13, Oxford University Press, Oxford, pp. 259–291.
- Namer, F. (2003). WaliM : valider les unités morphologiquement complexes par le web. In Fradin, B., Dal, G., Kerleroux, F., Hathout, N., Plénat, M., & Roché, M., editors, *Les unités morphologiques. Actes du 3ème Forum de Morphologie.*, Lille, pp. 142–150.
- Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., & Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the 2009 ACL-IJCNLP Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, Singapore, pp. 19–27.
- New, B. (2006). Lexique 3 : Une nouvelle base de données lexicales. In *Verbum ex machina. Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2006)*, Louvain-la-Neuve, pp. 892–900.
- Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL 2007 Shared Task on dependency parsing (EMNLP-CoNLL)*, Prague, pp. 915–932.
- Penta, D. J. (2011). The wiki-fication of the dictionary: defining lexicography in the digital age. In *Proceedings of the MIT7 Conference "unstable platforms: the promise and peril of transition"*, Cambridge.
- Pérennou, G. & de Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of the European Conference on Speech Technology, ECST 1987*, Edinburgh, pp. 1393–1396.
- Rajman, M., Lecomte, J., & Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF.

- Romary, L., Salmon-Alt, S., & Francopoulo, G. (2004). Standards going concrete : from LMF to Morphalou. In *Proceedings of COLING 2004: Enhancing and using electronic dictionaries*, Geneva, pp. 22–28.
- Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? In Meunier, F., De Cock, S., Gilquin, G., & Paquot, M., editors, *A Taste for Corpora. In honour of Sylviane Granger*, John Benjamins, pp. 257–282.
- Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In *Proceedings of the 15th EURALEX International Congress*, Oslo, pp. 47–92.
- Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, pp. 65–70.
- Sagot, B., Clément, L., De La Clergerie, E., & Boullier, P. (2006). The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, pp. 1348–1351.
- Sagot, B. & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of OntoLex 2008*, Marrakech.
- Sajous, F., Navarro, E., Gaume, B., Prévot, L., & Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., & Helgadóttir, S., editors, *Advances in Natural Language Processing*, volume 6233 of *LNCS*, Springer Berlin / Heidelberg, pp. 332–344.
- Sajous, F., Hathout, N., & Calderone, B. (2013a). GLÀFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, pp. 285–298.
- Sajous, F., Navarro, E., Gaume, B., Prévot, L., & Chudy, Y. (2013b). Semi-automatic Enrichment of Crowdsourced Synonymy Networks: the WISIGOTH System Applied to Wiktionary. *Language Resources and Evaluation, special issue on Collaboratively Constructed Language Resources*, pp. 1–34.
- Sajous, F., Hathout, N., & Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oripeau du web ! Études et réalisations fondées sur le dictionnaire collaboratif. In *Actes du 4e Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, pp. 663–680.
- Sérasset, G. (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, pp. 2466–2472.
- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II-Le Mirail.
- Weale, T., Brew, C., & Fosler-Lussier, E. (2009). Using the Wiktionary Graph Structure for Synonym Detection. In *Proceedings of the ACL-IJCNLP Workshop on The People's*

Web Meets NLP: Collaboratively Constructed Semantic Resources, Singapore, pp. 28–31.

Zesch, T. & Gurevych, I. (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words. *Journal of Natural Language Engineering.*, 16(01), pp. 25–59.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Using machine learning for language and structure annotation in an 18th century dictionary

Petra Bago, Nikola Ljubešić

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences,
University of Zagreb, Ivana Lučića 3, HR-10000
{pbago, nljubesi}@ffzg.hr

Abstract

The accessibility of digitized historical texts is increasing, which, consequently, has resulted in a growing interest in applying machine learning methods to enrich this type of content. The need for applying machine learning is even greater than in modern texts given the high level of inconsistency in historical texts even within the same document. In this paper we investigate the application of a supervised structural machine learning method on language and structure annotation of 18th century dictionary entries. Our research is conducted on the first volume of a trilingual dictionary ‘Dizionario italiano–latino–illirico’ (Italian–Latin–Croatian Dictionary) compiled by Ardellio della Bella and printed in Dubrovnik in 1785. We assume that by using this method, we can significantly reduce time for manual annotation and simplify the process for the annotators. We reach accuracy of approximately 98% for language annotation and around 96% for structure annotation. A final experiment on the time gain obtained by pre-annotating the data shows that only correcting the generated labels is roughly five times faster than full manual annotation.

Keywords: historical dictionaries; language annotation; structure annotation; supervised machine learning

1. Introduction

The accessibility of digitized historical texts is increasing, which, consequently, has resulted in a growing interest in applying natural language processing and machine learning methods for processing and enriching this type of content. Using these methods, some of the problems approached are mapping historical spelling variants to modern equivalents (Archer et al., 2015), identifying and extracting mentions of times present in historical resources (Foley and Allan, 2015), improving verb phrase extraction (Pettersson and Nivre, 2015) or developing a web-based application for editing manuscripts (Raaf, 2015). The need for applying machine learning is even greater than in modern texts given the high level of inconsistency in historical texts even within the same document (Piotrowski, 2012). In this paper we investigate the application of a supervised structural machine learning method on language and structure annotation of 18th century dictionary entries.

Our research is conducted on the first volume of a second edition of a trilingual dictionary ‘Dizionario italiano–latino–illirico’ (Italian–Latin–Croatian Dictionary) compiled by Ardellio della Bella and printed in Dubrovnik in 1785 (della Bella, 1785). The dictionary was intended for Italian Jesuit

missionaries to help them spread the faith in a national language i.e. Croatian language, but also other Slavic languages. For this reason a Croatian grammar can be found inside the dictionary preamble. The dictionary consists of 899 pages and two parts. The first part is a preamble written in Italian on 54 pages. The second part is the dictionary, containing around 19,000 headwords. The dictionary is printed in two volumes: the first volume contains the preamble and the dictionary part from letters A to H, while the second volume contains the dictionary part from letters I to Z. For the first time in Croatian lexicography, della Bella’s dictionary contains examples of uses of headwords in various literary works and oral literature.

In the paper we approach two separate annotation, i.e. enrichment problems, using the state-of-the-art supervised machine learning algorithm for labeling sequences – conditional random fields (CRFs). We first approach the problem of annotating each token with its corresponding language label which is a ternary classification task given the three languages that are represented in the dictionary. Having the language label at our disposal, we then approach the problem of annotating each token with the corresponding structure label. The structure level has 19 different labels based on the Text Encoding Initiative (TEI) encoding scheme for dictionaries (TEI Consortium, 2014).

We approach both annotation problems by determining first whether the original or lowercased tokens produce better results, defining that feature as our basic feature. Next, we measure the performance of adding several other features to the basic one like whether the token is originally lowercased, the frequency of a specific token trigraph, the previous and the next token, whether the previous and the next token is lowercased, etc. Finally, we combine all features that show increase over the results obtained with the basic feature.

2. Related work

Historical texts are written in historical languages that are natural languages, just like the modern languages found in modern texts. Consequently, both historical and modern languages share the same challenges when it comes to natural language processing (NLP) of these types of texts, such as homonymy and polysemy. However, historical texts have further characteristics that pose additional challenges to NLP tools trained on modern texts: the lack of a standard variant, the lack of a standard orthography, the lack of electronically available texts, and the lack of existing NLP resources and tools for this type of text (Piotrowski, 2012).

Nevertheless, machine learning methods have been applied to historical texts approaching various problems. (Buchler et al., 2014) address the issue of complication to historical text-reuse detection, because of its longer time span, thereby having a larger set of morphological, linguistic, syntactic, semantic and copying variations. (Mitankin et al., 2014) present an approach to historical text normalisation, achieving 81.79% normalisation accuracy of 17th century English texts in a fully unsupervised setup. Furthermore, (Kettunen et al., 2014) experimented with methods based on corpus statistics, language technology and machine learning in order to find ways to automate

the process of analyzing and improving the quality of a historical news collection. (Horton et al., 2009) trained a supervised machine learning algorithm to determine classes of knowledge of the articles in the the Encyclopédie of Denis Diderot and Jean le Rond d’Alembert. (Hendrickx et al., 2011) presented an approach to automatic text segmentation of historical letters in Portuguese in formal/informal parts using a statistical n-gram based technique, achieving the result of 86% micro-averaged F-score. Additionally, they presented an approach to semantic labeling of the formal parts of the letters using supervised machine learning, achieving the result of 66.3% micro-averaged F-score.

In the paper we approach two separate annotation, i.e. enrichment problems, using the state-of-the-art supervised machine learning algorithm for labeling sequences – conditional random fields (CRFs). Conditional random fields (CRFs) are a statistical method for structure prediction, that has the ability to predict labels based on several dependent variables. The models are applied to image labeling, e.g. (He et al., 2004), (Kumar and Hebert, 2003), various bioinformatics problems, e.g. (Sato and Sakakibara, 2005), (Liu et al., 2005), speech processing, e.g. (Yu et al., 2010), (Boonsuk et al., 2014), and, the most relevant to the paper, textual data, e.g. (Sha and Pereira, 2003), (McCallum and Li, 2003), (Taskar et al., 2002), (Pinto et al., 2003), (Shen et al., 2007), (Choi et al., 2005).

In digital humanities, annotating the structure of a digitized text is a manual task, that is time consuming and tedious, thereby paving the way for an annotator to introduce inconsistencies. By automating the process of annotation, we consider it to reduce cognitive load in annotators and time spent on the task. As far as we know, the present work is the first to apply conditional random fields on a historical text. Additionally, we have not come across an application of CRFs on language labeling on textual data, nor on structure labeling based on a *de facto* standard for encoding textual resources in digital form.

3. Dataset

Our research is conducted on the first volume of the second edition of a trilingual dictionary ‘Dizionario italiano–latino–illirico’ (Italian–Latin–Croatian Dictionary) compiled by Ardellio della Bella and printed in Dubrovnik in 1785 (della Bella, 1785). The digitization process of the printed 18th dictionary was conducted as part of the project ‘Croatian dictionary heritage and Croatian European identity’ and was not the scope of this research. However, we will briefly describe the digitization process in order to better describe the data used in this research. The dictionary was photographed and the images were processed with an optical character recognition software. Since the software produced many errors detecting characters, the text was manually compared and checked to corresponding pictures by undergraduate students. Furthermore, during the manual inspection, markup was added for distinct section breaks such as line breaks, new paragraphs, column breaks, and page breaks. Additional markup was manually inserted to encode the beginning and the end of the Latin parts of the entry, and the beginning and the end of the citations from works used as a corpus for dictionary compilation by della Bella. The manual part of the digitization process is the

most tedious and time-consuming. Aforementioned text is stored in a proprietary word processor that we converted into a plain text file for further processing.

The first volume of the trilingual dictionary consists of 7,972 dictionary entries starting with the letter A and ending with the letter H (Huquang), comprising 403,128 tokens that were automatically segmented. The average length of the dictionary entry is 50.57 tokens.

Following the tokenization phase, for our training sample we randomly selected 101 dictionary entries for manual annotation. The training sample comprises of 8,340 tokens (2,07%), while the unlabelled set contains 394,788 tokens (97,93%).

Every token out of the selected entries is annotated on two levels: the language level and the structure level. The language level has three distinct labels, while the structure level has 19. Label distributions of both levels are depicted in Tables 1 and 2. Altogether 8,340 labels are manually annotated on each level, that is 16,680 labels in total. The average length of the selected entries is 82.57 token, i.e. 32 tokens more than the average entry of the first volume of the dictionary.

There are three labels of the language level based on three languages that can be found in della Bella’s dictionary. The labels used for the language annotation, its explanation and frequency distribution are given in Table 1.

label	explanation	frequency
hr	a token in Croatian	4,395
it	a token in Italian	2,164
la	a token in Latin	1,781

Table 1: The labels used for the language annotation, its explanation and frequency distribution

In Table 1 it is interesting to note that more than half (53%) of the tokens are in Croatian language, while Italian is more frequent than Latin (26% vs. 21%). This can be interpreted as the lexicographer’s attempt to include all possible words with similar senses in the Croatian language, while for the Latin language there can usually be found only one word sense, probably because of the similarity between Italian and Latin.

There are 19 labels of the structure level that are based on the Text Encoding Initiative module for dictionaries (TEI Consortium, 2014). The labels used for the structure annotation, its explanation and frequency distribution are given in Table 2.

We perform two separate annotation problems: the problem of annotating each token with the corresponding language label and the problem of annotating each token with the corresponding structure label, having at that point the language label at our disposal.

label	explanation	frequency
abbr	an abbreviation	55
adj	a suffix for an adjective	2
adjf	a suffix for a feminine singular adjective	109
adjn	a suffix for a neuter singular adjective	118
bibl	a source of citation	90
cb	a column break when it is not separating a token ¹	7
citex	a citation	729
cittrans	a translation of the headword or another word within the dictionary entry	3,167
formlem	a headword ²	125
genpl	a suffix for a genitive plural noun	1
gensg	a suffix for a genitive singular noun	185
hint	a token that guides the sense of the headword or another word within the dictionary entry	415
lb	a line break when it is within one entry and does not separate a token ³	506
pb	a page break when it is not within one token ⁴	6
pc	a punctuation character that is not part of an abbreviation	2,329
pos	a part of speech (masculine, feminine and neuter gender of a noun, plural if a noun is in that form, adjective, adverb)	198
ref	a reference to another entry	35
v	a suffix for a verb form, usually first person singular present and first person singular perfect	230
xr	a token for a cross-reference phrase	33

Table 2: The labels used for the structure annotation, its explanation and frequency distribution

4. Experimental setup

In our experiment we use state-of-the-art supervised machine learning algorithm for labeling sequences named conditional random fields (CRFs) (Lafferty et al., 2001). CRFs are a statistical method for structure prediction, that has the ability to predict labels based on several dependent variables. These models are successfully applied in different fields, such as text processing, bioinformatics and computer vision (Sutton and McCallum, 2012).

We train and evaluate CRFs with the `CRFsuite tool` (Okazaki, 2007). The tool implements several different state-of-the-art methods of machine learning and we use the passive aggressive training algorithm since it obtained the best results. The software has features like fast training and tagging data, simple data format and the ability to design an arbitrary number of features for each item. Additionally the tool has the ability to compute performance evaluation of the model evaluated on test set (precision, recall and F_1 scores).

We perform two separate annotation problems: the problem of annotating each token with the corresponding language label and the problem of annotating each token with the corresponding structure label, having at that point the language label at our disposal. Our approach to both problems is similar. Firstly we define potentially interesting sets of features that could obtain better results than the data alone. Next we measure performance of the selected features. Finally, we combine all features that show an increase over the result obtained with the basic feature thereby achieving the best possible result with the defined features. We compute the usual metrics used for model evaluation in the field of natural language processing: precision, recall, F measure and accuracy.

Our experiment is conducted in three phases. The first phase consists of testing the most obvious feature, i.e. does the spelling of the token have an effect on the result: original spelling of the token and lowercased spelling of the token. We expect that one of the forms of spelling will yield a better result. Consequently we will be using the feature that achieved better results as the basic feature in further testing.

In the second phase of the experiment we test the effect of additional features on the results of machine learning. As the basic feature we use the one from the first phase of the experiment. On the language level as additional feature we measure a Boolean variable of whether the original token is lowercased or not. Next we measure the frequency of a specific trigraph. Furthermore we test the effect of N tokens before and after the specific token, for N ranging from 1 to 3. The final tested measure is a Boolean variable of whether tokens before and after are lowercased or not.

On the structure level as an additional feature we measure a Boolean variable of whether the original token is lowecased or not. Next we test the effect of N tokens before and after the specific token, for N ranging from one to four. Furthermore, we measure a Boolean variable of whether tokens before and after are lowercased or not. Since the dataset for this phase contains data about the language of the token, we test the effect of that feature on the results.

In the final phase of the experiment, we combine in one experiment all features that show an increase over the result obtained with the basic feature. Thereby we achieve the best possible result with the defined features.

To estimate how accurately our predictive model will perform on an independent dataset, we evaluate each parameter by calculating accuracy via a 10-fold cross-validation.

5. Results

5.1 The language annotation

The language annotation has a set of three labels. The experiment is conducted with the following features:

- **token**: a token in its original form,
- **ltoken**: lowercased token,
- **lcasebool**: a Boolean variable whether a token is lowercased or not,
- **trigraphfreq**: a frequency of a specific trigraph,
- **prevNtoken** and **nextNtoken**: N tokens before and after a specific token, for $N = 1..3$,
- **prevNlcasebool** and **nextNlcasebool**: a Boolean variable whether N tokens before and after are lowercased.

Below we depict 7 tokens labelled on both the language and the structure level:

```
radici  it hint
.       it pc
V.      it xr
Barbare it ref
.       it pc
Radicare it ref
.       it pc
```

The feature values for the token `Barbare` of the abovementioned sequence are as follows:

```
token=Barbare
ltoken=barbare
lcasebool=False
trigraphfreq=_ba:1
trigraphfreq=bar:2
trigraphfreq=arb:1
trigraphfreq=rba:1
trigraphfreq=are:1
trigraphfreq=re_:1
prev1token=V.
prev2token=.
prev3token=radici
next1token=.
next2token=radicare
next3token=.
prev1lcasebool=False
prev2lcasebool=True
prev3lcasebool=True
next1lcasebool=True
next2lcasebool=True
next3lcasebool=True
```

The results of the accuracy of the language annotation with specific features are given in Table 3. Since lowercased tokens perform better than the original ones, the remainder of the experiments use the lowercased tokens as the basic feature.

Additionally, the most informative features are token trigrams and tokens before and after the specific token. Using two tokens before and after a specific token gives slightly better results than using just one or three tokens before and after. This is why in the last parameter we combine the best performing features: lowercased tokens, token trigrams, a Boolean variable whether a token is lowercased or not, a Boolean variable whether tokens before and after are lowercased, and two tokens before and after a specific token. This selected feature set obtains the best results, i.e. the accuracy of the language annotation of 98.413%.

features	accuracy
token	0.93224
ltoken	0.94143
ltoken lcasebool	0.95405
ltoken trigraph	0.97107
ltoken prevNtoken nextNtoken N=1	0.97188
ltoken prevNtoken nextNtoken N=1..2	0.97997
ltoken prevNtoken nextNtoken N=1..3	0.97697
ltoken prevNlcasebool nextNlcasebool N=1	0.94475
ltoken prevNlcasebool nextNlcasebool N=1..2	0.95086
ltoken prevNlcasebool nextNlcasebool N=1..3	0.94142
ltoken lcasebool trigraph prevNtoken nextNtoken prevNlcasebool nextNlcasebool N=1..2	0.98413

Table 3: The accuracy of language annotation with various features

Table 4 gives the results of precision, recall and F_1 measure of the final language classifier by category. The classifier obtains the best results for the Latin language for all three measures: a precision (P) score of 0.99815, a recall (R) score of 0.99938, and an F_1 score of 0.99878. Since the Latin part of the dictionary entries is always wrapped in special markup, the results are expected. The classifier accomplishes better precision scores for the Italian language (0.9829) than for Croatian (0.97953). The reason for this could be due to the fact that the beginning of a dictionary entry is always in Italian. Better results of the recall scores are obtained for the Croatian language (0.99067) than for Italian (0.95831), which can be interpreted by the fact that over half (53%) of the tokens are labelled as Croatian, but just over one quarter (26%) as Italian.

lang	Precision	Recall	F_1
hr	0.97953	0.99067	0.98507
it	0.9829	0.95831	0.97045
la	0.99815	0.99938	0.99878

Table 4: The performance of the final language classifier by category

5.2 The structure annotation

The structure annotation has a set of 19 labels. The experiment on the structure level follows the same methodology as for the language level. The experiment is conducted with following features:

- **token**: a token in its original form,
- **ltoken**: lowercased token,
- **lcasebool**: a Boolean variable whether a token is lowercased or not,
- **prevNtoken** and **nextNtoken**: N tokens before and after a specific token, for $N = 1..4$,
- **prevNlcasebool** and **nextNlcasebool**: a Boolean variable whether N tokens before and after are lowercased.
- **lang**: a language label of the token,
- **suffixN**: a suffix of a specific token of length $N=4$.

The results of the accuracy of the structure annotation with specific features are given in Table 5. Since tokens in its original form perform better than lowercased tokens, the remainder of the experiment uses the original form of tokens as the basic feature.

features	accuracy
token	0.85993
ltoken	0.85538
token lcasebool	0.8934
token prevNtoken nextNtoken N=1	0.90388
token prevNtoken nextNtoken N=1..2	0.93794
token prevNtoken nextNtoken N=1..3	0.94994
token prevNtoken nextNtoken N=1..4	0.94219
token prevNlcasebool nextNlcasebool N=1	0.87586
token prevNlcasebool nextNlcasebool N=1..2	0.87706
token prevNlcasebool nextNlcasebool N=1..3	0.88755
token prevNlcasebool nextNlcasebool N=1..4	0.89588
token lang	0.86555
token suffixN N=1..4	0.87192
token lcasebool lang prevNtoken nextNtoken prevNlcasebool nextNlcasebool suffixN N=1..4	0.96111
token lcasebool prevNtoken nextNtoken N=1..3 prevNlcasebool nextNlcasebool suffixN N=1..4	0.96372

Table 5: The accuracy of the structure annotation with various features

Additionally, the most informative feature is four tokens before and after a specific token. However, when we combine the best performing features, the accuracy score increases almost 2% and totals 0.96372. Those features are: tokens in their original form, a Boolean variable whether a token is lowercased or not, four tokens before and after a specific token, a Boolean variable whether four tokens before and after a specific token are lowercased or not, a language label, and a suffix of a specific token of length N=1..4.

Table 6 gives the results of precision, recall and F_1 measure of the final structural classifier by category. The classifier obtains 100% precision for column breaks and line breaks, which is expected since these properties are explicitly tagged in the dictionary corpus. The next best accuracy score is 0.9981 for punctuation characters. Since in the dictionary corpus there is always a space before a punctuation character that is not part of an abbreviation, this result is likewise expected. The worst results obtained by the classifier are for labels that are rare in the manually annotated corpus. There is only one occurrence of the label `genp1`, and the precision score is 0.0. The same result is obtained for the label `adj`, that has only two occurrences. The third worse result (0.6) is obtained for the label `pb`, that has only six occurrences in the manually annotated corpus.

The classifier obtains 100% recall for the label `1b`, while the second best result (0.99762) is for the label `pc`. Both results can be interpreted as with the precision. The label `1b` is explicitly tagged in the dictionary corpus, while there is always a space before punctuation character that is not part of an abbreviation. The classifier obtains the third best result (0.99087) for the label `v` and the reason for this could be the fact that this label refers to the suffixes for verbs that regularly have the same form. The worst results obtained by the classifier are for the labels `genp1` and `adj`, like with the

precision scores, because the labels rarely occur in the manually annotated corpus. The recall score for the label **cb** is surprising and only totals to 0.57143. The column break is explicitly tagged in the dictionary corpus in two ways: it can be a standalone tag, but it can also be found within a token, where it is left as part of that token, and not separately tokenized. The assumption is that the tag within a token generates obstacles for the classifier to obtain higher recall score.

The classifier obtains the top three results for the F_1 measure for the labels **lb** (1.0), **pc** (0.99786) and **v** (0.9819). If observing all three measures combined, the classifier obtains the best result for the label **lb**, while the label **pc** is in top three results for all three measures. On the contrary, the worst results are obtained for the labels **adj**, **genpl** and **pb**, on account of the labels rarely occurring in the manually annotated corpus.

lang	Precision	Recall	F_1
abbr	0.85714	0.78261	0.81818
adj	0.0	0.0	0.0
adjf	0.97196	0.99048	0.98113
adjn	0.94595	0.92105	0.93333
bibl	0.95122	0.98734	0.96894
cb	1.0	0.57143	0.72727
citex	0.95477	0.95323	0.954
cittrans	0.97875	0.95736	0.96794
formlem	0.97087	0.9009	0.93458
genpl	0.0	0.0	0.0
gensg	0.97093	0.98235	0.97661
hint	0.76027	0.91484	0.83042
lb	1.0	1.0	1.0
pb	0.6	0.6	0.6
pc	0.9981	0.99762	0.99786
pos	0.97297	0.98361	0.97826
ref	0.96875	0.91176	0.93939
v	0.97309	0.99087	0.9819
xr	0.96875	0.96875	0.96875

Table 6: The performance of the final structural classifier by category

5.3 Testing the time reduction for the manual annotation

Our next experiment answers the question whether correcting automatically assigning language and structure labels reduces the time for the manual annotation, and if confirmed, by how much. The experiment has two 60-minute parts: a manual token annotation and a correction of automatically labelled tokens. Both parts are carried out by an annotator knowledgeable of della Bella’s dictionary. The results of the experiment are given in Table 7.

In the first part of this experiment, an annotator manually annotates tokens on the language and structure level for 60 minutes. The starting token is randomly chosen, after which the tokens are annotated in the order of their appearance in the corpus. During this period 741 tokens (i.e. 482 labels) are annotated. In one minute, 12.35 tokens can be manually annotated.

	number of tokens	tokens per minute
manual annotation	741	12.35
correction	3,439	57.32

Table 7: The number of tokens manually annotated and corrected

In the second part of this experiment, an annotator reviews and corrects the automatic labels on the language and structure level for 60 minutes. The starting token is randomly chosen, after which the tokens are reviewed and corrected in the order of their appearance in the corpus. During this period 3,439 tokens (i.e. 6,878 labels) are reviewed and corrected. In one minute, 57.32 tokens can be reviewed and corrected: specifically this method is 4.64 times faster than manual annotation, which we consider clearly more productive than the manual annotation.

Additional value of this experiment is 7,987⁵ tokens subsequently annotated or reviewed and corrected that can be incorporated into the training set, thereby possibly obtaining better accuracy scores with the classifier and yet further reducing the time for correction speed.

5.4 The final experiment on the test set

To closely analyse the performance of the classifier, we present the confusion matrices for the language level in Table 8 and the structure level in Table 9. The test set is the result of the experiment in the previous section.

The accuracy of the classifier for the language level is 0.97308. In the confusion matrix given in Table 8 it is evident that the classifier displays fewer problems with predicting the Latin text. Since the Latin part of dictionary entries is always wrapped in special markup, the results are expected. The classifier has the most problems with the Croatian–Italian language pair. The dictionary entries often do not follow the structure of a trilingual dictionary, thus the sequence of the languages appearing is not always Italian–Latin–Croatian. As mentioned before, the Latin part is always wrapped in special markup, which would be a great separator of the Italian from the Croatian. However, if there is a compound within an entry, then the Latin part is frequently absent, which creates a situation where the Croatian part follows the Italian part. Additionally, at the time the dictionary was created, there was no consensus over orthography for the Croatian language, so the lexicographer adopted the Italian practice to record Croatian phonemes. However, this practice introduces inconsistency in orthography within dictionary text. All of this could be the reason why the classifier has the most problems with the Croatian–Italian pair.

The accuracy of the classifier for the structure level is 0.954801. In the confusion matrix given in Table 9 it is evident that the classifier has the most problems with the label `cittrans`, and confuses it most frequently with the labels `hint`, `citex` and `v`. The reason behind this may be the fact that these parts of the entries contain free text. The classifier obtains the best results for the label `xr`,

⁵ The second part of this experiment had to be repeated 3 times due to the fatigue of the annotator. This is the reason this number is larger than the sum of the tokens in the first and the second part of this experiment.

	hr	it	la
hr	5,128	19	1
it	194	1,351	1
la	0	0	1,293
accuracy	0.973081257043		

Table 8: The confusion matrix for the language level

which is correctly predicted in all the cases, and for the labels **lb** and **bibl** that are only once incorrectly classified. Since these parts are explicitly tagged in the dictionary corpus, the results are expected. Three labels are not found in the test set: **cb**, **pb** and **adj**.

	cit	trans	ref	lb	bibl	hint	cb	v	pos	pb	pc	abbr	citex	adjn	xr	gensg	form	lem	adj	adjf
cit trans	2,129	0	0	0	18	0	3	0	0	4	0	19	2	0	1	0	0	3		
ref	17	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
lb	0	0	506	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bibl	2	0	0	80	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
hint	70	0	0	0	173	0	0	0	0	0	0	9	2	0	2	6	0	0	0	0
cb	5	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
v	32	0	0	0	5	0	568	0	0	0	0	0	19	0	1	0	0	2		
pos	2	0	0	1	0	0	0	253	0	1	1	0	0	0	0	0	0	0	0	0
pb	2	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
pc	0	0	0	0	1	0	0	0	0	2,618	0	1	0	0	0	0	0	0	0	0
abbr	2	0	0	0	4	0	0	2	0	2	19	0	0	0	0	0	0	0	0	0
citex	48	0	0	0	2	0	0	0	0	0	0	556	0	0	1	0	0	0	0	0
adjn	2	0	0	0	2	0	1	10	0	0	0	0	124	0	2	0	0	1	0	0
xr	4	0	0	0	0	0	0	0	0	0	4	0	0	44	0	0	0	0	0	0
gensg	5	0	0	0	0	0	1	0	0	0	0	1	1	0	245	0	0	0	0	0
form lem	2	0	0	0	13	0	1	0	0	1	0	0	1	0	0	135	0	0	0	0
adj	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
adjf	1	0	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	133		
accuracy	0.954801552523																			

Table 9: The confusion matrix for the structure level

5.5 The learning curve

The learning curve of the used algorithm is given in Figure 1. With regards to the language level having only three labels, while the structure level has 19, we expect the algorithm to generate better results for the former level than for the latter. Moreover, we expect that less data would be necessary for the algorithm to learn most rules for the language level, while the structure level will require more data.

In Figure 1 it is evident that the algorithm discriminates the language better than the structure. Most of the language discrimination is learned after 20% of the data seen, when it reaches accuracy of almost 96%. The final accuracy score is 98.59%, which we regard as an excellent result considering the text is from the 18th century when inconsistency in Croatian orthography was frequent and more than half (53%) of tokens in the manually annotated corpus are Croatian.

The most structure discrimination is learned at about 40% of the data seen, when it reaches accuracy of more than 94%. The final accuracy score is 95.92%, which is a result that exceeds our expectations considering the structure level has 19 labels.

Both curves are still significantly rising. By adding additional data to the training set from the experiment with speed comparison, we could improve accuracy scores for both the language and the structure level, but also decrease the time needed for manual processing of the data.

Finally, we consider the existing algorithm to be beneficial in the language and structure annotation of 18th century dictionary entries, with the accuracy scores being sufficiently high and considerably speeding up the process of the manual processing.

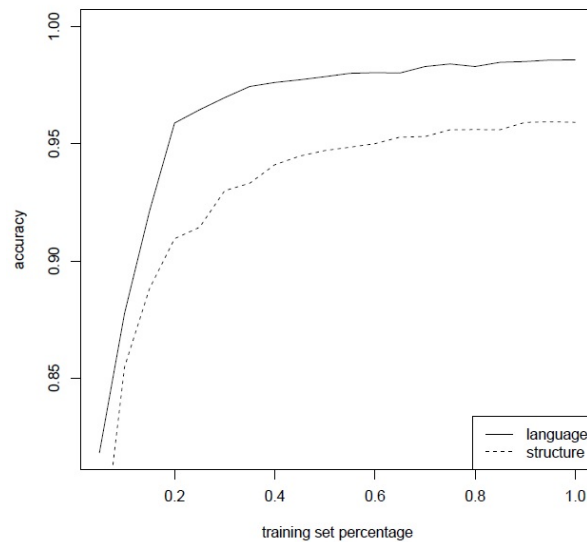


Fig. 1: The learning curve for the language and structure labels

6. Conclusion

In this paper we investigate the application of a supervised structural machine learning method on the language and structure annotation of 18th century dictionary entries. We use state-of-the art supervised machine learning algorithm for labeling sequences – conditional random fields (CRFs). Our research is conducted on the first volume of a trilingual dictionary ‘Dizionario italiano–latino–illirico’ (Italian–Latin–Croatian Dictionary) compiled by Ardellio della Bella and printed in Dubrovnik in 1785. The training sample comprises of 8,340 tokens out of 403,128 found in the whole of the dictionary corpus. We measure the performance of several features, finally combining all features that show increase over the results obtained with the basic feature for the best result.

We reach the accuracy of approximately 98% for the language annotation with three labels and around 96% for the structure annotation with 19 labels. We compute the usual metrics used for

model evaluation in the field of natural language processing (precision, recall, F measure and accuracy) for both levels of annotation.

In this paper we answered the question whether correcting automatically assigned language and structure labels reduces the time for the manual annotation, and if confirmed, by how much. This experiment confirmed that pre-annotating the data is roughly five times faster than the full manual annotation.

The learning curves for both the language and the structure level are still significantly rising. By adding additional data to the training set from the experiment with speed comparison, we could improve accuracy scores for both language and structure level, but also decrease the time needed for manual processing of data.

7. Acknowledgements

This work was partially supported by the Swiss National Science Foundation grant IZ74Z0_160501.

8. References

- Archer, D., Kytö, M., Baron, A. & Rayson, P. (2015). Guidelines for normalising early modern english corpora: Decisions and justifications. *ICAME Journal*, 39(1), pp. 5–24.
- Boonsuk, S., Suchato, A., Punyabukkana, P., Wutiwiwatchai, C. & Thatphithakkul, N. (2014). Language recognition using latent dynamic conditional random field model with phonological features. *Mathematical Problems in Engineering*, 2014.
- Buchler, M., Franzini, G., Franzini, E. & Moritz, M. (2014). Scaling historical text re-use. In *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 23–31.
- Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 355–362.
- della Bella, A. (1785). *Dizionario italiano-latino-illirico*. Nella Stamperia Privilegiata, prima edizione ragusea edition.
- Foley, J. and Allan, J. (2015). Retrieving time from scanned books. In *Advances in Information Retrieval*, Springer, pp. 221–232.
- He, X., Zemel, R. S. & Carreira-Perpindn, M. (2004). Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, IEEE, pp. 695–702.
- Hendrickx, I., Génereux, M. & Marquilha, R. (2011). Automatic pragmatic text segmentation of historical letters. In *Language Technology for Cultural Heritage*, Springer, pp. 135–153.

- Horton, R., Morrissey, R., Olsen, M., Roe, G., Voyer, R., et al. (2009). Mining eighteenth century ontologies: Machine learning and knowledge classification in the encyclopédie.
- Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J., et al. (2014). Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. In *IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly*.
- Kumar, S. and Hebert, M. (2003). Discriminative fields for modeling spatial dependencies in natural images. In *In NIPS*. MIT Press.
- Lafferty, J. D., McCallum, A. & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, pp. 282–289.
- Liu, Y., Carbonell, J., Weigele, P. & Gopalakrishnan, V. (2005). Segmentation conditional random fields (scrfs): A new approach for protein fold recognition. In *Research in Computational Molecular Biology*, Springer, pp. 408–422.
- McCallum, A. and Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4, CONLL '03*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 188–191.
- Mitankin, P., Gerdjikov, S. & Mihov, S. (2014). An approach to unsupervised historical text normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, DATeCH '14*, New York, NY, USA. ACM, pp. 29–34.
- Okazaki, N. (2007). Crfsuite: a fast implementation of conditional random fields (crfs).
- Pettersson, E. and Nivre, J. (2015). Improving verb phrase extraction from historical text by use of verb valency frames. In Megyesi, B., editor, *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*.
- Pinto, D., McCallum, A., Wei, X. & Croft, W. B. (2003). Table extraction using conditional random fields. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03*, New York, NY, USA. ACM, pp. 235–242.
- Piotrowski, M. (2012). *Natural language processing for historical texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Raaf, M. (2015). *Historical Corpora: Challenges and Perspectives*, chapter A web-based application for editing manuscripts, Gunter Narr Verlag, pp. 365–372.
- Sato, K. and Sakakibara, Y. (2005). Rna secondary structural alignment with conditional random fields. *Bioinformatics*, 21(suppl 2), pp. 237–242.
- Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03*, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 134–141.

- Shen, D., Sun, J.-T., Li, H., Yang, Q. & Chen, Z. (2007). Document summarization using conditional random fields. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., pp. 2862–2867.
- Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4), pp. 267–373.
- Taskar, B., Abbeel, P. & Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, UAI'02*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., pp. 485–492.
- TEI Consortium, T., editor (2014). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, chapter Dictionaries. TEI Consortium, 2.6.0 edition.
- Yu, D., Wang, S., Karam, Z. & Deng, L. (2010). Language recognition using deep-structured conditional random fields. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, IEEE, pp. 5030–5033.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



DEBWrite: Free Customizable Web-based Dictionary Writing System

Adam Rambousek, Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 60200 Brno, Czech Republic
{rambousek,hales}@fi.muni.cz

Abstract

Today, lexicographers can avail themselves of several commercial and freely distributed dictionary writing systems (DWS). Nevertheless, there is still a group of users whose requirements are not satisfied by existing DWSs. In various lexicographic forums, there is a growing demand for freely available DWS that allows customization of the dictionary microstructure. In accordance with such requests, a new project was developed as part of the DEB (Dictionary Editor and Browser) platform. DEBWrite is implemented as a multi-platform web application based on open standards. It allows users to create and share a new dictionary without any difficult configuration or advanced technical skills. According to a defined entry structure, the editing form and the public dictionary browser are generated automatically. DEBWrite supports small and larger team cooperation when working on the dictionary content. Access rights management for the created dictionary involves three levels of user roles: a manager, an editor, and a reader. It is possible to publish the resulting dictionary in various formats, both for human readers, and for external applications (e.g. NLP-related applications that need to work with lexicographic data). The dictionary may be published in an online form, or in formats suitable for print preparation.

Keywords: dictionary writing system; lexicographic platform; dictionary authoring; DEB platform

1. Introduction

There are several software tools available for dictionary creation and publication, both commercial (e.g. IDM DPS (IDM DPS, 2006) or TLex (Joffe and de Schryver, 2004)), and freely available (e.g. Mātāpuna (Moskovitz, 2004)). During the development of the DEB (Dictionary Editor and Browser) lexicographic platform (Horák and Rambousek, 2007; Horák et al., 2008), we have designed and implemented many lexicographic projects with complex entry structure or management. On the other hand, we have also experienced demand for dictionary writing software in the form of small size dictionaries with entry structure, usually by a small lexicographic team with limited resources for their project. For such teams, existing free tools are too limiting, and commercial tools are too expensive. Several such dictionaries were created using the DEB platform tools. For example, the Terminological Dictionary of Fine Arts by the Faculty of Fine Arts, Brno University of Technology (Horák and Rambousek, 2007), or the Czech-English Dictionary of Ethnological

Terminology by the The National Institute of Folk Culture¹. To fulfil the requirements for such range of dictionaries, a new application of the DEB platform was developed, called DEBWrite.

2. The DEB platform

Utilizing the experience from several preceding lexicographic projects, we have designed and implemented a universal dictionary writing system that can be exploited in various lexicographic applications to build distributed lexical databases. The system is called Dictionary Editor and Browser, or the DEB platform (Horák and Rambousek, 2007, 2010). Since 2005, the DEB platform was applied in more than 10 large international research projects. Large-scale applications based on the DEB platform include the lexicographic workstation for the development of the Czech Lexical Database (Horák and Rambousek, 2013) with detailed morpho-syntactic information on more than 213,000 Czech words, or the complex lexical database Cornetto combining the Dutch wordnet, an ontology, and an elaborate lexicon (Horák et al., 2008). Currently ongoing projects include Pattern Dictionary of English Verbs tightly interlinked with the corpus evidence (Maarouf et al., 2014), Family names in Britain and Ireland (Hanks et al., 2011) providing detailed investigations for over 45,000 surnames to be published by Oxford University Press, or the dictionary of the Czech Sign Language² with an extensive use of video recordings to present the signs (Rambousek and Horák, 2015).

The DEB platform is based on the client-server architecture, which brings along a lot of benefits. All the dictionary and interlinked data are stored on a server and a considerable part of the functionality is also implemented on the server-side, consequently the client application can be very lightweight. This approach provides very good tools for editor team cooperation; data modifications are immediately seen by all involved users. The DEB server also provides authentication and authorization tools.

The server part is built from small, reusable parts, called servlets, which allow a modular composition of all services. Each servlet provides different functionality such as database access, dictionary search, morphological analysis or a connection to corpora. The overall design of the DEB platform focuses on modularity. The data stored in a DEB server can use any kind of structural database (or consult several databases and join them into one compact dictionary storage) and prepare and combine complex results of answers to user queries without the need to use specific query languages for each data source. The main data storage is currently provided by the Sedna XML database (Fomichev et al., 2006), which is an open-source native XML database providing XPath and XQuery access to a set of document containers. Several DEB applications also work with connections to standard relational databases, such as PostgreSQL or MySQL, or to specialized data providers, such as the geographical information system GRASS or a morphological analyser.

¹ <http://www.nulk.cz>

² <http://www.dictio.info>

The user interface, which forms the most important part of a client application, usually consists of a set of flexible complex forms that dynamically cooperate with the server parts. Client applications can be implemented in any programming language that allows to interact with the DEB server using the available server interfaces.

Client applications communicate with servlets using standard HTTP requests in a manner similar to a popular concept in web development called AJAX (Asynchronous JavaScript and XML) or using the SOAP protocol³. The data are transported over HTTP in a variety of formats – RDF, XML documents, JSON-encoded data⁴, plain-text formats, or marshalled using SOAP.

The main assets of the DEB development platform can be characterized by the following points:

- All the data are stored on the server and a considerable part of the functionality is also implemented on the server, while the client application can be very lightweight.
- Very good tools for (remote) team cooperation; data modifications are immediately seen by all the users. The server also provides authentication and authorization tools.
- Server may offer different interfaces using the same data structure. These interfaces can be reused by many client applications.
- Homogeneity of the data structure and presentation. If an administrator commits a change in the data presentation, this change will automatically appear in every instance of the client software.
- Integration with external applications.

2.1 Linked Data

The term Linked Data refers to a methodology for publishing and interlinking structured data online. This methodology was proposed by Berners-Lee in 2006 (Berners-Lee, 2006; Bizer et al., 2009), who outlined four rules of how data are required to meet for easy sharing and interconnecting:

1. objects are identified by an URI⁵ (e.g. <http://dbpedia.org/page/Brno>),
2. URI identifiers are HTTP links, where people or software tools can access the data,
3. useful information are provided on given URI, using the appropriate standards (like RDF) (the previously mentioned page contains links to the same information in multiple formats, RDF is provided at <http://dbpedia.org/data/Brno.rdf>),
4. other objects are referenced using their URIs to get more information (e.g. link from the [Brno.rdf](http://dbpedia.org/data/Brno.rdf) to http://dbpedia.org/resource/South_Moravian_Region).

All resources stored in the DEB platform can be published using the Linked Data methodology. The DEB platform provides the tools for Linked Data presentation and the decision how to release the data lies with the author. Linked Data requirements are satisfied in the following manner:

³ <http://www.w3.org/TR/2007/REC-soap12-part0-20070427/>

⁴ <http://www.json.org/xml.html>

⁵ Uniform resource identifier (Berners-Lee et al., 2005)

1. use URIs as names – each entry has a unique URI identifier,
2. use HTTP URIs – through the DEB platform API, entries are accessible on HTTP URI,
3. provide useful information using standards – when linking to an entry URI, the data are displayed either in raw XML format, or converted to RDF or other defined format,
4. link to other URIs – the DEB platform enables to link to other resources if provided by the data author.

These requirements are fully embraced in DEB-based projects, DEBVisDic (Horák et al., 2006) and the KYOTO project (Horák and Rambousek, 2010, 2009), where all the information were released as Linked Data.

Berners-Lee later published a rating system for the distributed data, while expanding the term Linked Data to Linked Open Data – which means Linked Data that are released under an open licence. This rating system is aimed especially at government agencies to encourage them to publish valuable (and reusable) information. The importance of Linked Open Data is acknowledged for example by the European Union, funding projects like *LOD2* (large integrating project to develop tools, standards and management methods for Linked Open Data) or *Open Data Portal* (catalogue of data available for reuse). The rating system follows these principles:

- 1 star – the data are available on the web in any format, with an open licence.
- 2 stars – the data are published in machine-readable structured format.
- 3 stars – the data use non-proprietary format.
- 4 stars – W3C open standards (RDF and SPARQL) are used to identify objects for linking.
- 5 stars – the data contain links to other resources to give context.

The DEB platform offers a full support to the dictionary publisher to disseminate the dictionary content as Linked Open Data:

1. published online with an open licence – this has to be decided by the data authors, but the DEB platform enables releasing data on the web.
2. available as machine-readable structured data – documents in the DEB platform are stored in an XML format which is machine-readable.
3. non-proprietary format – XML is a standardized format.
4. use open standards from W3C (RDF and SPARQL) – XML format itself is the W3C standard, but to conform with this requirement more precisely, documents are converted to RDF format.
5. link to other resources – the DEB platform enables interlinking to other resources.

As demonstrated, the only limitation is the decision of the data authors regarding the licensing. When this is resolved, the DEB platform enables to publish all documents as Linked Open Data.

element	<input type="text" value="hw"/>	label	<input type="text" value="headword"/>	multiple	<input type="checkbox"/>	type	<input type="text" value="text"/>	<input type="button" value="remove"/>
element	<input type="text" value="gram"/>	label	<input type="text" value="grammar"/>	multiple	<input type="checkbox"/>	type	<input type="text" value="text"/>	<input type="button" value="remove"/>
element	<input type="text" value="recording"/>	label	<input type="text" value="pronunciation audio"/>	multiple	<input checked="" type="checkbox"/>	type	<input type="text" value="file"/>	<input type="button" value="remove"/>
element	<input type="text" value="mean"/>	label	<input type="text" value="meaning"/>	multiple	<input checked="" type="checkbox"/>	type	<input type="text" value="container"/>	<input type="button" value="remove"/>
element	<input type="text" value="nr"/>	label	<input type="text" value="meaning nr"/>	multiple	<input type="checkbox"/>	type	<input type="text" value="number"/>	<input type="button" value="remove"/>
element	<input type="text" value="def"/>	label	<input type="text" value="explanation"/>	multiple	<input type="checkbox"/>	type	<input type="text" value="textarea"/>	<input type="button" value="remove"/>
element	<input type="text" value="ex"/>	label	<input type="text" value="usage"/>	multiple	<input checked="" type="checkbox"/>	type	<input type="text" value="textarea"/>	<input type="button" value="remove"/>

Figure 1: Setting the entry structure.

3. The DEBWrite application

The DEBWrite application is implemented as a multi-platform web application, utilizing HTML5 and JavaScript standards⁶ that allow full interoperability and dynamic adaptations to current dictionary interfaces. The DEBWrite application allows users to create and share a new dictionary without any complicated configuration or advanced technical skills. Based on experience with dictionaries in the DEB platform, a default entry structure is proposed that fits many dictionaries (also with terminological dictionaries in mind). Each entry is composed of a top level information (headword and its variants, grammatical information, domain/category) and any number of meanings (each containing explanation and usage examples). Translations to various languages, cross-references to other entries (with relation type), collocations, and external references may be included on the entry level or meaning level. Within the dictionary definition form, users may alter the entry structure in a graphical interface (see Figure 1) – deleting unnecessary information or adding new entry fields, changing labels, or altering the option lists (relation types, languages for translations, domains...).

According to the updated entry structure, the editing form and the public browser are generated automatically. See Figure 2 for an example of the editing form. The dictionary website design is fully customizable via CSS stylesheets or templates that are used for output generation. XSLT templates are used as a default option, however HandlebarsJS template engine⁷ is also evaluated. Based on the user feedback, the preferred template engine might be changed in the future DEBWrite updates. The authors may either edit the source code of the output generating files, or select some of the variables (e.g. colours and font styles) in the graphical interface (see Figure 3). In future versions, more detailed graphical interface to change the output layout will be added. Each dictionary may use multiple output templates to provide different dictionary previews based on user settings.

⁶ with jQuery, <https://jquery.com/>, and jQuery UI, <https://jqueryui.com/>, libraries.

⁷ <http://handlebarsjs.com/>

headword
grammar
+ **pronunciation audio** No file chosen

+ **meaning**
meaning nr
explanation
+ **usage**

- + **meaning**
meaning nr
explanation
+ **usage**

Figure 2: Example of the editing form automatically generated from the settings.

The DEBWrite dictionary editor also supports upload of multimedia attachments (e.g. large figures, audio or video recordings) to supplement the entries. The authors need to specify a special field type in the entry structure for file uploads. The server detects the attachment type (e.g. image, video, audio) and displays the multimedia content in an appropriate form for the output. See Figure 4 for an example of multimedia file upload and output.

In cases, when the lexicographers have some information prepared in advance, DEBWrite can simplify the start of the dictionary creation process. A common scenario includes the situation, where DEBWrite imports a list of headwords and automatically creates corresponding empty entries prepared for expert editing. Another scenario works with the requirement of moving rich existing structured data to DEBWrite. In such cases, DEBWrite can import a (part of the) full dictionary in the XML format. As of now, the imported file must follow the XML structure used in the DEBWrite application internally. However, a conversion between different (compatible) XML structures is a matter of applying an XSLT template conversion. Future versions of DEBWrite will support also import of data in custom XML format.

The application also supports an export to standard XML file. Preprocessed XSLT templates are included to export converted dictionary data into an HTML format for online publishing. For printed or electronic edition in PDF, the data are converted to L^AT_EX and subsequently to PDF format.

To enhance the possibility to share and re-use lexicographic resource sharing, DEBWrite also provides the data in the form compliant with the Linked Data methodology (see section 2.1). The decision about the data licensing and access control lies entirely on the dictionary authors, however DEBWrite provides the tools needed to make the sharing easy.

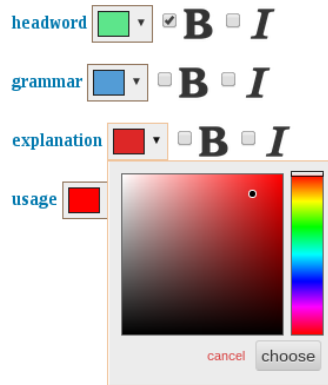


Figure 3: Example of output design customizations.

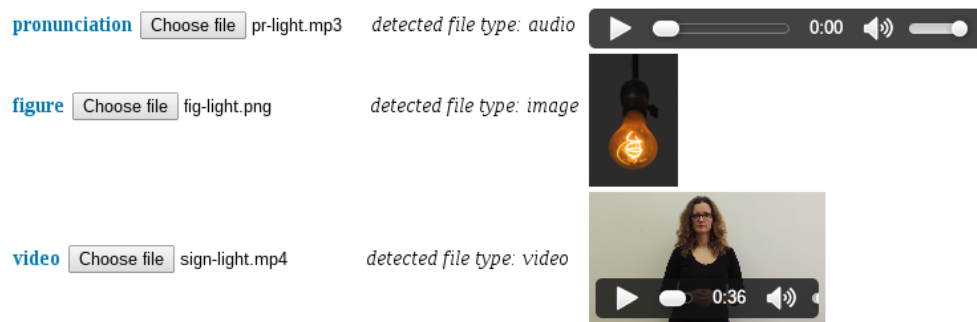


Figure 4: Output representation of various media attachment types.

One of the major advantages of the DEBWrite application lies in its support of a team cooperation on the dictionary preparation process. DEBWrite classifies authorized users into one of three possible user roles: a manager, an editor, or a reader (see Figure 5 for example of user access management).

- The user who created the dictionary is the dictionary *manager*. Managers may alter any dictionary settings. They may grant access to the dictionary to other users, specifying their role. Managers are able to edit all the dictionary entries and set an entry for publication. The manager may also decide to make published entries publicly available, which means that no password is needed to browse the dictionary (this might be regarded as a fourth user role in the dictionary access management).
- An *editor* may edit entries before they are set to be published.
- *Readers* may browse and navigate through the published entries and their attachments with advanced search capabilities.

username	xrambous	role	manager ▼	e-mail	xrambous@fi.muni.cz	update	reset password	
username	scruffy	role	editor ▼	e-mail		update	remove access	reset password
username	wloki	role	editor ▼	e-mail		update	remove access	reset password
username	dheiko	role	read ▼	e-mail		update	remove access	reset password
<input type="checkbox"/> make dictionary public								
<input type="button" value="add user"/>								

Figure 5: User access management.

4. Conclusions

We have introduced a new customizable and freely available dictionary writing system named DEBWrite. The application prototype is currently in public testing, available at <http://deb.fi.muni.cz/debwrite>. As a part of testing, the Terminological Dictionary of Fine Arts was converted to DEBWrite from the original application (where the editing form functionality was originally limited to the Firefox browser only), allowing multi-platform editing and providing better user experience.

5. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

6. References

- Berners-Lee, T. (2006). Design Issues: Linked Data.
- Berners-Lee, T., Fielding, R. & Masinter, L. (2005). Uniform Resource Identifier (URI): Generic Syntax. STD 66 (INTERNET STANDARD).
- Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data-The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3), pp. 1–22.
- Fomichev, A., Grinev, M. & Kuznetsov, S. (2006). Sedna: A Native XML DBMS. *Lecture Notes in Computer Science*, 3831:272.
- Hanks, P., Coates, R. & McClure, P. (2011). Methods for Studying the Origins and History of Family Names in Britain. In *Facts and Findings on Personal Names: Some European Examples*, Uppsala. Acta Academiae Regiae Scientiarum Upsaliensis, pp. 37–58.
- Horák, A., Pala, K., Rambousek, A. & Povolný, M. (2006). DEBVisDic – First Version of New Client-Server Wordnet Browsing and Editing Tool. In *Proceedings of the Third International WordNet Conference - GWC 2006*, Jeju, South Korea. Masaryk University, Brno, pp. 325–328.
- Horák, A. & Rambousek, A. (2007). DEB Platform Deployment – Current Applications. In *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*, Brno, Czech Republic. Masaryk University, pp. 3–11.

- Horák, A. & Rambousek, A. (2009). Using Wordnets and Ontologies for Text-Meaning Assignment - Implementation Details of the KYOTO Project First Phase. In *Proceedings of the 4th International Conference on Software and Data Technologies, Volume 2*, Portugal. INSTICC, pp. 303–307.
- Horák, A. & Rambousek, A. (2010). Using DEB Services for Knowledge Representation within the KYOTO Project. In *Principles, Construction and Application of Multilingual WordNets, Proceedings of the Fifth Global WordNet Conference*, New Delhi, India. Narosa Publishing House, pp. 165–170.
- Horák, A. & Rambousek, A. (2013). PRALED – A New Kind of Lexicographic Workstation. In Przepiórkowski, A., Piasecki, M., Jassem, K. & Fuglewicz, P., editors, *Computational Linguistics: Applications*, Springer, pp. 131–141.
- Horák, A., Vossen, P. & Rambousek, A. (2008). A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In *Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing*, Haifa, Israel. Springer-Verlag, pp. 1–15.
- IDM DPS (2006). IDM Dictionary Production System. http://www.idm.fr/products/dictionary_writing_system.
- Joffe, D. & de Schryver, G.-M. (2004). TshwaneLex – Professional off-the-shelf lexicography software. In *Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts*, Brno, Czech Republic. Masaryk University, Faculty of Informatics. <http://tshwanedje.com/tshwanelex/>.
- Maarouf, I. E., Bradbury, J., Baisa, V. & Hanks, P. (2014). Disambiguating verbs by collocation: Corpus lexicography meets natural language processing. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. & Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Moskovitz, D. (2004). Mātāpuna Dictionary Database System. In *Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts*, Brno, Czech Republic. Masaryk University, Faculty of Informatics. <http://matapuna.thinktank.co.nz/matapuna/>.
- Rambousek, A. & Horák, A. (2015). Management and Publishing of Multimedia Dictionary of the Czech Sign Language. In Biemann, C., Handschuh, S., Freitas, A., Mezziane, F. & Métais, E., editors, *Natural Language Processing and Information Systems, NLDB 2015*, Lecture Notes in Computer Science, Springer, pp. 399–403.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Automatically Linking Dictionaries of Gallo-Romance Languages Using Etymological Information

Pascale Renders¹, Gérard Dethier², Esther Baiwir³

¹FNRS/University of Liège

²University of Liège

³FNRS/University of Liège

pascale.renders@ulg.ac.be, g.dethier@alumni.ulg.ac.be, ebaiwir@ulg.ac.be

Abstract

How could we link together digital dictionaries which have no common lexical units, but deal with the same linguistic area? And how could we do that automatically, in order to ensure that all future updates of these dictionaries are taken into account in the linking process?

This contribution exposes the solutions that we propose in the field of French and Gallo-Romance historical lexicography. The digitalisation currently in progress of a work of scientific reference, i.e. the *Französisches Etymologisches Wörterbuch* (FEW), gives us a mean to link together other dictionaries, such as the *Dictionnaire Etymologique de l'Ancien Français* (DEAF), the *Dictionnaire du Moyen Français* (DMF), the *Anglo-Norman Dictionary* (AND), or the *Atlas Linguistique de la Wallonie* (ALW), through the use of the references of these dictionaries to the FEW. Concrete examples of linking lexical data are discussed in this context.

We also describe a simple peer-to-peer protocol allowing e-dictionaries to be automatically linked in a distributed way using the references of their articles. An implementation based on a simple REST API is suggested to let teams maintaining different e-dictionaries keep their own technologies and data schema.

Keywords: Linked lexical data; Gallo-Romance lexicography; FEW; Exploitation of language resources

1. Introduction

Etymology is an information that is not systematically available in all dictionaries. However, it might be used to link together digital dictionaries which have no common lexical units, but deal with the same linguistic area. In the field of French and Gallo-Romance lexicography, the digitalisation currently in progress of a reference dictionary, the *Französisches Etymologisches Wörterbuch* (FEW), gives us the opportunity to automatically link dictionaries such as the *Dictionnaire Etymologique de l'Ancien Français* (DEAF), the *Dictionnaire du Moyen Français* (DMF), the *Anglo-Norman Dictionary* (AND) or the *Atlas Linguistique de la Wallonie* (ALW).

The questions that will be addressed are (1) how can we link these resources and what is to be linked exactly; (2) how can this be done automatically? This contribution gives some examples of lexical units that could be linked in French and Gallo-Romance lexicography, exposes the linking process we imagine in theory and explains the way in which this could be implemented in practice.

2. A Case Study: Gallo-Romance Lexicography

The FEW has the particularity to gather lexical units of French, Gascon, Occitan, Francoprovençal and their dialects, according to their common ancestry (etymon). Each FEW article provides, under an etymon lemma, the history of one lexical family. Lexical units whose etymology is not known are gathered in the volumes 21–23, with an onomasiologic classification.

As a thesaurus and a reference for the etymology of all lexical units in the area under consideration, the FEW works as a “*lieu de synthèse*” in this linguistic area, see (Buchi and Renders, 2013). Consequently, the FEW is systematically cited in many historical dictionaries of these languages and dialects. This provides a wonderful opportunity to link dictionaries together by putting the FEW at the center of a lexicographic network, through the use of etymological information.

The linking process has another purpose. The dictionaries mentioned above not only mention, but regularly update, the FEW, for instance by providing a new etymology to FEW units from volumes 21–23. Unfortunately, providing an updated version of the FEW integrating these contributions is not possible in practice, because of the complex structures of the FEW. Linking the FEW with all the lexicographic resources available would provide users and lexicographers with a facilitated and easy access to these updates. In this context, it is necessary to implement an automatic linking process, in order to ensure that all future updates of these dictionaries are actually taken into account.

Gallo-Romance dictionaries that could be involved are, for example, the DEAF, the AND, the ALW, the TLF, and all the resources provided by the ATILF (TLF-Etym etc.). Some of the historical or etymological dictionaries of Romance languages, such as the DERom, could also be added to this network. These dictionaries mention for each lexical unit a “FEW reference” i.e. the exact location in the dictionary (volume, page and column) where this lexical unit can be found. For example, ALW 17 provides a new etymology for 21 lexical units that are described as from “uncertain origin” in FEW. For each of them, the ALW mentions the exact location where it appears in the FEW and provides the new location where it should be moved according to its new etymology. The wallonish verbs “*zam’ter*”, “*cham’ter*” were, for instance, marked “from uncertain origin” in the FEW and therefore put in the volume 21 (FEW 21, 342a). However, ALW 17, 206a defines “*examen*” (FEW 3, 258a) as their common etymon. Updating the FEW means that these lexical units should be moved from FEW 21, 342a to FEW 3, 258a under the “*examen*” lemma. The same applies for the wallonish term *fournakeye* (f.) “*ribambelle*” (ALW 17, 73a and 75a), which should be added to FEW 3, 907b.

3. Linking E-Dictionaries

This section describes an automated method of linking e-dictionaries. The method is first described from a theoretical point of view. Then a suggestion of implementation is proposed.

3.1 Definitions

From a Computer Science point of view, a dictionary is a set of entries (k, v) where k is a key and v a value. An additional property that is commonly accepted is the unicity of the keys in a given dictionary i.e. in the set of all entries, it is not possible to find two entries (k_1, v_1) and (k_2, v_2) with $k_1 = k_2$.

Let v_1 be the article of a dictionary d_1 having the key k_1 and v_2 be the article of another dictionary d_2 having the key k_2 . If the article v_2 references the article v_1 , the reference can be represented by the tuple (d_2, k_2, d_1, k_1) . A reference can also be noted $v_2 \rightarrow v_1$ or $(d_2, k_2) \rightarrow (d_1, k_1)$.

Although the above definition is straightforward, the keys and articles for a particular dictionary are not always easily defined. In the case of the FEW, the FEW reference can be used as the key (e.g. FEW 3, 258a). As previously stated, the FEW reference is a location (the column of a particular page in a given volume). In some cases (when several articles have the same location), the location has still to be augmented with the etymon to uniquely identify one article. This is also true for ALW references which also represent locations where a particular notice can be found.

Let D be the set of all the dictionaries complying to the rules described above (set of entries with unique keys), K_i the set of all the keys of a dictionary d_i and R the set of all references $(d_i, k_{i,j}, d_k, k_{k,m})$ where $d_i, d_k \in D$, $k_{i,j} \in K_i$ and $k_{k,m} \in K_k$. In a perfect world, when reading the article v of dictionary d_i with key $k_{i,j}$, we would have access to all references $(d_j, k_{j,l}, d_i, k_{i,j})$ with $d_j \in D$ and $k_{j,l} \in K_j$ and therefore all the information available on the article: its content but also links to other articles (and their content) referencing it. If one of these articles suggests an update, the reader would be aware of it and always have access to the latest “version” of an article.

The above model can be applied to the task of linking dictionaries of Gallo-Romance languages exposed in the previous section. Indeed, if the FEW, the ALW, etc. can be considered as part of D , then we can model references between articles of these dictionaries using above framework. For instance, let d_{FEW} be the FEW and d_{ALW} be the ALW. The example of update of the FEW by the ALW from previous section actually implies two distinct references:

1. ALW 17, 206a \rightarrow FEW 21, 342a (removing the lexical unit)
2. ALW 17, 206a \rightarrow FEW 3, 258a (adding the lexical unit)

with $ALW\ 17,\ 206a \in K_{ALW}$, $FEW\ 21,\ 342a \in K_{FEW}$ and $FEW\ 3,\ 258a \in K_{FEW}$.

We can define an e-dictionary as a system able to provide the content of an article v given its key k . We suppose that an e-dictionary represents a single dictionary of D . In the following, we will note e_i the e-dictionary system hosting a dictionary $d_i \in D$. In order to link several e-dictionaries together, we only need a way to implement R . In this case, someone reading an article through an e-dictionary would have access to the content of the article and to all the articles referencing it or referenced by it only by querying the e-dictionary and the system hosting R .

Implementing that kind of system is not trivial. The most obvious solution is a centralised platform maintained by an independent organisation. However, building this kind of organisation and platform is neither simple nor efficient: it requires substantial funding in order to maintain R , a huge set that continuously evolves. Also, it is not scalable nor secure from a technical point of view as it represents both a potential bottleneck and a single point of failure.

An alternative is to let the e-dictionaries build a distributed representation of R in a collaborative way. Indeed, each e-dictionary does not need to be aware of the whole R set. Let $R_{i,j}$ be the subset of R containing all references implying keys from either $d_i \in D$ or $d_j \in D$. An e-dictionary representing d_i only needs to be aware of $S_i = \bigcup_{j \in E} R_{i,j}$ where E is the set of dictionaries to which d_i refers (i.e. the dictionaries to which d_i 's articles refer).

Next section describes the protocol that enables e-dictionaries to build $R_{i,j}$ in a collaborative way. Some technological choices are also suggested to build a practical solution.

3.2 The Linking Protocol

In this section, we will describe a simple protocol allowing e-dictionary systems to build their S_i set in a distributed way. Concrete technologies are suggested to actually implement the protocol.

3.2.1. Theory

Let d_i be a dictionary represented by an e-dictionary system e_i . In order to build S_i , e_i will send and receive messages representing the creation of references. When a reference from article v with key k_v of d_i is made to an article w with key k_w of d_j , e_i sends a message notifying e_j of the new reference (d_i, k_v, d_j, k_w) being created, in addition to storing the new reference in its own representation of $R_{i,j}$ (and therefore S_i). When e_j receives the message sent by e_i , it updates its representation of $R_{i,j}$ (and therefore S_j). When $R_{i,j}$'s representation is updated on both e_i and e_j , both e-dictionaries are aware of the reference being made from article v to w and are therefore able to expose this reference to their users.

With this protocol, creating a reference in a e-dictionary enriches automatically the set of references in all other relevant e-dictionaries. This incremental approach also allows the continuous improvement of the existing set of references with a minimum effort as the maintenance of the global references set is automated.

It is to be noted that letting e-dictionaries build their set of references actually leads to the emergence of a network of e-dictionaries connected by their references.

The protocol described here implies a peer-to-peer architecture where e-dictionaries are the peers. This is good news as peer-to-peer architectures are well known for their good scalability and ro-

bustness. We did not address the security and robustness problems that may arise. Although these must be tackled in a real world implementation, they are beyond the scope of this paper.

3.2.2. Implementation

As already stated, most e-dictionaries are developed by different teams from different organisations. The technologies used by these teams to actually implement the e-dictionaries might strongly differ (PHP, Java, Node.js, etc.). Our suggestion is for all these e-dictionaries have their own internal representation and technology stack, but for them to expose a common yet minimal API allowing the exchange of messages as exposed at the beginning of this section. In this way, the coupling between different projects and teams is minimised and allows more flexibility, robustness and scalability from the technical point of view, as well as from the point of view of project management.

A modern approach is to implement the API using web-oriented technologies, and our suggestion would be to implement a simple REST API based on HTTP request and using JSON-encoded data¹. The advantage of this approach is that this kind of interface can be implemented using a wide range of technologies, thus imposing almost no constraints to the teams developing the different e-dictionaries.

Each e-dictionary must be hosted under a different hostname which can therefore be used to uniquely identify the e-dictionary system itself. Let `my-edict.org` be the hostname of e-dictionary `my-edict`. Below REST resources should be exposed in order to let the e-dictionary receive messages coming from external systems and let other e-dictionaries access the content of hosted articles.

In the following, we will use `cURL`² syntax to express HTTP requests in a formal and precise way. Each section starts with a summary of the HTTP request composed of the HTTP method (GET, POST, etc.) and the URL pattern (parameters are prefixed with a colon) e.g. `GET http://www.google.com/:service/` where `service` is a parameter.

Creating References

`POST http://my-edict.org/api/reference`

Posting (i.e. doing an HTTP POST request with) the following data to this resource should lead to the addition of a reference in `my-edict`:

```
{
  "source_dict": "http://other-edict.org",
  "source_id": "a-key-in-other-edict",
```

¹ JavaScript Object Notation, see <http://json.org/> for a full specification of this data-interchange format.

² Curl is a command line tool and library for transferring data with URL syntax, see <http://curl.haxx.se> for more details.

```
"dest_dict": "http://my-edict.org",
"dest_id": "a-key-in-my-edict",
}
```

The following cURL command (or some equivalent implementation) should be executed by e-dictionary `other-edict` when represented reference is actually created:

```
curl "http://my-edict.org/api/reference" -X POST -H "Content-Type:
application/json" -d @data.json
```

where `data.json` is a file containing above data.

On reception of this kind of message, `my-edict` should ensure that:

1. `dest_dict` does contain the identifier of `my-edict`,
2. `dest_id` is the identifier of an existing article hosted by `my-edict`.

If above conditions are true, the reference can be inserted in `my-edict`'s database. In this way, when a user wants to read the article of `my-edict` identified by `a-key-in-my-edict`, `my-edict` will be able to expose the incoming reference from article of `other-edict` identified by `a-key-in-other-edict`.

Accessing Articles

```
GET http://my-edict.org/api/articles/:article-id
```

Getting (i.e. doing an HTTP GET request on) this resource should return the following data describing the article identified by `:article-id` (which is a placeholder for a real ID) in `my-edict`, for instance:

```
{
  "article-id": "a-key-in-my-edict",
  "url": "http://my-edict.org/a-key-in-my-edict"
}
```

where `article-id` is the unique identifier of the article in `my-edict` and `url` is the URL at which the article can be accessed. It is to be noted that the URL scheme used to let users access articles is totally up to the team implementing the e-dictionary.

The following cURL command (or some equivalent implementation) should be executed when accessing an article:

```
curl "http://my-edict.org/api/articles/a-key-in-my-edict"
```

Listing References of an Article

```
GET http://my-edict.org/api/articles/:article-id/references
```

Getting (i.e. doing an HTTP GET request on) this resource should return the list of references associated to the article identified by `article-id` (which is a placeholder for a real ID) in `my-edict`, for instance:

```
[
  {
    "source_dict": "http://other-edict.org",
    "source_article_id": "a-key-in-other-edict",
    "dest_dict": "http://my-edict.org",
    "dest_id": "a-key-in-my-edict",
  },
  {
    "source_dict": "http://my-edict.org",
    "source_article_id": "a-key-in-my-edict",
    "dest_dict": "http://other-edict2.org",
    "dest_id": "a-key-in-other-edict2",
  }
]
```

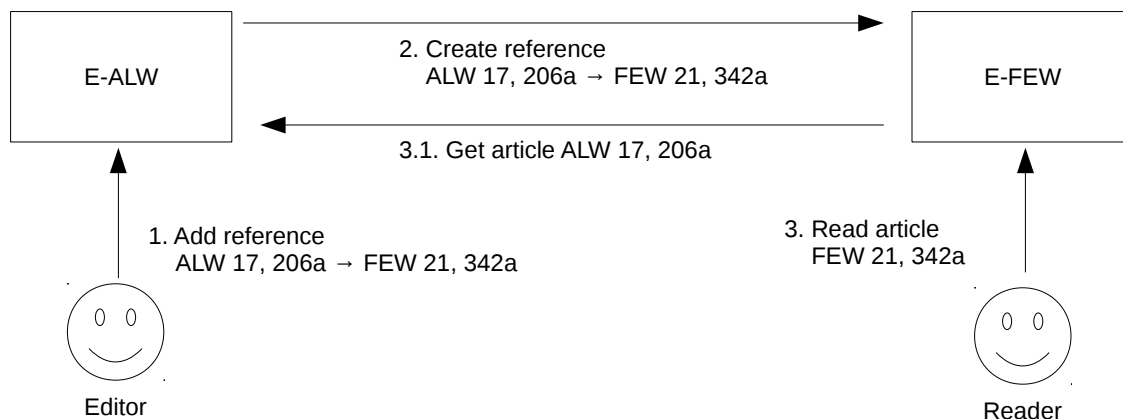
The references of an article include both incoming and outgoing references.

The following cURL command (or some equivalent implementation) should be executed when accessing the list of references of an article:

```
curl "http://my-edict.org/api/articles/a-key-in-my-edict/references"
```

3.2.3. Example

The following figure illustrates the interactions between users and e-dictionaries and the requests these interactions imply. The scenario described here uses the example given in section 3.1: an editor creates the reference ALW 17, 206a → FEW 21, 342a and, after that, a reader of the FEW displays the article FEW 21, 342a and has access at the same time to the update made by the article ALW 17, 206a.



1. An editor of the ALW adds the reference ALW 17, 206a → FEW 21, 342a by a means that is dependent on the way the e-ALW is implemented e.g. using a web interface.
2. The e-ALW notifies the e-FEW that a new reference has been created using the request described in section “Creating References”.
3. A reader of the FEW accesses the article FEW 21, 342a and, in a transparent way, the e-FEW builds a consolidated view of the article by retrieving also the article ALW 17, 206a (step 3.1) using the request described in section “Accessing Articles”.

The request described in section “Listing References of an Article” is not used in above scenario. However, it might make sense in more elaborated scenarios where a user wants to explore a graph of references that might span several e-dictionaries.

4. Conclusion

This paper discussed the question of linking together digital dictionaries which deal with the same linguistic area, some of these dictionaries giving additional or updated information about lexical units from other dictionaries. The update of the FEW through the references made by the ALW is given as a case study and highlights the need for linking.

We exposed a simple peer-to-peer protocol allowing several e-dictionaries to connect and maintain together the set of references involving the articles they host without the need for a central organization or system, preventing a potential bottleneck and a single point of failure. We also suggested an implementation of this protocol implying a small REST API that should be exposed by all e-dictionaries willing to be connected. This approach allows the teams responsible for the maintenance of the various e-dictionaries to keep their own technologies and representation for their data.

The described protocol allows us to link lexical units on the basis of any criteria. In the particular case of Gallo-Romance lexicography, the etymological information and the systematic mention of the FEW allow a quick linking process. At the same time, this linking process enables the update of the FEW by giving direct access to updates made by other dictionaries.

5. References

Buchi, E. & Renders, P. (2013). 41. Gallo-Romance I: Historical and etymological lexicography. In Gouws, R. H., Heid, U., Schweickard, W. & Wiegand, H. E., editors, *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, Handbooks of Linguistics and Communication Science (HSK) 5/4, De Gruyter Mouton, Berlin, pp. 653–662.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Improving the use of electronic collocation resources by visual analytics techniques

Roberto Carlini¹, Joan Codina-Filba¹, Leo Wanner^{2,1}

¹Natural Language Processing Group, Dept. of Information and Communication Technologies
Pompeu Fabra University, C/Roc Boronat, 138, 08018 Barcelona

²Catalan Institute for Research and Advanced Studies (ICREA), Passeig Lluís Companys, 23, 08010 Barcelona
Email: {roberto.carlini,joan.codina,leo.wanner}@upf.edu

Abstract

With the increasing prominence of the electronic medium in lexicography, the face of collocation resources also changed. Collocation dictionaries have been extended by additional material (e.g., examples from a corpus and interfaces for targeted access to information), and tools such as Sketch Engine have been developed, which query a corpus and display the collocational (and grammatical) behaviour of a specified word. However, the paradigm of consulting, viewing and exploring the resources still follows to a major extent the traditional dictionary look up philosophy: the user enters a keyword and obtains an outcome in a sequential text format. This implies significant limitations if the user wants to contrast information concerning different keywords or their collocates, view information in incremental detail, etc. Studies on the presentation of information argue that visualization techniques facilitate comprehension. It is thus not by chance that visualization of linguistic information and data has become a popular research topic. In our work, we aim to go one step further: we research how Visual Analytics (VA), which deals with the development of techniques that support the exploration, analysis and interpretation of information, can be used to explore collocation resources in the context of learning Spanish as second language.

Keywords: collocations; active learning; visual analytics

1. Introduction

With an increasing prominence of the electronic medium in lexicography, the face of collocation resources has also changed. Collocation dictionaries have been extended by additional material such as examples from a corpus and interfaces, which allow for targeted access of information; cf., e.g., DICE <http://www.dicesp.com>. Also, tools such as Sketch Engine (Kilgarriff et al., 2014) have been developed, which query a corpus and display the collocational (and grammatical) behaviour of a specified word. However, the paradigm of consulting, viewing and exploring the resources still predominantly follows the traditional dictionary look up philosophy: the user enters a keyword and obtains an outcome in a sequential text format. This implies significant limitations if the user wants to see which bases share a given collocate, contrast information concerning different keywords or their collocates, view in incremental detail some information, etc. Studies on the presentation of information (Tufte, 1983; Smith,

2005) argue that visualization techniques facilitate comprehension. It is thus not by chance that visualization of linguistic information and data has become a popular research topic; cf., e.g., (Collins et al., 2008, 2009; Penn and Carpendale, 2009; Feng and Lapata, 2010).

In our work, we aim to go one step further: we research how VA (Keim et al., 2008; Wong and Thomas, 2004) can be used to explore collocation resources in the context of second language learning. VA deals with the development of techniques that support the exploration, analysis and interpretation of information (in our case, collocation resources) via interactive visual interfaces.

In the context of second language learning, it is important to offer to the user the opportunity to (i) contemplate the possible collocates of a given keyword and compare the information concerning the frequency and context of their use; (ii) study the appearance of a collocation in different contexts; (iii) explore which of the keywords share the same collocate(s) and which ones do not; (iv) retrieve the syntactic structure of a collocation; etc. We explore VA techniques that account for these needs. Our resource is a large Spanish newspaper syntactic dependency corpus treebank. The corpus is indexed and processed for efficient computation of “collacability” between binary word co-occurrences, hence there holds a direct syntactic dependency and efficient access to supportive and illustrative information (such as samples of the use of a collocation in context).

In what follows, we first discuss the needs of a learner user of a collocation dictionary (Section 2). In Section 3, we then introduce the notion of VA and briefly show how it can be used for dynamic interactive exploration of collocation information. In particular, we present the VA techniques that we use in the context of the visualization of collocation information. Section 4 describes the application of these techniques to Spanish resources and illustrates their use through several examples, before Section 5 draws some conclusions from the described work and presents our future work in this area.

2. Needs of a Learner User of a Collocation Dictionary

Before we discuss the needs of the user, we shall briefly introduce the information that we assume to be available in a complete online collocation dictionary and the way it is presented.

2.1 The Content of a Collocation Dictionary

We take the Spanish collocation dictionary DiCE (Alonso et al., 2010) as an example of a complete online collocation dictionary. An entry of the DiCE contains the following main information:¹

¹ For a complete list of the information provided in a DiCE entry, see <http://www.dicesp.com/paginas/index/1>.

1. The corresponding list of disambiguated lexemes of the lemma of the keyword (or base), together with their part of speech (PoS) and semantic category; in the case of nouns, instead of the noun tag, the grammatical gender tag is given. Consider, for illustration, the information provided for *afecto* ‘affection’:

afecto1 m. sentimiento ‘sentiment’, *afecto2a* m. sentimiento ‘sentiment’,
afecto2b m. manifestación ‘manifestation’, *afecto3a* adj. estado ‘state’,
afecto3b adj. estado ‘state’, *afecto3c* adj. estado ‘state’.

2. For each lexeme, as, e.g., for *afecto1*:

- (i) its argument structure
 afecto de individuo X por hecho Y ‘affection of individual X for a fact Y’;
- (ii) its (quasi-)synonyms and antonyms
 emoción, estado de ánimo, pasión1, sentimiento1a;
- (iii) its subcategorization (government) structure
 1 →X *de* N | A_{pos}
 2 →Y *por* N | *ante* N | *hacia* N.

which states that the first semantic actant of *afecto1* is projected onto its first syntactic actant, which is realized either as a noun with a preposition *de* ‘of’ or as a possessive adjective, and that its second semantic actant is projected onto its second syntactic actant, which, in turn, is realized as a noun with one of the prepositions *por* ‘for’, *ante* ‘before’, or *hacia* ‘towards’.

- (iv) its collocates, categorized first according to the PoS of the collocate and its default location relative to the base (i.e.,: <verb>+BASE, BASE+<verb>, <adjective>+BASE, etc.), and then, within each of these categories, according to the semantics of the collocate in combination with the base
 manifestar ~ ‘manifest ~’
 expresar ‘express’

The use of the individual lexemes and of the individual collocations is illustrated by examples, mainly from a corpus of Spanish of the Spanish Royal Academy (<http://corpus.rae.es/creanet.html>); consider Figure 1 for illustration.

2.2 The Needs of the User

Online collocation dictionaries of the type of DiCE facilitate information when the intention of the user is to look up the collocates of a base (in order to then choose one of them), to verify a collocation they had in mind, or to learn about the use of a specific collocation in context. They may also provide some detailed information on the base lexeme—e.g., its argument and subcategorization structures or its (quasi-)synonyms or antonyms. To obtain

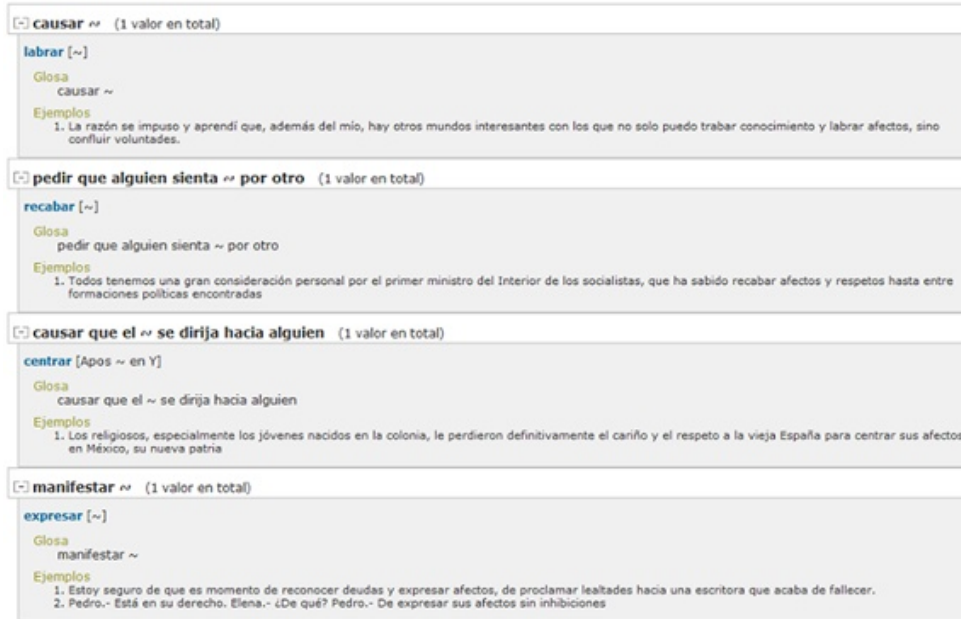


Figure 1: Display of the verb+noun collocations of *afecto1* ‘affection’ in DiCE

the desired information (in a sequential text format), the user needs either to introduce the base into an interface or select it (possibly in a cascaded menu) from a list (as is the case in DiCE). However, this traditional dictionary look up philosophy is not sufficient when the user is a language learner and the dictionary is supposed to serve as an instrument that supports active learning. Active learning is closely related to exploration and even more so in the context of active learning of collocations: collocations are idiosyncratic in that two bases with similar meanings may have different collocates (possibly with the same semantics; cf., e.g., *labrar afecto* ‘produce affect’ vs. *inspirar simpatía* ‘inspire sympathy’) or share the same collocates (as, e.g., *té* ‘tea’ and *café* ‘coffee’: *tomar un té/café*), deviate from a literal translation from L1 (as, e.g., *take [a] walk* vs. *dar [un] paseo*, lit. ‘give a walk’) or not (as, e.g., *give [a] talk* vs. *dar [una] conferencia*), etc. This can only be learned by navigating in the collocation spaces, by comparing, clustering, etc.

The most intuitive questions to explore in view of a collocation include, for instance:²

- Which other lexemes collocate with the base of this collocation and how common are these collocations (either compared to the given collocation or in absolute terms)?
- Which other bases take the collocate of this collocation (and, again, how common are these collocations)?
- What is the overlap of collocates of the given base with semantically similar bases?
- What is the typical context of this collocation?

² These and further similar questions can be derived from the didactic studies related to collocation learning; see, among others, (Hausmann, 1984; Lewis, 2000; Higuera García, 2011).

We shall now investigate how VA can help to explore these or similar questions and to provide the information that the user expects to encounter when consulting a collocation dictionary such as DiCE.

3. Visual Analytics Techniques and Collocation Information

In what follows, we first give a short introduction to Visual Analytics and then discuss techniques that we consider appropriate for the display and exploration of collocation information.

3.1 What is Visual Analytics?

Visual Analytics (VA) is a recent research area that emerged within the field of information visualization as a response to the need of (possibly unexperienced) users to explore new (usually large) information spaces; cf.: “Visual analytics is the formation of abstract visual metaphors in combination with a human information discourse (interaction) that enables detection of the expected and discovery of the unexpected within massive, dynamically changing information spaces.” (Wong and Thomas, 2004). Indeed, this is exactly what is expected by a learner who actively explores the “collocation space”. A great number of different visual metaphors have been proposed by the VA community for the exploration of different types of information spaces; see, e.g., <http://d3js.org/> for an extensive library. Among the most common visual metaphors are various types of networks (to visualize the connectivity between the elements of the explored space), trees (to visualize hierarchical relations between the elements of the space), flows (to visualize the change of the information space over a time line), glyphs (to visualize multidimensional data), etc. Figure 2 presents a fragment of a radial tree taken from <http://bl.ocks.org/mbostock/4063550>, cited in (Butt and Culy, 2014).³ The tree is interactive in that it can be collapsed, expanded, zoomed-in, etc.

The general principle underlying nearly all metaphors is “Overview first, zoom and filter, then details-on-demand” (Shneiderman, 1996). To facilitate an overview, data tend to be aggregated (clustered) with respect to specific features. The zoom allows for inspection of specific patterns or subsets of data by applying filtering. The “details-on-demand” displays the individual features, examples, etc. related to individual entities in the information space.

We shall now discuss how VA can be used for visualizing and exploring collocation information.

³ Note that Figure 2 does not represent any collocation-oriented information; it is displayed just for the sake of the illustration of the notion of a radial tree.

intended for specialists (e.g., data analysts), not (potentially formally untrained) end users such as language learners.

Therefore, to ensure an “easy-to-follow” dynamic interaction, we use *Gephi* as a library. We first generate static visualizations of graphs, export the resulting *Gephi* graph in *gefx* format, and subsequently visualize it by means of a web interface that we developed using the `sigma.js` JavaScript library⁵. `sigma.js` is among the best graph drawing JavaScript libraries available. Besides being easily customizable and having a lot of built-in features, such as Canvas and WebGL renderers or mouse and touch support, it provides a plugin system, so anybody can add code to implement any other functionality.

4. Towards Visual Exploration of Collocation Spaces

In order to be able to offer the functionality of the exploration of collocation resources as sketched in Subsection 2.2 above, these resources need to be preprocessed in several terms. Therefore, before we embark on the description of the implementation of VA, we present the preprocessing of the resources we use.

4.1 Preprocessing of the Collocation Resource

Our resource is a large Spanish newspaper syntactic dependency corpus treebank. The treebank has been indexed in Solr, with each sentence being captured in the index in three different ways in order to be able to retrieve the following kinds of information:

1. The sentence as it appears in the original corpus.
2. The sentence as a sequence of “PoS|lemma” tags, to allow for searches based on the lemma with a given PoS.
3. Sequences of lemmas with their parents in the dependency tree: For each lemma in the sentence, an element that includes the term, its PoS, the PoS of the parent, the lemma of the parent, the syntactic relation between the lemma and the parent and the position of both in the sentence. When there is a preposition between a verb and a noun, the preposition is removed and a direct relationship is created. This structure allows for searches such as a “lemma being a noun related to any verb” and finds these verbs and how they are related.

For the treebank’s binary word co-occurrences between which a direct syntactic dependency holds, the *collocate-weighted normalized pointwise mutual information* (NPMI_c) has been calculated as a measure of “collocality”; cf. (Carlini et al., 2014).⁶ Solr’s faceted search has been

⁵ <http://sigmajs.org/>.

⁶ In contrast to the standard PMI, as commonly used in Corpus Lexicography since (Church and Hanks, 1989), NPMI_c takes the asymmetric nature of collocations into account.

used in order to retrieve the information needed for the computation of the NPMI_c s, which are precomputed and stored in a relational database. The use of the relational database and Solr facilitates efficient access of individual tokens, lemmas, token/lemma – co-occurrences with NPMI_c s, syntactic dependencies, and example sentences (with their dependency structures) and real-time delivery of the corresponding information (including examples) via the user interface.

4.2 Realizing VA for Collocation Resource Exploration

In Subsection 2.2, we listed some questions concerning both individual collocations and collocation collections the exploration of which should be facilitated by use of a VA tool. In what follows, we present some of our realizations aimed to fulfil this demand.

4.2.1. Exploring the collocation space of a base

In order to help the learner to explore the collocability of a given base, the collocates of this base are clustered with respect to their context (and thus with respect to their distributional semantics) and displayed in terms of coloured circles. The size of a collocate’s circle indicates the commonality of the collocation formed by the base–collocate co-occurrence (more precisely, its size is proportional to its NPMI_c). Each cluster is displayed in a different colour. Figure 3 illustrates this kind of visualization for the collocation space of *té* ‘tea’. *Beber* ‘drink’ (cf. *beber té* ‘drink tea’) and *tomar* ‘take’ (cf. *tomar té*, lit. ‘take tea’) form one cluster (as a matter of fact, *beber* and *tomar* are synonymous in their role as collocates of *té*). A second, considerably more heterogeneous, cluster is formed by *preparar* ‘prepare’, *ofrecer* ‘offer’, *pedir* ‘ask for’, *servir* ‘serve’, and *compartir* ‘share’.

Café ‘coffee’ can be expected to share as base its collocates with *té*. However, given that, on the one hand, drinking coffee in Spain is much more common than drinking tea and, on the other hand, *café* is polysemous in that it can also refer, e.g., to a location or to a drink after lunch in general, the graph for *café* is considerably richer; cf. Figure 4. Thus, it also contains clusters related to breakfast (*desayunar* ‘have breakfast’), to the social event of drinking coffee (*invitar* ‘invite’, *compartir* ‘share’), which overlaps with the cluster of *café* as location (*frecuentar* ‘frequent’), and to coffee as a plant (*plantar* ‘plant’), etc.

To obtain a graph such as that of *té* or *café*, we first generate a weighted graph of nodes centred on the base, with all of its collocates that show a NPMI_c over a given threshold.⁷ The weighted graph is then clustered using the modularity algorithm presented in (Blondel et al., 2008) and as implemented by (Lambiotte et al., 2008) in *Gephi*.

⁷ We set the threshold to 0.2 since even if an NPMI_c higher than 0 indicates that the relation between both elements is beyond randomness, more significance is needed for the two elements to become a collocation and avoid noise.

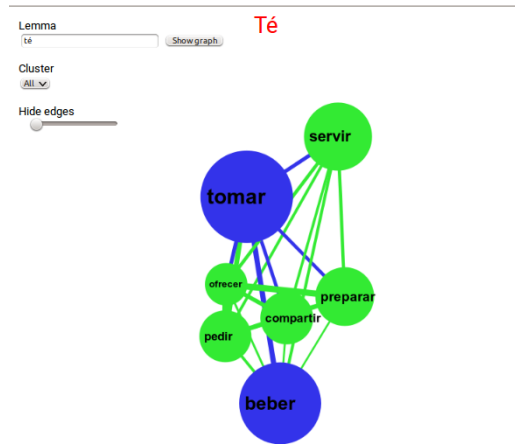


Figure 3: Collocation space of the base *té* ‘tea’

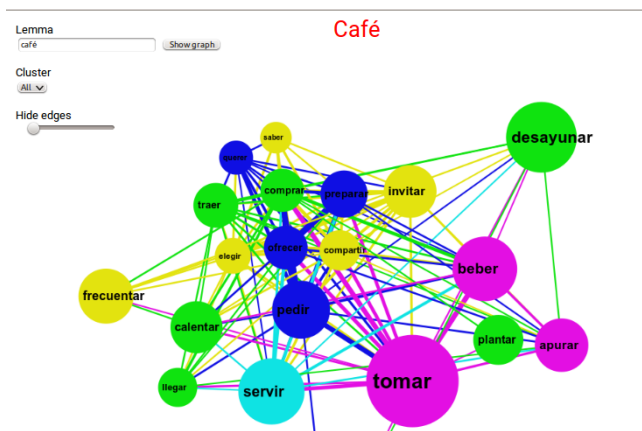


Figure 4: Collocation space of the base *café* ‘coffee’

4.2.2. Collocation space of bases sharing collocates

In order to move on from the exploration of the collocation space of a single base to a (contrastive) exploration of the space of several bases in parallel, the weighted graph from above is expanded by all bases of the collocates that are related to them with a $NPMI_c$ above the threshold via a specific syntactic dependency relation (e.g., direct object).⁸ With this action we obtain a bipartite graph of bases and collocates.

In a second step, we use *Gephi*'s multimodal transformation to find for every pair of collocates how many bases they have in common.⁹ This produces a reduced graph where only the collocates are present. However, as a rule, it is still a high density graph that is difficult to

⁸ In the current initial version of our VA experiments, we work by default with some prominent dependency relations such as ‘direct object’, ‘indirect object’ and ‘subject’. It is foreseen that the learner can choose the relations interactively via the interface.

⁹ <https://marketplace.gephi.org/plugin/multimode-networks-transformations-2/>

view and inspect. Therefore, the edges under a certain threshold are pruned to simplify the graph.¹⁰ For the spatial distribution of nodes, a force atlas is used and labels are adjusted to avoid label superposition. Once the collocates are clustered, the graph is expanded again with the bases.

Finally the elements in the graph are scaled such that:

- the size of the bases is the sum of the $NPMI_c$ s they have with the different collocates with the $NPMI_c$ of the collocate; in this way, bases that highly correlate with the source base appear bigger;
- the strength of the edges between collocates indicate how many bases they have in common;
- the strength of the edges between bases and collocates is proportional to their $NPMI_c$ s.

In Figure 5, the collocate selected to be in focus (*beber* ‘drink’) is represented as a hexagon, the other collocates as circles and the bases as triangles. The bigger the size of a base triangle the more collocates it shares with *té*.

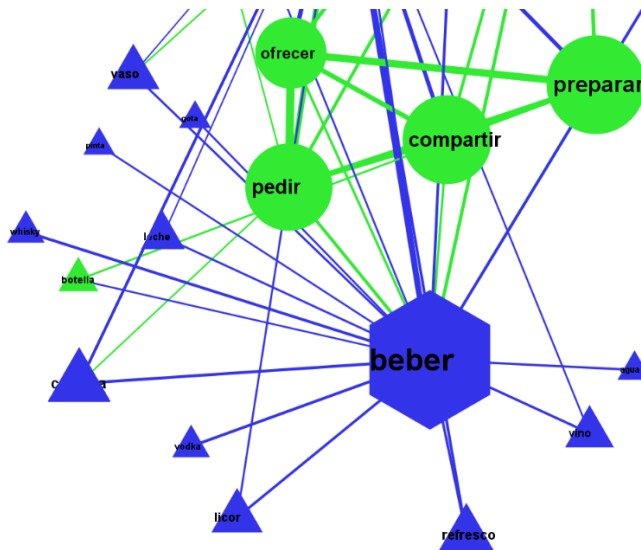


Figure 5: Collocation space of several bases related with *beber* and the original base *té*

4.2.3. Zooming in on the details of a collocate or collocation

The user may also want to further explore individual elements of the graph. This can be done using the “zoom-in” function. Thus, the user can, e.g., click on a collocate and obtain

¹⁰ After a series of tests, we set this threshold to 0.3, as it sufficiently reduces the density of the data.

information about it, get sample sentences with the use of the collocation formed by the collocates and the corresponding base, and the information regarding which other bases this collocates co-occurs (as in the initial setting, only those bases are displayed that have an $NPMI_c$ above the threshold). For instance, if we click on *apurar* ‘[to] drain’ we learn (see Figure 6) that *apurar* co-occurs with such bases as *cerveza* ‘beer’, *vaso* ‘cup’, and *copa* ‘glas’. Several examples from the corpus illustrate the use of *apurar* in context (in this case, its co-occurrence with *café*). Also, the learner can learn about the frequency of the collocates in the corpus and its $NPMI_c$.

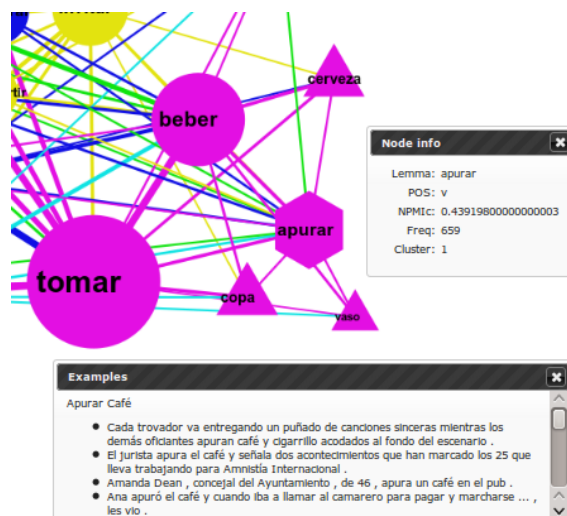


Figure 6: Zooming in on the collocates ‘apurar’ ‘[to] drain’

The user can also zoom-in on the link between a base and a collocates (i.e., on a specific collocation) to obtain examples; see Figure 7, where the user clicked on the link between *vaso* ‘cup’ and *apurar* ‘drain’ to obtain sample sentences in which *vaso* and *apurar* appear as collocation. The information regarding which collocates belong to the same cluster as *apurar* (namely *tomar* ‘take’ and *beber* ‘drink’) and with which prominence, and which other prominent clusters are involved in the collocation space of *vaso* (in this case, the cluster consisting of *servir* ‘serve’), is also displayed.

4.2.4. Navigation within collocation spaces

The user can navigate starting from the graph centred around a given base to a graph centred around one of the bases with which this base shares some of the collocates. This is done by double-clicking on the base the user wishes to look at. The obtained graph is obviously different from the starting graph because it is centered around the new base. Figure 8 shows the graph for *copa* ‘cup’, obtained departing from the graph of *café* ‘coffee’. The most prominent collocates for *copa* remain (as already for *café*) *beber* ‘drink’ and *tomar*

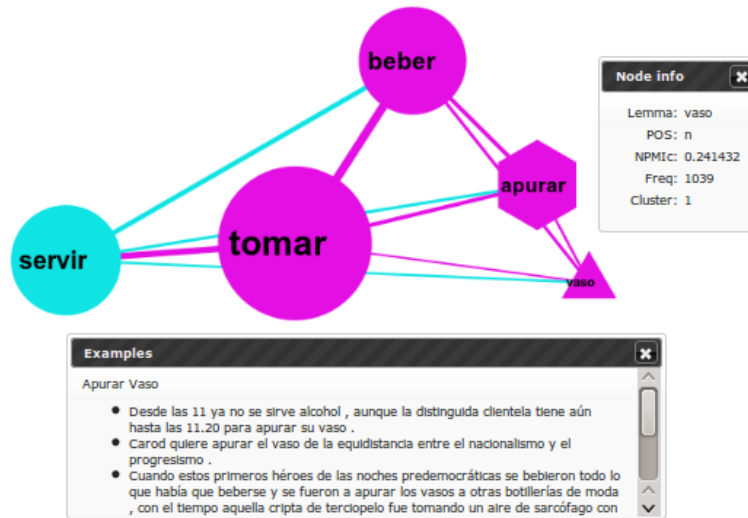


Figure 7: Zooming in on *apurar* [el] *vaso* ‘drain [the] cup’

‘take’, but it can also be observed that *copa* has a number of collocates not shared with *café*. In this context, it is important to notice that in all of the given graphs, the similarities and correspondences between bases are always calculated and displayed with respect to the subset of the collocates of the base in focus, not with respect to the full language model.

4.2.5. Exploration of collocate clusters

In Subsection 4.2.1, we already mentioned that collocates are clustered in accordance with their distributional semantics. An ideal clustering algorithm would group collocates with respect to a theoretically well-defined semantic collocate typology—as, e.g., *the lexical functions* (LFs) (Mel’čuk, 1996) or a generalization thereof. In DiCE, the glosses of the collocate groups in the individual entries for the bases (see Figure 1) are, in fact, LFs.¹¹ For automatic classification of given collocation lists in terms of LFs, see, e.g., (Wanner, 2004; Wanner et al., 2006; Moreno et al., 2013). In the current implementation of our VA tool, collocates are clustered according to the strength of the relationships between them (number of common bases) using the “Louvain algorithm” (Blondel et al., 2008) for community detection. This algorithm is graph-based and tries to optimize the modularity of the community.¹² Applied to the collocates, it groups those collocates that share more bases between them than with the other collocates.

Each base is assigned to the cluster of the collocates which show, in the co-occurrence with it, a $NPMI_c$ higher than the threshold. The user can restrict the visualization of the graph

¹¹ The interface of the DiCE also allows for the display of actual LF labels, along with the glosses.

¹² Modularity measures the relation between the density of edges inside communities to edges outside communities.

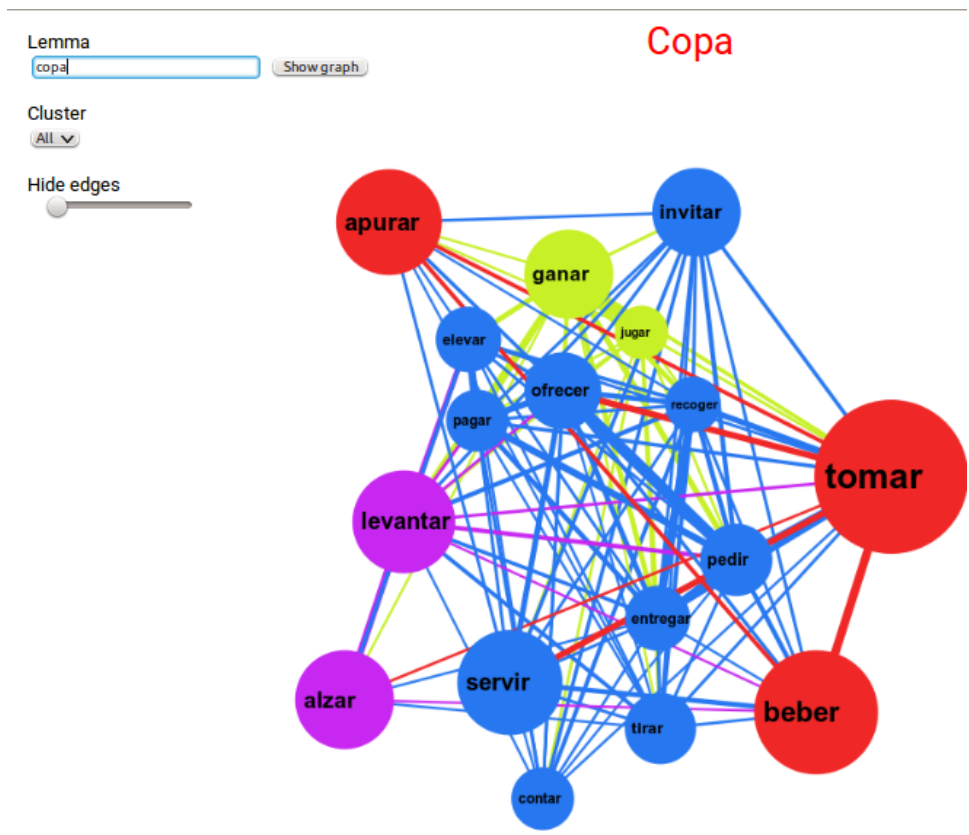


Figure 8: Navigating from *café* to *copa* ‘cup’

to a subset of nodes belonging to a single cluster. Figure 9 shows the resulting graph after selecting one of the clusters from the graph for *copa* ‘cup’.

5. Conclusions and Future Work

In this paper we presented some VA techniques for dynamic interactive exploration of collocation information, starting from the collocation space of a single base and either expanding it to the space of several bases or zooming in on the details of a single collocation. We believe that VA is crucial in all active learning environments, but particularly so in a collocation learning environment since collocations are idiosyncratic in their nature and thus require extra support for memorization.

The interface of the current implementation of our VA tool has been first realized as a standalone web application. It is now about to be built into the HARENES project interface (Wanner et al., 2013), where it will be integrated with other functionalities—for instance, that the learner can introduce a collocation, validate its correctness and obtain correction suggestions in case it is not correct, or introduce a whole text and receive correction sugges-

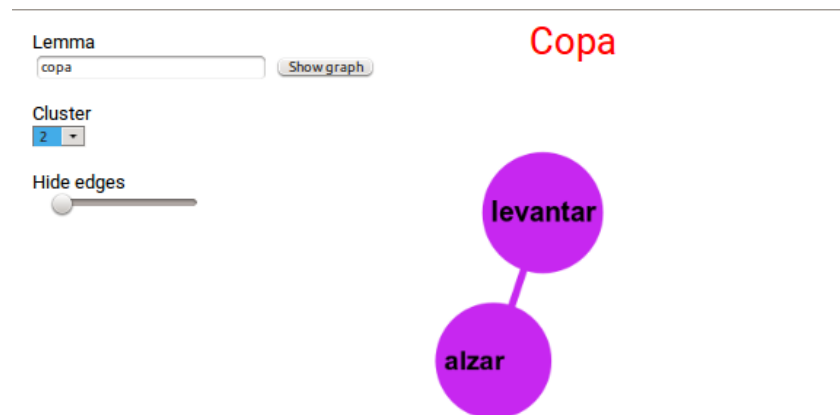


Figure 9: Selection of the collocates that belong to a cluster composed of the collocates *levantar* ‘raise’ and *alzar* ‘lift’ (in connection with *copa* ‘cup’ as a glass or as a trophy)

tions for the detected miscollocations. In this context, we plan also to experiment with the use of other collocate clustering (or classification) algorithms than the one that is used in the current VA tool—for instance, the one described in (Moreno et al., 2013).

The presented tool can be built into the interface of any online collocation dictionary such as DiCE, where it could be used to better visualize and explore the information that is available in this dictionary. However, prior to this integration, it must be evaluated—ideally in real language learning settings, involving students and teachers. Furthermore, it should be kept in mind that its current design does not necessarily follow rigorous didactic and/or visualization optimization considerations. A collaboration of specialists from these fields will be necessary to make the presented prototypical implementation a valuable aid in second language collocation learning.

6. Acknowledgements

The work presented in this paper has been supported by the Spanish Ministry of Economy and Competitiveness under the contract number FFI2011-30219-C02-02 in the framework of the HARENES Project, carried out in collaboration with the DiCE Group of the University of La Coruña.

7. References

- Alonso, M., Nishikawa, A., & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In Granger, S. & Paquot, M., editors, *Proceedings of eLex 2009, Cahiers du Cental 7*. Louvain-la-Neuve. Presses universitaires de Louvain, pp. 367–368.

- Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Menlo Park. AAAI Press.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*.
- Butt, M. & Culy, C. (2014). Visual analytics for linguists. ESSLI '14 Course material <http://www.sfs.uni-tuebingen.de/~cculy/courses/ESSLI2014/>.
- Carlini, R., Codina-Filba, J., & Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. *NEALT Proceedings Series Vol. 22*.
- Church, K. & Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pp. 76–83.
- Collins, C., Carpendale, S., & Penn, G. (2009). DocuBurst: Visualizing Document Content Using Language Structure. In *Proceedings of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis '09)*. Eurographics Association, pp. 1039–1046.
- Collins, C., Penn, G., & Carpendale, S. (2008). Interactive visualization for computational linguistics. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Morristown, NJ, USA. Association for Computational Linguistics.
- Feng, Y. & Lapata, M. (2010). Visual Information in Semantic Representation. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL and the Conference on Human Language Technologies*. Morristown, NJ, USA. Association for Computational Linguistics, pp. 91–99.
- Hausmann, F.-J. (1984). Wortschatzlernen ist kollokationslernen. zum lehren und lernen französischer wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(1), pp. 395–406.
- Higuera García, M. (2011). Lexical collocations and the learning of Spanish as a foreign language. In Cifuentes Honrubia, J. and Rodríguez Rosique, S., editors, *Spanish Word Formation and Lexical Creation*. Benjamins Academic Publishers, Amsterdam, pp. 439–464.
- Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual Analytics: Scope and Challenges. In Simoff, S., editor, *Visual Data Mining*, LNCS 4404. Springer Verlag, Berlin, pp. 76–90.
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). *The Sketch Engine: ten years on*. *Lexicography: Journal of ASIALEX*, 1(1), pp. 7–36.
- Lambiotte, R., Delvenne, J.-C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*.
- Lewis, M. (2000). *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

- Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in the lexicon. In Wanner, L., editor, *Lexical Functions in Lexicography and Natural Language Processing*. Benjamins Academic Publishers, Amsterdam, pp. 37–102.
- Moreno, P., Ferraro, G., & Wanner, L. (2013). Can we determine the semantics of collocations without using semantics? In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013: Electronic Lexicography in the 21st Century*. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 106–121.
- Penn, G. & Carpendale, S. (2009). Linguistic information visualization. ESSLI '09 Course material. http://esslli2009.labri.fr/documents/carpendale_penn.pdf.
- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pp. 336–343.
- Smith, K. L. (2005). *Handbook of visual communication: theory, methods, and media*. Rutledge, Oxford.
- Tufte, E. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CN, USA.
- Wanner, L. (2004). Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering Journal*, 10(2), pp. 95–143.
- Wanner, L., Bohnet, B., & Giereth, M. (2006). Making Sense of Collocations. *Computer Speech and Language*, 20(4), pp. 609–624.
- Wanner, L., Verlinde, S., & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of eLex 2013: Electronic Lexicography in the 21st Century*. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 472–487.
- Wong, P. C. & Thomas, J. (2004). Visual analytics. *IEEE Computer Graphics and Applications*, 24(5), pp. 20–21.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License

<http://creativecommons.org/licenses/by-sa/4.0/>



Predicting corpus example quality via supervised machine learning

Nikola Ljubešić, Mario Peronja

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences
University of Zagreb, Ivana Lučića 3, HR-10000 Zagreb
{nljubesi,mperonja}@ffzg.hr

Abstract

In this paper we present a supervised-learning approach to extracting good dictionary examples from corpora. We train our predictor of quality on a dataset of corpus examples annotated with a four-level ordinal variable, ranging from a very bad to a very good example. Each of the examples is formally described through 23 variables; the dependence of the quality of which is modelled using a regression model. The evaluation of the ranked results for each of the collocations in the annotated dataset shows that we obtain precision on 10 top-ranked examples of ~80% and a precision of ~90% on the three top-ranked examples. Our approach is highly language independent as well, suffering almost no loss on the 10 top-ranked examples and a loss of ~4% on the three highest-ranked examples once the language-dependent and knowledge-source-dependent features are removed.

Keywords: dictionary example; corpus extraction; supervised machine learning

1. Introduction

Corpus examples are a very welcome part of a dictionary entry. If a dictionary entry includes an example which is a good match for the context in which the user has encountered a word, or for the context in which they want to use it, then the user generally gets what they want in a quick and straightforward way. (Kilgarriff et al., 2008)

Finding good examples manually by looking through concordances in a corpus is very tedious and ranking concordances by the automatically estimated quality of the example is a very welcome addition to lexicographic processes.

The best known tool for finding good examples from a corpus is GDEX (Kilgarriff et al., 2008), part of Sketch Engine (Kilgarriff et al., 2004) where the lexicographer defines criteria for good examples using variables like sentence length, word frequency, pronouns, start and ending of a sentence etc., and has been adapted for a series of languages (Kosem et al., 2011).

In this paper we propose predicting the quality of a corpus example through the paradigm of supervised machine learning where we:

1. manually annotate a sample of examples for a given headword / collocation with its corresponding quality,
2. define features we consider informative for predicting the quality of a corpus example,
3. train a predictor, using features as explanatory variables and the manually assigned quality as our response variable, and finally
4. use that predictor to rank corpus examples of a headword / collocation by descending predicted quality of the examples.

Beside the prediction task, we measure the informativeness of each feature with the goal of better understanding the underlying phenomenon of what makes a good dictionary example extracted from a corpus.

We run our machine learning experiments by writing feature extractors in Python and performing all supervised learning tasks in scikit-learn (Pedregosa et al., 2011).

2. Dataset

The *conditio sine qua non* of our approach to predicting dictionary example quality is the sample of corpus examples, each of which is human-annotated with a quality score. On this dataset we extract variables, i.e. features we consider informative for predicting the quality of a corpus example for dictionary use. We use those variables and human scores to perform supervised machine learning, i.e. statistical modelling, in which we model the dependence of the response variable (the quality of an example) to the explanatory variables (the features extracted from each of the corpus examples) with the idea of predicting the quality of previously unseen corpus examples.

We extracted our corpus examples from the web corpus of Croatian (Ljubešić and Klubička, 2014). To produce a real-world-scenario sample, we built the dataset from sentences containing one of the 16 collocations chosen as a basis for building this dataset. The collocations were sampled from the hrMWELex lexicon of Croatian multiword expressions (Ljubešić et al., 2014). These 16 collocations consist of four mid-frequency lexemes, each belonging to an open-class part-of-speech: noun, verb, adjective and adverb. Given that we, as will later be described in detail, use shallow features such as sentence length and number of upper-cased tokens for predicting the quality of examples, and therefore do not try to model the deep, semantic criteria for a good example, we consider our dataset to be representative for predicting corpus quality of both collocations and single word units.

We finally produced a dataset of 1094 sentences randomly picked from all the sentences of the corpus containing any of the 16 collocations. Each collocation is thereby covered by 14 to 99 examples, which successfully mimics the scenario of extracting collocation examples from a medium-sized corpus.

It is important to note that, since the web corpus is annotated on the morphosyntactic and dependency syntax level, for each of the chosen sentences we had those two annotation layers at our disposal as well.

We annotated each of the 1094 sentences by the following four-level schema:

- 1 – very bad example, the example is useless
- 2 – bad example, most of the example should be rewritten
- 3 – good example, minor changes are necessary
- 4 – very good example, no changes at all are required

The *very bad* score was given to 14% of sentences, the *bad* score to 41.7% of sentences, while the *good* and *very good* scores were given to 33.3% and 11.1% of sentences respectively. This distribution of scores shows that the human annotator considered more than the half of the corpus examples as bad examples. A likely explanation for such a rather high percentage of examples being perceived as bad is that the data, although cleaned, still comes from the web where different types of noise are present on a regular basis.

3. Features

To be able to perform a quality prediction on our potential dictionary examples, i.e. sentences from a corpus, we have to transform each of those sentences into a set of variables. Given that these variables are used for performing the prediction, we refer to them as explanatory variables or features.

We defined altogether 23 features from three different categories: string-based features encoding properties of text on the string level, corpus-based features measuring the coverage of an example by the most frequent words from a corpus and linguistic features that use the linguistic annotation of the candidate example.

The string-based features are the following:

- `sent_len`: length of the sentence
- `avg_len`: average token length
- `gte10_perc`: percentage of tokens longer or equal to 10 characters
- `lt3_perc`: percentage of tokens shorter than 3 characters
- `alphanum_perc`: percentage of tokens being alphanumeric
- `alphanumpunc_perc`: percentage of tokens being alphanumeric or standard punctuations
- `startswithucase`: whether the sentence starts with an uppercase letter
- `endswithpunc`: whether the sentence ends with a punctuation

- `diac_perc`: percentage of tokens containing diacritics
- `lcase_perc`: percentage of lowercased tokens
- `ucase_perc`: percentage of uppercased tokens
- `tcase_perc`: percentage of titlecased tokens
- `headpos_perc`: relative position of the start of collocation

The corpus-based features were extracted with the help of a token frequency list compiled from the whole hrWaC web corpus. These are the features:

- `mf1k_perc`: percentage of tokens among the 1k most frequent corpus tokens
- `mf10k_perc`: percentage of tokens among the 10k most frequent corpus tokens
- `mf100k_perc`: percentage of tokens among the 100k most frequent corpus tokens

Finally, the linguistic features calculated from the two annotation layers present in the corpus, and thereby in each of our 1094 annotated examples, are thus:

- `pron_perc`: percentage of pronoun tokens
- `pn_perc`: percentage of proper noun tokens
- `num_perc`: percentage of numeral tokens
- `sub_num`: number of subordinating conjunctions
- `co_num`: number of coordinating conjunctions
- `subco_num`: number of conjunctions
- `syntcomplex`: syntactic complexity as the average length of the dependency arcs

To obtain the first insight into the informativeness of the features for the task at hand, we calculated the p-values for t-tests on each feature given the response variable transformed to a binary *good example / bad example* variable. In other words, for each feature we calculated the probability that the difference in the distribution mean of the feature among the good examples and the distribution mean of the feature among the bad examples occurred by chance. The results are given in Table 1.

Among the string-based features we can observe that the `sent_len` and `endswithpunc` features are the strongest predictors of the quality of the example. On the other hand, the only statistically insignificant differences are obtained with the `gte10_perc` and the `tcase_perc` features.

In corpus-based features, the coverage by the 1,000 most frequent words is shown to be statistically insignificant as well. As the number of the most frequent words increases, the p-value drops off.

Among the linguistic features, the probability of the difference in the means of the percentage of pronouns among good and bad examples is shown to be at very high 40%, indicating that

string-based	p-value	corpus-based	p-value
sent_len	7.0e-18	mflk_perc	0.0687
avg_len	5.7e-05	mf10k_perc	0.0008
gte10_perc	0.1087	mf100k_perc	1.7e-05
lt3_perc	9.9e-05		
alphanum_perc	4.1e-09	linguistic	p-value
alphanumpunc_perc	5.1e-05	pron_perc	0.4039
startswithucase	3.5e-04	pn_perc	0.0018
endswithpunc	2.7e-20	num_perc	0.0037
diac_perc	0.0064	sub_num	5.7e-08
lcase_perc	0.0063	co_num	7.4e-16
ucase_perc	0.0045	subco_num	1.3e-15
tcase_perc	0.0760	syntcomplex	8.2e-12
headpos_perc	0.0007		

Table 1: T-test p-values for each feature calculated on the feature distributions of good and bad examples

this feature is a bad predictor of the quality of an example. On the other hand, the number of coordinating conjunctions is shown to be a very good predictor. It is interesting to observe that the syntactic complexity of the example has also a very low p-value. One has to be cautious about drawing the conclusion that it is a strong predictor of example quality as it correlates very strongly (0.82) with the feature encoding the sentence length which has an even lower p-value.

4. Experiments and results

The first experiment focused on optimising our regressor. We performed a randomised search hyperparameter optimisation of our Random Forest regressor by doing 10-fold cross-validation. Our scoring function on the regressor was mean absolute error, i.e. the average absolute difference between the human-given quality and the predicted quality. The optimised regressor misses the human score on average by 0.52 points, while the non-optimised regressor produces a mean absolute error of 0.55 points.

In the second set of experiments we measured the ranking performance of our optimised regressor. We evaluated the ranked results via the precision-at-N metric which calculates the precision of the N highest ranked examples. We consider good and very good examples to be positive results and the bad and very bad examples to be negative results.

Since there are examples for 16 different collocates, we ran 16 iterations, during each we trained our regressor on examples of 15 collocates, and used the regressor to produce the ranked result for the left-out collocate examples. We calculated the final precision as the arithmetic means of the precisions of each collocate.

We compared the obtained result with a baseline system which orders the examples randomly and a ceiling system which orders the examples by the score given by the human annotator.

The results of this set of experiments are presented in Table 2. While the baseline gives a precision of around 50%, as expected, given the distribution of scores in the annotated dataset, the ceiling shows that each of the 16 collocations has at least five good examples, while the precision drops slightly when we consider the 10 highest-ranked examples of each collocate.

The result obtained with the four-level regressor is regressor_4. It has precision of $\sim 80\%$ to $\sim 90\%$, depending on the number of candidates taken into account, which is much closer to the ceiling than to the baseline.

The regressor_2 system is the one trained on two levels of the response variable only, i.e. it does not use the information about the difference between good and very good examples on one side, and bad and very bad examples on the other side. We can observe a minimal drop, showing that manually annotating the data with a two-level categorical variable is almost as informative for this task as our four-level ordinal variable.

	P@10	P@5	P@3
baseline	48.7%	48.7%	48.7%
ceiling	98.8%	100.0%	100.0%
regressor_4	78.8%	86.6%	89.3%
regressor_2	78.2%	86.2%	89.1%

Table 2: Precision on first N candidates obtained with the random baseline, the ceiling, and a regressor trained on 4-level and 2-level response variables

In the next experiment we considered using subsets of features only. We envisaged the following scenarios:

- regressor – using all features
- regressor_string – using string features only, i.e. not having (large) corpora at our disposal and the possibility of a linguistic analysis
- regressor_langind – using string features only without the percentage of diacritics as it could be considered specific for the Croatian language, thereby assessing how well our system could work on any other language

The results are presented in Table 3. The drop is surprisingly low when removing outer knowledge sources like the corpus and tools for linguistic analysis, showing a minor drop if 10 candidates are taken into account and a 3.7% drop on the first three candidates. Making the predictor language-independent adds an additional below-1% loss. It is important to

stress that the language-independent predictor would still need annotated data in the other language. Measuring the predictor performance in another language without retraining it on the data of that language could be very interesting and is left, as we do not have testing data for other languages, for future work.

	P@10	P@5	P@3
regressor	78.8%	86.6%	89.3%
regressor_string	79.0%	83.9%	85.7%
regressor_langind	78.4%	83.2%	85.0%

Table 3: Precision on first N candidates obtained with the regressor using all features, the regressor using string features only and the language-independent regressor

We finally depict the probability distribution of the examples of a specific quality obtained when using the baseline, and when taking into account the first 10, five or three top-ranked examples. These distributions are given in Figure 1.

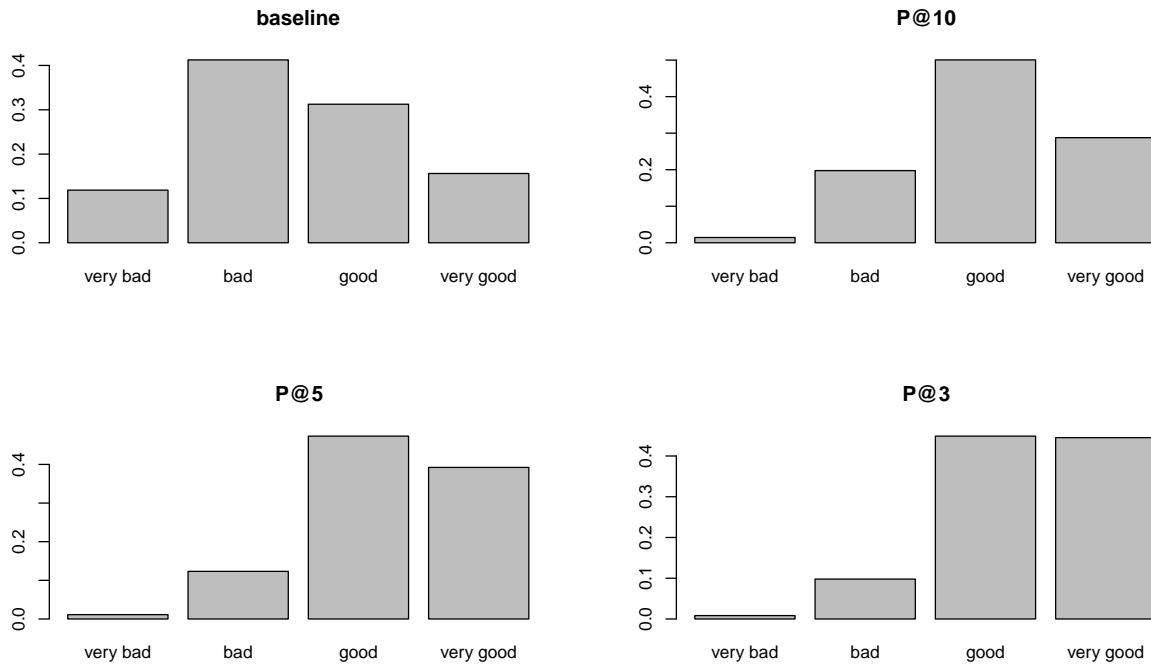


Figure 1: Probability distribution of the quality of the examples for the random baseline and the system taking into account first 10, 5 and 3 top-ranked examples

Having a better ratio between good and very good examples as we consider a lower number of highest-ranked candidates is expected. It points to the conclusion that the ranker manages

to produce the best results on the top of each output and that the results deteriorate as we move down the ranked output.

We can observe that we drastically outperform our baseline. While the best represented category in the baseline are bad examples, in P@10 and P@5 the good category is the most prominent one, the P@3 output having a similar amount of good and very good examples in the output.

5. Conclusion

In this paper we have presented an approach to extracting good corpus examples for dictionary use by using supervised learning, i.e. building a prediction model on a dataset on which the corpus example quality was already attested by a human. We argue that this approach is much more convenient than that used in GDEX where a lexicographer defines criteria for good examples by hand. Specifically, examples have to be annotated, or chosen, anyway, and such prediction algorithms have a steep learning curve, meaning that after annotating just a few examples, the ranking of the candidate examples improves drastically.

We have inspected the informativeness of each of the features used, showing that shallow features, such as the length of the example and the use of punctuation, and some less shallow features that are dependent on the shallow ones, such as the number of coordinating conjunctions, is most informative for the task.

In the ranking experiments we have shown to produce precision of $\sim 80\%$ on the first 10 candidates and $\sim 90\%$ on the first three candidates, which outperforms the random baseline of $\sim 50\%$ precision drastically.

We have shown that removing all external information sources, such as the corpus and its annotation, and language-dependent features, such as the percentage of diacritics, deteriorates our results among the first three top-ranked candidates slightly, lowering precision by $\sim 4\%$.

Our future work will involve two main directions of research. The first direction is testing the system on different languages and checking the language independence of the approach in both cases, when training data (i.e. annotated examples) in the new language is present, or when it is not and the model built on one language is applied directly onto the sentences of another language.

The second direction of our future work is the comparison of this approach with the rule-based approach, such as GDEX, where the (probably computational) lexicographer defines the criteria for a good dictionary example by hand.

6. Acknowledgements

The research leading to these results has received funding from the European Fund for Regional Development 2007-2013 under grant agreement no. RC.2.2.08-0050 (project RAPUT).

7. References

- Kilgarriff, A., Husák, M., Rundell, M., McAdam, K. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, Barcelona. Institut Universitari de Lingüística Aplicada. pp. 425–432.
- Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Information Technology*, 105:116.
- Kosem, I., Husák, M. & McCarthy, D. (2011). GDEX for Slovene. In *Electronic lexicography in the 21st century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011*. pp. 151–159.
- Ljubešić, N., Dobrovoljc, K., Krek, S., AntoniĆ, M. P. & Fišer, D. (2014). hrMWELex – A MWE lexicon of Croatian extracted from a parsed gigacorporus. In Erjavec, T. and Gros, J. Ž., editors, *Language technologies: Proceedings of the 17th International Multiconference Information Society IS2014*, Ljubljana, Slovenia.
- Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, Gothenburg, Sweden. Association for Computational Linguistics. pp. 29–35.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp. 2825–2830.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for specialized e-dictionaries

Ina Rösiger¹, Johannes Schäfer¹, Tanja George¹, Simon Tannert¹,
Ulrich Heid², Michael Dorna³

¹Institute for Natural Language Processing, University of Stuttgart, Germany

²University of Hildesheim, Germany

³Robert Bosch GmbH, Germany

[roesigia|schaefjs|georgeta|tannersn]@ims.uni-stuttgart.de,
heidul@uni-hildesheim.de, michael.dorna@de.bosch.com

Abstract

We report on ongoing experiments in data extraction from German texts in the domain of do-it-yourself (DIY) instructions, where the objective is (i) to extract nominal term candidates with high quality; (ii) to extract predicate-argument structures involving the term candidates, and (iii) to relate German word formation products with syntactic paraphrases: we focus on the analysis of compounds and on relating them with their syntactic paraphrases, in order to provide evidence for the (semantic) relationship between compound heads and non-heads (*Holzbohrer* (wood drill) \leftrightarrow *Holz_{Object} bohren* ([to] drill wood)). The extracted material is collected in order to provide structured data input for the creation of specialized dictionaries that are richer than standard terminological glossaries. For the creation of taxonomic knowledge (*Bandsäge -is-a* \rightarrow *Säge* (bandsaw \rightarrow saw)), we analyze subtypes of compounds.

Keywords: terminology extraction; raw material for specialized dictionary creation; lexical resources; German language; parsing

1. Introduction

There is a growing need for tools to extract terminology and relational data from text of specialized domains. Relational data involve verbal or adjectival predicates, their subjects, objects, complements, or preferred adjuncts; together with (mostly nominal) term candidates, they serve as a basis for ontology building and for the creation of raw material for dictionaries of the language of specialized domains.

The objective of the work described in this paper is the collection of German terminological data from heterogeneous corpora from the domain of do-it-yourself instructions. We use standard corpus linguistic technology for terminology extraction, as well as additional procedures for collecting and grouping related data with a view to the creation of a specialized lexical

resource. The procedures are based on automatic word formation analysis and on dependency parsing. While the use of parsing for term extraction is not new, dependency parsing for German of an appropriate quality has only been available for five years (Bohnet, 2010).

The remainder of this paper is structured as follows: Section 2 describes the specialized and general-language corpora used as a text basis for the extraction of term candidates. Section 3 presents the NLP tools and methods involved, and Section 4 gives an overview of the approaches designed to link the extracted term candidates, in order to collect raw material for a dictionary of specialized vocabulary.

2. Corpus data

Since our term extraction procedures rely, among other factors, on the comparison of specialized and “general language” texts, we work with corpora of both kinds.

As a domain-specific corpus, we use a corpus containing both expert and user-generated German texts from the DIY domain, which is composed, among other things, of manuals, practical tips, marketing texts and DIY project descriptions. The basic version of the corpus contains ca. 2.7 M tokens; in the course of this work, the corpus has been extended to 17.9 M tokens (see Tables 1 and 2 for details). The current versions of the corpus are not yet publicly available.

Text type:	#	tokens:	authors:
DIY manual	62,131		experts
DIY encyclopedia	6,868		experts
DIY practical “tricks”	15,104		experts
Marketing texts	35,302		experts
DIY project descriptions	2,160,008		UGC
FAQs (forum)	5,150		UGC
Wiki content	444,381		UGC
Total	2,728,944		

Table 1: DIY corpus

Text type:	#	tokens:	authors:
DIY manual	62,131		experts
DIY encyclopedia	6,868		experts
DIY practical “tricks”	15,104		experts
Marketing texts	35,302		experts
DIY project descriptions	4,479,437		UGC
FAQs (forum)	128,906		UGC
Wiki content	896,267		UGC
DIY articles	2,807,487		experts
Test descriptions	239,238		experts
DIY web encyclopedia	21,562		experts
Forum articles	296,242		UGC
DIY forum posts	7,873,115		UGC
Builders’ diaries	22,715		UGC
Video descriptions	2,280		UGC
Tool manuals	69,123		experts
Keyword lists	15,940		experts
Varia (no metadata)	961,236		-
Total	17,932,953		

Table 2: Extended DIY corpus

Our corpora are heterogeneous, as far as authorship and intended readership, text types and the level of specificity of the texts are concerned: while the manuals and the “tips and tricks” documents are written by experts (mostly for semi-experts or lay persons), a large portion of the texts comes from user-generated content (UGC) available in forums and thus likely

authored by semi-experts and/or lay persons. The corpus is intended to be a sample of the domain-related material available on the internet with a ratio of roughly 1:4 of expert vs. user generated content. In future work, we intend to separately analyze forum data and texts authored by experts, to assess specificities of each subcorpus.

As for the general-language corpus, we rely on the SdeWaC corpus (cf. (Faaß and Eckart, 2013)), a web corpus covering a wide range of topics and text styles, that contains around 880 M words. SdeWaC is a subset of deWaC (Baroni and Kilgarriff, 2006); it only contains sentences that can be parsed by the rule-based dependency parser FSPar (Schiehlen, 2003).

3. Computational linguistic technology used

The procedures used in our experiments are based on existing generic tools:

- A hybrid term extractor based on the prototype designed in the EU project TTC (*Terminology Extraction, Translation Tools and Comparable Corpora*, FP-7, STREP 248005, (Gojun et al., 2012a), (Gojun et al., 2012b) cf. Section 3.1);
- the dependency parser included in the *mate* tools (Bohnet, 2010), (Björkelund et al., 2010), as well as a tool that annotates syntactic phrases (and their boundaries, implicitly), cf. Section 3.2 and 3.3;
- the compound splitting tool CompoST (Cap, 2014), cf. Section 3.4.

We intend to combine the output of the tools in such a way as to be able to accumulate, from the corpus, the raw material for lexical entries that cater for term variation, partial taxonomies and the description of other, non-taxonomic relationships between concepts denoted by terms of the domain.

In the following, we briefly describe the three types of computational linguistic tools mentioned above.

3.1 Term extraction tools

The term extractor used in our work is a prototype based on a tool for German developed in the TTC project (Gojun et al., 2012b). It is a hybrid tool combining linguistic corpus preprocessing with statistical domain specificity ranking. Figure 1 schematizes the main steps of the tool pipeline.

The pipeline involves the following components:

- Preprocessing:



Figure 1: Steps in term candidate extraction: overview

- Tokenization: sentence and word form delimitation and markup;
- word class tagging and preliminary lemmatization: annotation by means of the RF-Tagger (Schmid and Laws, 2008), including an annotation as “unknown” of word forms absent from the tagger lexicon;
- lemmatization: specific treatment of the word forms absent from the tagger lexicon, with a view to guessing their lemma, by use of word form similarity, inflection-based rules and compound splitting; this component provides lemma forms for most of the “unknowns” which remained after the first lemmatization step.

The preprocessing steps of POS-tagging and lemmatization involve a simple form of domain adaptation: as the tagger used in the first run marks which word forms are not contained in its dictionary (“unknowns”, with respect to the data acquired in standard training from newspaper texts), these can be handled in the above mentioned specific lemmatization step which uses morphological knowledge and similarity data to guess lemma values. In future work, this set of procedures will be combined with Named Entity Recognition tools to make it more robust to new domains.

The preprocessing annotations are stored in a one word per line format.

- Pattern-based term candidate extraction: use of simple as well as extended POS-based patterns to identify term candidates; typical basic patterns are simple nouns, adjective+noun groups and nouns followed by genitive or prepositional modifiers. For verbal term extraction, patterns based on dependency parses are used, cf. Section 3.2.
- Ranking: sorting of the candidate lists produced by the preceding step, according to different measures: a basic approach uses (Ahmad et al., 1992)’s “weirdness ratio” (quotient of relative domain corpus frequency by relative general-language corpus frequency), while more advanced versions involve further measures, such as the C-Value measure ((Frantzi and Ananiadou, 1999); cf. (Schäfer, 2015) for details).

The output of the above steps are term candidate lists by patterns; examples of each pattern are given below:

N	<i>Bohrmaschine</i> (drill)
Adj+N	<i>oszillierende Säge</i> (oscillating saw)
N+Det+N _{genitive}	<i>Kopf einer Schraube</i> (head of a screw)
N+Prep+N	<i>Handkreissäge mit Führungsschiene</i> (skill saw with guide rail)

In addition to the basic patterns, and in line with Daille’s notion of term variants (Daille, 2007), more complex patterns are processed in the same way. The set of extended patterns is described by the regular expressions given below:

- ((Adv)? (Adj)? Adj)? N
- (N Det)? ((Adv)? Adj)? N Prep (Det)? ((Adv)? Adj)? N
- ((Adv)? Adj)? N Det ((Adv)? Adj)? N_{genitive}

3.2 Extracting verb object pairs from dependency parsed text

Standard term candidate extraction typically focuses on nouns and nominal phrases as they cover the objects of the domain (see patterns above). For the extraction of relational knowledge and to put the domain objects into context, verbally expressed relations are needed as well. We thus want to apply a variant of the above mentioned term extraction pipeline, i.e. the selection of candidates via linguistic preprocessing combined with a statistical ranking, also to verbal term candidates. The problem that arises is that the POS-based tool has no information about syntactic phrases and their boundaries, such that a part-of-speech-based approach is not sufficient, particularly for a language like German that has three models of verb placement and allows flexible word order.

For the verbal candidate extraction, pre-processing thus includes a separate dependency parsing step, followed by a script that extracts verb object (or subject verb) pairs which are then processed by the statistical filtering step. This treatment leads to local information which can be considered as a combination of dependency syntactic and constituent structural knowledge; it is thus richer than mere dependency annotations as provided, for example by Constraint Grammar.

To find suitable verb candidates and their corresponding subjects and objects, we use the dependency parser contained in the *mate* tool package (Bohnet, 2010), (Björkelund et al., 2010) to annotate the texts with dependency syntactic analyses; the parser is trained on a dependency version of the TiGer treebank (Brants et al., 2004), (Seeker and Kuhn, 2012) which contains newspaper texts; there is no domain-specific treebank available. However, the tool profits from the domain adaptation of the pre-processing steps, i.e. lemmatization and POS-tagging. We are currently investigating ways to adapt the dependency parser to the domain without the rather expensive creation of manual gold data.

As we are interested in verb+object (or subject+verb) pairs irrespective of whether the pair occurred in the active or passive voice, we apply an approach that annotates passive sentences with grammatical functions that correspond to the active voice version so that all corpus sentences can be handled in the same way in the pattern-based term extraction step.

3.3 Annotation of syntactic boundaries

The dependency parser can also be used to improve nominal term extraction by making sure that noun phrase candidates are syntactically valid. Term candidates covering excessively long spans typically occur in NPs followed by a PP, when part of the extracted candidate is actually attached to the verbal phrase, e.g. in (1) and (2). The invalid term candidates are underlined and marked with an asterisk. In these cases a phrase boundary ([NP][PP]) is found within the extracted string, and the (terminological) NP and the subsequent PPs are sisters. Valid term candidates would consist of a complex NP where the PP is embedded. We filter the output of the POS-pattern based extraction by using *mate* to find start and end points of NPs.¹

- (1) die *Vorlage mit Sprühkleber besprühen (spray the *template with paint)
- (2) ein *Loch in die Wand bohren (drill a *hole into the wall)

The boundary violation filter works as follows: if one or more words of the selected term candidate go beyond the phrase boundary, the candidate is not counted as a valid occurrence of this particular lemma sequence. The candidate sequence is not removed from the list of possible candidate terms, as other occurrences might not violate syntactic boundaries. The filter is thus a “soft” one as it only affects the frequency of the lexeme combination candidate. We also experiment with a “hard” filter, where the lexeme combination candidate is removed altogether as soon as an invalid candidate occurrence is found.

3.4 Compound splitting

For compound splitting we use CompoST (Compound Splitting Tool, (Cap, 2014)), a compound splitter which combines the use of a rule-based morphology system (SMOR, (Schmid et al., 2004)) with subword (i.e. morpheme) verification in corpus data, thereby extending and improving on the approach proposed by (Koehn and Knight, 2003) for statistical machine translation: for all components of a compound, including those which are complex themselves, the tool verifies the presence and number of occurrences in a (set of) texts; in our application, the do-it-yourself corpus is used as a knowledge source for this check, in addition to a (newspaper-based) general language corpus. Splits that involve implausible or rare components are dispreferred.

¹ In current experiments only for NPs in subject or object position; work towards covering all relevant construction types is ongoing. We are aware that *mate* has not been optimized to solve the PP attachment problem.

For specialized terms, taking a domain corpus as the basis for the computation of probable splits often has the effect that wrong splits based on general-language frequencies (*Betonverbinder* (*concrete connector*) split into *Beton(concrete)|verb(verb)|inder(indian)*) are avoided and the right splits are produced (*Beton(concrete)|verbinder(connector)*). The tool allows a set of parameters, such as to show all possible splits or just the most probable one, and to decide whether the output should contain surface forms or lemmatized forms, to name only a few.

3.5 Quality of the term candidate extraction

The performance of the basic pipeline (cf. Section 3.1) has been evaluated on a gold standard data collection created from the 2.7 M words corpus described above in Section 2.

The gold standard (GS) was annotated manually by three independent experts; only term candidates with a minimum frequency of four and pertaining to one of the basic patterns (Section 3.1) were annotated, following predefined guidelines (cf. (George, 2014)). The candidates based on the extended patterns and the verbal candidates have not yet been evaluated against a gold standard.

We obtained a strict and a liberal version of the gold standard, where the strict GS only contains items for which full agreement on their term status was found. The total GS contains 4,238 single-word terms and 859 multi-word terms. The strict GS contains 2,777 terms, while the liberal GS includes additional 2,320 term candidates. The inter-annotator agreement ranges between moderate and substantial agreement (Landis and Koch, 1977), cf. Table 3.

annotators:	κ of N+“von”+N:	κ of N+Det+N _{gen} :	κ of N:	κ of Adj+N:	κ of N+Prep+N:
A1&A2	0.69	0.47	0.50	0.55	0.63
A2&A3	0.65	0.60	0.54	0.54	0.65
A3&A1	0.71	0.48	0.48	0.52	0.60
A1, A2&A3	0.68	0.52	0.51	0.54	0.63

Table 3: Inter-annotator agreement for the gold standard data. Interpretation of the kappa values: 0.41 – 0.6 = moderate agreement; 0.61 – 0.8 = substantial agreement.

We automatically evaluated the output of our pipeline computing precision, recall and f-measure for each of the basic patterns. Table 4 contains the results obtained on the liberal gold standard.

We furthermore compared the term candidates extracted from our corpus with a commercial tool (SDL MultiTerm Extract, version May 2014²) which is based exclusively on statistical

² <http://www.sdl.com/de/exc/language/terminology-management/multiterm/extract.html>

	N+“von”+N	N+Det+N _{gen}	N	Adj+N	N+Prep+N
Precision	72%	65%	52%	38%	55%
Recall	84%	91%	85%	55%	73%
F-measure	78%	76%	65%	45%	63%

Table 4: Precision, recall and f-measure values for the basic patterns compared with the liberal gold standard

procedures; while that tool is applicable to many languages without any need for language-specific knowledge, it is clearly outperformed on the German data by our prototype (George, 2014).

So far, no extensive GS-based evaluation of the effect of the phrase boundary check has been performed. However, tendencies can be observed: for the 107 terms of the GS which show the POS pattern “Noun+Preposition+Noun”, an improvement in precision is found both with the “soft” and with the “hard” filter. For the term candidates extracted on the basis of the extended patterns, we also checked the top-500 candidates that contained a preposition, and we determined whether the removal from the candidate list which was suggested by the filter was justified: it achieved, on that sample, 83% precision. This means in four out of five cases the removed candidate was indeed violating syntactic boundaries.

4. Collecting raw material for a dictionary of specialized vocabulary

In this section we show how the corpus data and the above mentioned processing tools can be used to relate the term candidates extracted, with a view to the provision of a maximal amount of structured raw data for subsequent (manual) lexicographic work.

We do not aim to automate the creation of a specialized dictionary, but we intend to provide rich input for the lexicographic process. The focus in this paper is on term variants (in the sense of (Daille, 2007)) and on partial taxonomies. We explain different procedures used for this purpose, and we give examples of the output of each one. As we report on ongoing work, no quantitative evaluation of these procedures is yet available.

4.1 Analyzing variation in multi-word terms

As discussed in Section 3.1, we use basic POS patterns for the extraction of multi-word term candidates as well as extended ones which we relate in a meaningful way to the basic patterns, as suggested by (Daille, 2012). We consider a term candidate with an extended pattern to be a variant of a term candidate with a basic pattern if it contains the tokens of the basic one (in the same order). The term candidates with basic patterns are in turn retrieved by seeding the extractor with the nouns from our gold standard.

The relationships observed in the data can be subdivided into the following three types:

(1) Variation:

– Example:

Verkleidung aus Rigipsplatten (cladding made of plasterboard) ↔
Gipskartonplatten als Verkleidung (plasterboard as cladding)

(2) Subtype relations:

– Example: Adj N → Adv Adj N:

weiße Farbe (white paint) ↔
matt weiße Farbe, normal weiße Wandfarbe, weißlich durchsichtige Farbe
(flat white paint, normal white wall paint, whitish sheer paint)

– Example: N → Adj N:

Schraube (screw) →
spezielle Schraube, passende Schraube, kleine Schraube, lange Schraube
(particular screw, appropriate screw, small screw, long screw)

(3) Relations of non-taxonomic type, e.g. focusing on aspects of an item:

– Examples:

* Adj₁ N₁ → N₂ ((Det₁) Adj₁ N₁)_{genitive}:

bodengleiche Dusche (walk-in shower) → *Aufbau einer bodengleichen Dusche*
(construction of a walk-in-shower)

* Adj₁ N₁ → N₂ Prep ((Det₁) Adj₁ N₁):

bodengleiche Dusche (walk-in shower) → *Anschluss an die bodengleiche Dusche*
(connection to the walk-in-shower)

4.2 Analyzing compounds for the creation of taxonomic knowledge

Many specialized compounds are transparent, compositional determinative compounds and thus their head denotes their hypernym: *Kreissäge* (buzzsaw) “is-a” *Säge* (saw). On this (simplistic) assumption, compound splitting and the identification of heads allow for a grouping of items according to subtype relations. For example, starting from a simplex term (e.g. *Säge*, saw), all compounds could be identified that have this term as a head (e.g. *Bandsäge* (band-saw), *Kreissäge* (buzzsaw), etc.), and a subtype relation could be assigned. This strategy could be applied recursively to create a partial hierarchy from more general to more specific terms (such as, e.g. *Säge* → *Bandsäge* → *Horizontalbandsäge* (horizontal bandsaw)).

The implementation differs from this principle, in order to correctly cover multimorphemic non-head elements: it takes a compound, splits it into morphemes, removes the first one and tries to find occurrences of the remaining part in the corpus. If, for example, it starts from *Eigenbaubandsäge* (self-made bandsaw) (split as *Eigen·bau·band·säge*), it will check the corpus for ^{??}*Baubandsäge*, and it will not find any occurrence. It then skips the element *-bau-*

and checks for *Bandsäge*, where a sufficient number of occurrences are found. As we work on compounds from the domain, not finding an item in the corpus will most often mean that this item does not exist (as the hypothetical form ??*Baubandsäge*); obviously, a few cases may also be due to data sparsity. The full set of subtypes of *Bandsäge* (bandsaw), as found in our data, is summarized in Table 5. An exemplary hierarchy for the term *Säge* (saw) is given in Figure 3.

Eigenbaubandsäge (self-made bandsaw)	Eigen Bau Band Säge
Elektro-Bandsäge (electric bandsaw)	Elektro Band Säge
Hand-Bandsäge (hand bandsaw)	Hand Band Säge
Horizontalbandsäge (horizontal bandsaw)	Horizontal Band Säge
Vertikalbandsäge (vertical bandsaw)	Vertikal Band Säge
Metallbandsäge (metal bandsaw)	Metall Band Säge
Minibandsäge (mini bandsaw)	Mini Band Säge
Bandsäge (bandsaw)	Band Säge

Table 5: Subtypes of *Bandsäge* (bandsaw) in the corpus

For the term *Säge* (saw) we gathered and manually verified the partial ontology constructed from the compounds analyzed in this way. Of 213 compound candidates, 36 candidates are not found in the corpus, because the compounds do not exist in German or because the forms used as an input to the procedures contain typographic errors.

4.3 Analyzing syntactic paraphrases of compounds

We use the parsed version of the corpora to identify potential syntactic paraphrases of German noun compounds; examples include nouns with genitive attributes (*Holzmaserung – Maserung des Holzes* (grain of wood)) and nominals with PPs (*Wasserkontakt, Kontakt mit Wasser* (contact with water)) as well as verb+object collocations (*Temperaturerhöhung – Temperatur+erhöhen* (increase (in) temperature)).

4.3.1. Compounds with nominal heads

We acquire paraphrases for compounds with nominal heads by querying noun+preposition+ noun or noun+determiner+noun (in genitive case) patterns in the 17.9 M corpus. Searching for syntactic paraphrases (synt) of nominal compounds (cmpd) serves two different purposes of lexicographic relevance:

- (i) quantitative aspects: to find more instances of an item, by grouping term variants together:

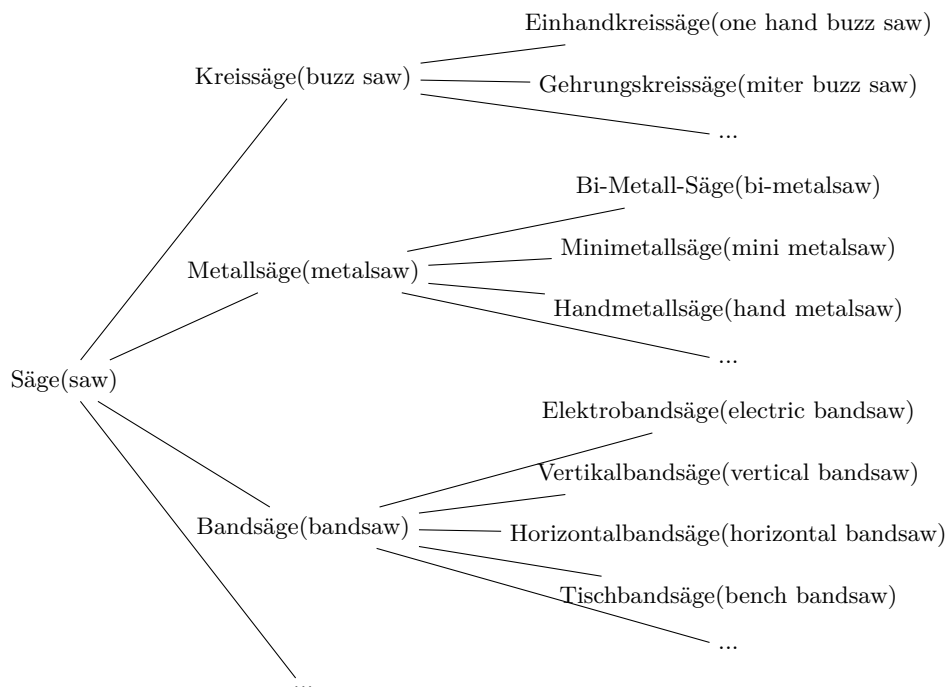


Figure 3: Sample of a partial hierarchy of the term candidate *Säge* (saw)

	f_{compd}	f_{synt}	\sum
- <i>Schraubenloch</i> (screw+hole) ↔ <i>Loch für Schraube</i> (hole for screw)	441	15	456
- <i>Raummitte</i> (room+centre) ↔ <i>Mitte des Raumes</i> (centre of the room)	37	57	94
- <i>Holzmaserung</i> (wood+grain) ↔ <i>Maserung des Holzes</i> (grain of the wood)	136	56	192
- <i>Brettkante</i> (board+edge) ↔ <i>Kante des Brettes</i> (edge of the board)	79	41	120

(ii) to derive the semantic relation existing between the compound head and the non-head:

	f_{compd}	f_{synt}	\sum
- <i>location: Fliesenfuge</i> (slab+joint) ↔ <i>Fuge zwischen Fliesen</i> (joint between slabs)	110	17	127
- <i>material: Teakmöbel, Teakholzmöbel</i> (teak(wood)+furniture) ↔ <i>Möbel aus Teak</i> (furniture made of teak)	7(+8)	21	28
- <i>material: Beton-Fundament, Betonfundament</i> (concrete+basement) ↔ <i>Fundament aus Beton</i> (basement made of concrete)	127(+22)	21	148

With respect to the first objective, a simple case is the collection of all possible “genitive” forms: next to the rare item *Loch bohren* (drill a hole) ($f = 7$), we find *Bohren des Lochs* (drilling of the hole) (103), *Bohren eines Lochs* (drilling of a hole) (6), *Bohren von Löchern* (drilling of holes) (8). These procedures allow us to collect all morphosyntactic variants of a collocation, i.e. verb+object (*Temperatur erhöhen* (increase temperature)), nominalisation of the verb+genitive (*Erhöhung der Temperatur*), compound (*Temperaturerhöhung*) and, if the lexicographer regards this as a separate type, attributive participle (*erhöhte Temperatur*). We are aware that these “variants” are not necessarily fully synonymous. Specialized languages in addition tend to be highly selective with respect to the choice among these variants as shown by (Fritzingler and Heid, 2009) for a subdomain of juridical language. A more difficult task is that of relating compounds with appropriate noun+PP paraphrases.

While some compounds only have one paraphrase, or only one statistically prominent paraphrase, others have several potential paraphrases, especially those which are truly polysemous. An example of this last case is *Holzfarbe* (wood+colour): it is polysemous and denotes (a) the colour of wood or (b) (synthetic) colours designed to paint wood. Both readings show up in our corpus, but the first reading is most prominent in the syntactic paraphrase data. For a disambiguation of the compound occurrences (e.g. to provide example sentences for the lexicographer), we intend to rely on indicator items from the context, e.g. (semantic) types of adjectives preceding *Holzfarbe* (*graue* (*gray*), *weiße* (*white*), ... → colour to paint wood; *originale* (*original*), *natürliche* (*natural*), ... → colour of wood).

The taxonomy of compounds with a specific head noun (as in Figure 3) can now be enriched with the semantic relations acquired from the noun+PP paraphrases, which makes it possible to group the subtype items. Table 6 presents an excerpt from a detailed analysis of compounds of the noun *Schraube* (screw) and their paraphrases where the compounds are grouped by the semantic relation between the compound head and the non-head.

material:	preposition: <i>aus</i> (made of)
<i>Stahlschraube</i>	↔ <i>Schraube aus Stahl</i> (steel screw)
<i>Edelstahlschraube</i>	↔ <i>Schraube aus Edelstahl</i> (stainless steel screw)
<i>Kupferschraube</i>	↔ <i>Schraube aus Kupfer</i> (copper screw)
application:	preposition: <i>für</i> (for)
<i>Rigips-Schraube</i>	↔ <i>Schraube für Rigips</i> (screw for plasterboard)
type:	preposition: <i>mit</i> (with)
<i>Senkkopf-Schraube</i>	↔ <i>Schraube mit Senkkopf</i> (countersunk head screw)
purpose:	preposition: <i>als/zu</i> (as/to)
<i>Führungsschraube</i>	↔ <i>Schraube als Führung</i> (screw as a guide)
<i>Befestigungsschraube</i>	↔ <i>Schraube zu Befestigung</i> (screw as a fixing)

Table 6: Compounds with the head *Schraube* (screw) and their paraphrases

Finally, there are cases where the compound is not paraphrased adequately in the corpus; equally, more work needs to be done to remove spurious paraphrase candidates:

- *Treppenraum* (*stairwell*) ↔ *Raum unter der Treppe* (*room under stairs*),
↔ *Raum zwischen Treppe und Wand* (*room between stairs and wall*)

Overall, the simple procedures sketched above produce relatively good results; a precision evaluation of a sample is planned.

Compound	Object +	Verb
Temperaturerhöhung (temperature rise)	Temperatur (temperature)	to rise (erhöhen)
Temperaturmessung (temperature measurement)	Temperatur	messen (to measure)
Temperaturregelung (temperature control)	Temperatur	regeln (to control)
Temperaturüberwachung (temperature monitoring)	Temperatur	überwachen (to monitor)
Dübellochbohrer (dowel hole drill)	Dübelloch (dowel hole)	bohren (to drill)
Fliesenbohrer (tile drill)	Fliesen (tile)	bohren
Holzbohrer (wood drill)	Holz (wood)	bohren
Kreisbohrer (circle cutter)	Kreis (circle)	bohren
Kunststoffbohrer (plastic drill)	Kunststoff (plastic)	bohren
Langlochbohrer (deep-hole drill)	Langloch (deep hole)	bohren
Maschinenbohrer (machine drill)	?? Maschinen (machine)	bohren
Nagelbohrer (nail drill)	?? Nagel (nail)	bohren
Pfostenbohrer (jamb drill)	?? Pfosten (jamb)	bohren
Diamantbohrer (diamond drill)	NOT: *Diamant (diamond)	bohren

Table 7: Deverbal compounds and their syntactic paraphrases for *Temperatur* (temperature) and *Bohrer* (drill)

4.3.2. Compounds with verbal heads

For deverbal compounds, we aim to distinguish different relations between the head and the non-head by analyzing the presence (or absence) of certain syntactic paraphrases, e.g. verb object pairs. The following section describes our experiments on linking deverbal compounds and their corresponding verb object pairs. In the future, we also plan to investigate subject verb pairs or other constructions that put the involved term candidates into context, such as predicative expressions.

For deverbal heads and their respective non-heads, there is a variety of possible relations between the two. If we take *Bohrer* (drill), for example, we can find a number of different semantic relations: *Diamantbohrer* (diamond drill) exemplifies an **is-made-of** relation where the non-head describes the material of which the drill is made, whereas a *Holzbohrer* (wood drill) is used to drill wood. Here, the non-head specifies the object to be drilled.

Thus, in our ongoing work, we first extract all deverbal compounds and the corresponding verb (a total of 8,750 compound types with verbal head and nominal non-head are present in our corpus) and then look for the respective verb object pairs in the dependency parses where the object equals the non-head of the compound. We then sort the extracted paraphrases by the nominal non-head (as in the first example in Table 7) and find events involving the noun, or we can sort by the deverbal head (as in the second example in Table 7) and find typical objects of the verb.

Table 7 shows the compounds and their matching paraphrases for two examples, *Temperature* (temperature) as a non-head and *Bohrer* (drill) as a head. When we find a verb object pair for a certain compound, e.g. *Kunststoffbohrer* (plastic drill), we now know that it is used to *drill plastic*. For *Diamantbohrer* (diamond drill) we do not find such a paraphrase. This confirms

our claim that the relation between the head and the non-head in this case is a different one, i.e. a **is-made-of** relation. In some cases, Noun+PP-evidence confirms this classification, cf. *Hartmetallbohrer* (tungsten carbide drill) \leftrightarrow *Bohrer aus Hartmetall* (drill made of carbide).

While a quantitative analysis of this automatic linking approach has not yet been performed, we have found a total of 7,411 occurrences of verb object pairs for our 8,750 compound types (1,381 unique verb object pairs). The reported links have been created on the basis of the 2.7 M corpus. We are currently performing experiments on the 17.9 M corpus, which will increase the coverage of matching paraphrases for the candidate terms extracted by the term extractor. We think that the number of links found is large enough to be beneficial for the creation of a specialized dictionary.

4.4 Lexicographic use of the collected data

The procedures discussed in section 4 of this paper are all meant to support human lexicographers in the preparation of entries of an online dictionary. The targeted dictionary is meant to be both a resource for human use and a knowledge source of automatic or semi-automatic tools, e.g. for e-mail routing, knowledge extraction from texts, as well as passage retrieval.

A possible interactive version of the dictionary would be characterized, among other factors, by the following properties: (i) it is a monolingual specialized dictionary allowing both semasiological and onomasiological access (the latter through the (partial) taxonomies constructed according to the procedures described in section 4.2); (ii) it goes beyond the structure and descriptive programme of terminological databases, insofar as it has not only nouns, but also verbs as lemmata and because it relates action-denoting verb+object pairs with terms; (iii) we foresee the possibility to add other languages to the dictionary.

The raw material gathered by means of the devices discussed in section 4 will serve the lexicographers as an input: it is not intended to create the lexicographic product fully automatically. The objective is to combine all evidence gathered for a given nominal or verbal element and to present this synthetically to the lexicographer. Furthermore, we intend to experiment with possibilities to propose collocation candidates on the assumptions (i) that most compounds in the domain are compositional and transparent and (ii) that in such cases compounds “inherit” collocational preferences from the heads of their bases: thus, as we have *Schraubenloch* (screw+hole) and *Loch für Schraube* (hole for screw) (section 4.3.1), as well as *Loch bohren* (drill a hole) and *Bohren des Lochs* (drilling of a hole), we provide *Schraubenloch bohren* and *Bohren des Schraubenlochs* as candidates, even though these are not covered by our current corpora, but may well be found in other corpora of the domain.

As of the summer of 2015, we are in the process of enhancing the tools; while experimental lexicographic work is going on to assess the usefulness of the tools, no large-scale lexicographic activity has yet been carried out.

5. Conclusion and future work

In this paper we presented tools and procedures for the extraction of term candidates from German specialized language texts, and for grouping the extracted data in a meaningful way, in order to provide raw material for the interactive construction of specialized dictionaries.

Since we intend these dictionaries to be used especially for semi-automatic document classification in the context of electronic communication between experts and lay persons or semi-experts, as well as for text production, we based our extraction procedures on both expert and user-generated text.

We consider that term variants, taxonomic relations, as well as other relations, such as purpose or material are crucial. To provide hints at such semantic relations, we use different morphological, morphosyntactic and syntactic extraction tools and relate their results. The setup is similar to that of the *Sketch Engine* (Kilgarriff et al., 2004), in so far as we extract syntagmatic data by means of pattern-based search, we are able to combine the results to make relations between the elements of German compounds explicit. We can go beyond the functions of *Sketch Engine* by exploiting nominal compounds and their syntactic paraphrases, and by interpreting e.g. noun+PP co-occurrences semantically.

The use of existing semantic lexicons, such as WordNet (Fellbaum, 1998)³, to seed the semantic classification, as well as the use of domain-specific hierarchies (e.g. provided by relevant manufacturers) is being investigated; a first inspection of WordNet data for the types of drills discussed in Table 7 showed mixed results: at an abstract level, “diamond” and “wood” are both materials, and disambiguation on WordNet data alone seems less powerful than the paraphrase-based approach discussed.

Future work will include broader coverage experimentation on the 17.9 M words corpus, the use of domain-specific taxonomic data from manufacturers, more paraphrase-based interpretation rules and quantitative evaluations of subsets of the data produced. Furthermore, the extraction procedures themselves will be fine-tuned, and experiments into low-cost domain-adaptation will be made.

6. Acknowledgements

The work reported in this paper has been carried out in the framework of the collaborative research project “Terminologieextraktion und Ontologieaufbau” financed by the corporate research department of Robert Bosch GmbH. We gratefully acknowledge this support.

³ <http://wordnetweb.princeton.edu/perl/webwn>

7. References

- Ahmad, K., Davies, A., Fulford, H. & Rogers, M. (1992). What is a term?—the semi-automatic extraction of terms from text. In *Translation Studies – An Interdiscipline*, pp. 267–278.
- Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *Processing of EACL Conference 2006*.
- Björkelund, A., Bohnet, B., Love, H. & Pierre, N. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, Beijing, China. Coling 2010 Organizing Committee, pp. 33–36.
- Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China, Association for Computational Linguistics, pp. 89–97.
- Brants, S., Dipper, S., Eisenberg, P., König, E., Lezius, W., Rohrer, C., Smith, G. & Uszkoreit, H. (2004). Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation*, 2, pp. 597–620.
- Cap, F. (2014). Morphological processing of compounds for statistical machine translation. Dissertation, Institute for Natural Language Processing (IMS), University of Stuttgart.
- Daille, B. (2007). Variations and application-oriented terminology engineering, pp. 163–177.
- Daille, B. (2012). Building bilingual terminologies from comparable corpora: The ttc termsuite. In *Proceedings, 5th Workshop on Building and Using Comparable Corpora with special topic “Language Resources for Machine Translation in Less-Resourced Languages and Domains”*, co-located with *LREC 2012*, Istanbul, Turkey.
- Faaß, G. and Eckart, K. (2013). Sdewac – a corpus of parsable sentences from the web. In *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25–27, Proceedings*, pp. 61–68.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.
- Frantzi, K. and Ananiadou, S. (1999). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal of Digital Libraries*, 6, pp. 145–179.
- Fritzinger, F. and Heid, U. (2009). Automatic grouping of morphologically related collocations. In *Proceedings of the Corpus Linguistics 2009 Conference*, Liverpool/UK.
- George, T. (2014). Comparing a commercial term extraction tool with a research prototype: an evaluation study on DIY instruction texts. Bachelor thesis, Institute for Natural Language Processing (IMS), University of Stuttgart.
- Gojun, A., Heid, U., Blancafort, H., Loginova, E., Guégan, M. & Gornostay, T. (2012a). Reference lists for the evaluation of term extraction tools. In *Proceedings of Terminology and Knowledge Engineering Conference*, pp. 651–656.
- Gojun, A., Heid, U., Weissbach, B., Loth, C. & Mingers, I. (2012b). Adapting and evaluating a generic term extraction tool. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 651–656.

- Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The sketch engine. *Information Technology*, 105, pp. 116.
- Koehn, P. and Knight, K. (2003). Feature-rich statistical translation of noun phrases. In *Proceedings of ACL 2003*.
- Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159–174.
- Schäfer, J. (2015). Statistical and parsing-based approaches to the extraction of multi-word terms from texts: implementation and comparative evaluation. Bachelor thesis, Institute for Natural Language Processing (IMS), University of Stuttgart.
- Schiehlen, M. (2003). A Cascaded Finite-State Parser for German. In *Proceedings of EACL 2003*, pp. 163–166.
- Schmid, H., Fitschen, A. & Heid, U. (2004). Smor: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 1263–1266.
- Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pp. 777–784.
- Seeker, W. and Kuhn, J. (2012). Making ellipses explicit in dependency conversion for a german treebank. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey, pp. 3132–3139.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Linked Terminologies: Applying Linked Data Principles to Terminological Resources

Philipp Cimiano¹, John P. McCrae^{1,4}, Víctor Rodríguez-Doncel², Tatiana Gornostay³, Asunción Gómez-Pérez², Benjamin Siemoneit¹, Andis Lagzdins³

¹Cognitive Interaction Technology, Excellence Cluster, Bielefeld University, Germany

²Ontology Engineering Group, Universidad Politecnica de Madrid, Spain

³Tilde, Latvia

⁴National University of Ireland, Galway, Ireland

{cimiano,jmccrae}@cit-ec.uni-bielefeld.de, bsiemone@techfak.uni-bielefeld.de,

{vrodriguez,asun}@fi.upm.es, {tatjana.gornostaja,andis.lagzdins}@tilde.lv

Abstract

In this paper we present an approach to publishing and linking terminological resources using linked data principles. We describe how terminologies can be represented in the Resource Description Framework (RDF), and as proof-of-concept we describe the application of these principles to two well-known terminologies, that is the InterActive Terminology for Europe (IATE) and the European Migration Network (EMN) glossary. We further present a simple yet effective method for inducing links between terminologies and present a small evaluation of the quality of the automatically induced links. We also present a publicly available service to transform TBX documents into RDF that we have used for the conversion of IATE to RDF.

Keywords: terminology; linked data; TBX; IATE; EMN

1. Introduction

Terminological resources (*terminologies* further in the text) play an important role in many applications where terminological consistency needs to be achieved or content needs to be described in multiple languages, for different audiences, levels of expertise, etc. So far, however, it is not trivial to discover, combine and exploit multiple terminologies within one application, nor is it easy to bootstrap the creation or extension of existing terminologies with content from other terminologies. To support such scenarios, an important step is to ensure that terminologies do not exist independently of each other, but are mutually linked to form a larger ecosystem of many (linked) terminologies comprising many domains, languages, etc.

Providing a first step towards creating such an ecosystem of linked terminologies, in this paper we propose a novel approach to publish and manage terminological datasets as linked data. Linked data represents a new paradigm for publishing data on the web relying on Semantic Web standards (RDF¹ and SPARQL²) in such a way that data is linked across

¹ <http://www.w3.org/RDF/>

² SPARQL is the query language for the RDF data model, see <http://www.w3.org/TR/rdf-sparql-query/>

datasets and sites. The main principles of Linked Data as defined by Tim Berners-Lee, the inventor of the World Wide Web, are as follows³ (Heath and Bizer, 2011):

1. Entities in the data should be named via unique URIs;
2. These URIs should be HTTP URIs and resolve using standard web protocols;
3. When these URIs are resolved, they should return useful information about the resource;
4. They should contain links to other URIs so people can discover related resources.

We apply linked data principles to terminological datasets and present an approach to transform term bases in TBX format to RDF. Our approach is based on the *lemon* model⁴ (McCrae et al., 2011), an RDF model developed to support publishing lexical resources as linked data. The proposed methodology has been implemented as an online service named TBX2RDF. We provide proof-of-concept for this transformation using the well-known InterActive Terminology for Europe (IATE) term base as well as the European Migration Network (EMN) glossary. While IATE was already available in TBX format, the EMN glossary was not, and it was directly converted from HTML into RDF format. The Linked Data version of IATE is available at <http://tbx2rdf.lider-project.eu/data/iate>, and the Linked Data version of the EMN glossary is also available online: <http://data.lider-project.eu/emn>. An implementation of the four linked data principles mentioned above can be exemplified with the URI <http://tbx2rdf.lider-project.eu/data/iate/competence+of+the+Member+States-en>, it uniquely identifies the lexical entry ‘*Competence of the Member States*’ within IATE, it is resolvable, and the returned message provides information on the resource, being additionally linked to other URIs.

We also present an automatic method to link different terminological datasets to each other. This contributes to the creation of a seamless ecosystem of terminologies that can be easily accessed and navigated and creates added value by allowing applications to access and exploit a network of linked terminologies. To show the advantages of this linking, we include the links directly into the Linked Data version of IATE as well as the EMN dataset, so that users exploring one of these can navigate to related terms of the other resource. By linking also to the Manually Annotated Subcorpus (MASC) of the American National Corpus (ANC), we also show that our approach can be extended to linking terminologies to the mentions of the terms in a corpus.

It is important to mention that we are not proposing to replace TBX by a new format. In fact, we regard our work as providing an alternative serialization of terminologies in RDF format. We assume that terminologies will be natively stored and managed using the TBX data model, but that in addition they will be exposed in RDF to support the linking of terminologies across datasets, thus supporting the creation of the above mentioned ecosystem.

³ <http://www.w3.org/DesignIssues/LinkedData.html>

⁴ <http://lemon-model.net/>

When we started this project, we were surprised to see that there was no standard and agreed-upon format for publishing terminologies as RDF. One possibility would have been to develop an RDF model that is faithful to the original TBX model, reusing essentially the data schema behind TBX. However, this would have reduced interoperability with other lexical resources published as Linked Data including bilingual dictionaries, monolingual dictionaries, wordnets, etc. To support this, we have reused existing vocabularies for representing lexical information in connection to ontologies (e.g. the lexicon model for ontologies or lemon for short) as well as vocabularies to describe provenance of data and transaction information (e.g. the PROV-O ontology).

In essence, the main advantage we see in publishing terminologies as RDF is that this supports linking across datasets. While one might argue that the links in some sense are already *'hidden'* in the data as they are induced automatically on the basis of information available in the data in our approach, these links are made explicit as a result of this, so that others can directly exploit these links instead of having to recompute them. Further, in case links are provided by a third party between for example TBX and IATE, to where would these links be added? The third party might not have the right to add these links to the original dataset, so the links themselves would then have to be published as Linked data, clearly creating an added value that was not previously there.

In addition, RDF represents a very flexible data model that supports the flexible organisation of terminologies as a (directed) graph, allowing direct representation of terminological relations (such as *broader term*, *narrower term*, etc.) as edges in the RDF model. Second, using RDF as a data model eases the manipulation and handling of terminological data as standard tasks in terminology management can be broken down to SPARQL queries, such as: i) selecting the term entries in a particular language, ii) selecting corresponding terms in two given languages, iii) selecting the subset of a term base for a given subject field, iv) finding duplicate term entries, or v) selecting all deprecated terms in a particular resource. Further, moving to a datamodel such as RDF offers additional flexibility in that copyright and licensing information can be specified at the level of each term and term entry (Cabrio et al., 2014; Rodriguez-Doncel et al., 2014), allowing to include terms with different status and provenance within one resource, thus supporting fine-grained specification of provenance and licensing information.

The paper is structured as follows: we describe our proposed model for representing terminologies in RDF in Section 2. We then discuss in Section 3 how two terminologies have been migrated into RDF based on the lemon model as proof-of-concept. Section 4 describes our methodology for linking the terminologies to each other as well as to BabelNet and MASC, and includes a small evaluation in terms of precision of the induced links. We present a publicly available service for transforming terminologies in TBX format into RDF in Section 5, concluding in Section 6.

2. Representation of terminologies in RDF

In this section, we describe how terminologies can be represented using the Resource Description Framework (RDF). For the sake of presentation, we assume that terminologies are given in the TBX format, which is an open XML format for terminologies originally specified by the now defunct Localization Industry Standards Association (LISA)⁵, and now available as an ISO standard (ISO, 2008). This does not represent any restriction as other formats can be converted to the proposed representation. This is corroborated by the fact that the European Migration Network terminology that we consider in Section 3 was not natively available in TBX, but only via HTML, which we transformed into lemon/RDF.

Our proposed representation for terminologies in RDF, fully described online⁶, relies on the *lemon* vocabulary. *Lemon* stands for the *Lexicon Model for Ontologies* (McCrae et al., 2011) and was designed to represent lexical information in combination with ontologies. *lemon* meets the needs for representing terminologies in RDF as the conceptual backbone of a terminology can be regarded as an ontology. The terms themselves can be regarded as lexical elements, and are represented in *lemon* as *lexical entries*.

In what follows, we describe the representation of terminologies in RDF in a step-by-step fashion. For the purpose of this section we will discuss the conversion to RDF using the sample terminology in TBX format in Figure 1. We start by describing how terminological concepts are represented in our RDF representation.

The term entry in lines 3–7 would be represented in RDF by a `skos:Concept`. The Simple Knowledge Organization System (SKOS) is a vocabulary for representing knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading and taxonomies in RDF. The fundamental element of a SKOS vocabulary are *concepts*, defined as ‘*units of thought, ideas, meanings, or (categories of) objects and events, which underlie many knowledge organization systems*’. As terminologies can be seen as a special case of a knowledge organization system, using SKOS concepts to represent terminological concepts seems appropriate.

This is shown by the following RDF snippet, where the the subject field of the terminological concept is specified via the property `subjectField`:

```
:IATE_84
  a skos:Concept ;
  tbx:subjectField "1011"^^xsd:string .
```

Our TBX document as shown in Figure 1 has two language sets for English and German. In the *lemon* model, a lexicon is regarded as language-specific and as comprising lexical entries

⁵ <http://www.ttt.org/oscarStandards/tbx/>

⁶ http://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF

```

1 <text>
2   <body>
3     <termEntry id="IATE-84">
4       <descripGrp>
5         <descrip type="subjectField">1011</descrip>
6       </descripGrp>
7     </termEntry>
8     <langSet xml:lang="en">
9       <tig>
10        <term>competence of the Member States</term>
11        <termNote type="termType">fullForm</termNote>
12        <descrip type="reliabilityCode">3</descrip>
13      </tig>
14    </langSet>
15    <langSet xml:lang="de">
16      <ntig>
17        <termGrp>
18          <term>Zuständigkeit der Mitgliedstaaten</term>
19          <termNote type="termType">fullForm</termNote>
20          <descrip type="reliabilityCode">3</descrip>
21          <termCompList type="lemma">
22            <termCompGrp>
23              <termComp>Zuständigkeit</termComp>
24              <termNote type="partOfSpeech">noun</termNote>
25              <termNote type="grammaticalNumber">singular</termNote>
26            </termCompGrp>
27            <termCompGrp>
28              <termComp>der</termComp>
29              <termNote type="partOfSpeech">other</termNote>
30            </termCompGrp>
31            <termCompGrp>
32              <termComp>Mitgliedstaat</termComp>
33              <termNote type="partOfSpeech">noun</termNote>
34              <termNote type="grammaticalNumber">plural</termNote>
35            </termCompGrp>
36          </termCompList>
37          <admin type="status">approved</admin>
38          <transacGrp>
39            <transac type="transactionType">origination</transac>
40            <transacNote type="responsibility">PC</transacNote>
41            <date>2014-05-08</date>
42          </transacGrp>
43        </termGrp>
44      </ntig>
45    </langSet>
46  </body>
47 </text>

```

Figure 1: An example TBX document.

for a single language. Thus, in order to represent lexical entries in different languages, one lexicon per language needs to be created. In our example, as there are terms for English and German, two lexica need to be created. These lexica contain one lexical entry each, corresponding to the terms ‘*Zuständigkeit der Mitgliedstaaten*’ and ‘*competence of the Member States*’. The English entry generated from lines 8–14 would look as follows:

```

1 <http://tbx2rdf.lider-project.eu/data/iate/en> a ontolex:Lexicon ;
2   ontolex:entry      :competence+of+the+Member+States-en ;
3   ontolex:language   "en" .
4
5 :competence+of+the+Member+States-en
6   a                   ontolex:LexicalEntry ;
7   tbx:reliabilityCode "3"^^xsd:string ;
8   tbx:termType        tbx:fullForm ;
9   ontolex:canonicalForm :competence+of+the+Member+States-en#CanonicalForm ;
10  ontolex:language     "en" ;
11  ontolex:sense         :competence+of+the+Member+States-en#Sense .
12
13 :competence+of+the+Member+States-en#CanonicalForm
14   ontolex:writtenRep  "competence of the member states"@en .
15
16 :competence+of+the+Member+States-en#Sense
17   ontolex:reference   :IATE_84 .

```

Note that the entry specifies the reliability code (i.e. 3), the type of term (i.e. *full form*), the canonical form (i.e. ‘*competence of the member states*’), and the language (i.e. *en*). Each lexical entry is assumed to have a `LexicalSense` that represents the meaning of the entry. In this case the meaning is established by `reference` to the terminological concept `:IATE_84`.

We would generate a similar entry for German, which is identified by the URI `:Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de` and is an entry in the corresponding German lexicon. Note that both entries have a `reference` to `:IATE_84` and are thus cross-lingual equivalents.

So far, we have not yet discussed how composite terms are supposed to be represented. The individual words that make up a term are represented as `constituents` of the composite term. A component is linked to its corresponding lexical entry by way of the `correspondsTo` relation. In the example below, the lexical entry `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de` is linked to an object `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList` representing its decomposition via the property `correspondsTo`. This object `Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList` is linked to its components via the property `constituent`. For each component, its part-of-speech and grammatical number (if applicable) are indicated. The decomposition of the German entry for *Zuständigkeit der Mitgliedstaaten* (lines 21–36 in the sample TBX document) is represented in RDF as indicated below:

```

1 <http://tbx2rdf.lider-project.eu/data/iate/de> a ontolex:Lexicon ;
2   ontolex:entry      :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de ;

```



```

3   ontalex:language "de" .
4
5   :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de
6     a      ontalex:LexicalEntry ;
7     tbx:reliabilityCode "3"^^tbx:reliabilityCode ;
8     tbx:termType      tbx:fullForm ;
9     ontalex:canonicalForm :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#CanonicalForm ;
10    ontalex:language    "en" ;
11    ontalex:sense       :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#Sense .
12
13   :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#CanonicalForm
14     ontalex:writtenRep "Zuständigkeit der Mitgliedstaaten"@de .
15
16   :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de#ComponentList decomp:identifies
17     :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de ;
18     decomp:constituent :component1, :component2, :component3 .
19
20
21   :component1 decomp:correspondsTo :Zust%C3%A4ndigkeit-de .
22   :component2 decomp:correspondsTo :der-de .
23   :component3 decomp:correspondsTo :Mitgliedstaaten-de .
24
25   :Zust%C3%A4ndigkeit-de
26     a      ontalex:LexicalEntry ;
27     rdfs:label "Zuständigkeit"@de ;
28     tbx:grammaticalNumber tbx:singular ;
29     tbx:partOfSpeech      tbx:noun .
30
31   :der-de
32     a      ontalex:LexicalEntry ;
33     rdfs:label "der"@en ;
34     tbx:partOfSpeech tbx:other .
35
36   :Mitgliedstaaten-de
37     a      ontalex:LexicalEntry ;
38     rdfs:label "Mitgliedstaat"@en ;
39     tbx:partOfSpeech tbx:singular ;
40     tbx:grammaticalNumber tbx:plural

```

Finally, we discuss how to represent provenance information, in particular that as expressed via transaction elements in TBX. We rely on the PROV ontology⁷ for this, as this is the W3C recommended vocabulary to ‘*represent and interchange provenance information generated in different systems and under different contexts.*’ Some provenance information is given on lines 37–42 of Figure 1 and from this we generate the following representation:

```

1   :Zust%C3%A4ndigkeit+der+Mitgliedstaaten-de
2     tbx:reliabilityCode "3"^^tbx:reliabilityCode ;
3     tbx:transaction    :Transaction .
4
5   :Transaction
6     a      prov:Activity , tbx:Transaction ;
7     tbx:transactionType "origination"@en ;
8     prov:endedAtTime    "2014-05-08"^^<http://www.w3.org/2001/XMLSchema#date> ;
9     prov:wasAssociatedWith :Agent .
10
11  :Agent
12    a      prov:Agent ;
13    rdfs:label "PC" .

```

⁷ <http://www.w3.org/TR/prov-o/>

3. Application to IATE and EMN

In this section, we describe how IATE and the European Migration Network (EMN) datasets were converted into RDF. Table 1 provides information about the size of the generated RDF resources.

Resource	Size (terms)	RDF Triples
IATE	8,081,142	74,023,248
EMN	8,855	106,283

Table 1: Size of the resources described in this paper (without links)

3.1 Converting IATE to RDF

IATE is the current EU’s inter-institutional terminology database and successor of several preexisting databases like EURODICAUTOM (Commission), TIS (Council) and EUTERPE (Parliament), among others. IATE is managed by a management group with representatives from different institutions including the European Parliament, the European Commission, the Council of the European Union, the European Court of Justice, the European Central Bank and the Translation Centre for the Bodies of the European Union. Published in 2007, IATE contains more than 8 million terms in all official 24 EU languages and it is still growing at a pace of 300 new terms added every day⁸. It covers a broad spectrum of domains: politics, law, economics, science, energy, etc. The IATE database can be queried online⁹, and the web receives about 3600 visits per hour, with 70 million queries a year.

IATE data exports are available as a single dump file for download on the IATE website¹⁰, or on the EU Open Data Portal¹¹ and, since February 2015, via the tool IATEExtract that permits choosing the languages of interest¹². This dump is provided in TBX format, described in the previous section. The TBX data fields used by IATE are very well documented¹³ and are fully compatible with the TBX specification. Data is structured in three levels: (i) abstract “concepts” which are language independent, (ii) language level with specific info for each language and (iii) term level. IATE has been integrated in different CAT tools and

⁸ According to https://tke2014.coreon.com/slides/2014_06_19_104_1150_Maslias_et_al.pdf

⁹ <http://iate.europa.eu/>

¹⁰ <http://iate.europa.eu/tbxPageDownload.do>

¹¹ <https://open-data.europa.eu/en/data/dataset/iate>

¹² Dealing with a huge files supposes a hurdle for average computer users and translators had found simpler but lengthier manners e.g. <http://multifarious.filkin.com/2014/07/13/what-a-whopper/>.

¹³ <http://iate.europa.eu/tbx/IATE%20Data%20Fields%20Explained.htm>

databases¹⁴ (Babelnet, Linguee, MateCat, MemoQ, SDL Trados Studio, DVX2/3, CafeTran), and is also accessible from a Firefox plugin¹⁵, Wordpress widget¹⁶ etc.

We converted the data dump for IATE into RDF using the TBX2RDF converter described below in section 5. Each terminological concept in IATE was transformed into a `skos:Concept`. One lexicon was generated for each of the 24 languages and each term was represented as one lexical entry in the corresponding lexicon. Decomposition and provenance information was represented as described above in Section 2.

3.2 Converting EMN to RDF

The EMN glossary describes terminology for use in the immigration and asylum domain. We implemented a crawler to download the HTML pages for the EMN and implemented an ad-hoc converter directly into *lemon*-based RDF format. It was converted into *lemon* in a manner that follows that of IATE, in that a `Lexicon` was created for each language and then for each of the available terms a `LexicalEntry` was created. The forms of the EMN datasets were preprocessed by removing elements in brackets as well as elements separated from the main term by special characters. In this way we created in total of 338 concepts with 8,855 terms in 22 European languages. Furthermore, we also included a concept definition, semantic relations, explanatory comments and references to other terms.

4. Linking Experiments

In order to link the different terminologies to each other in addition to Babelnet¹⁷, we established links between `skos:Concepts` across datasets by matching the canonical form (lemma) of the corresponding lexical entries in different languages. The number of languages for which the lexical entries for a given concept match, is regarded as an indicator of the quality of the match; that is, the more languages yield a match, the higher the quality of the induced link is expected.

In particular, EMN concepts were linked to IATE concepts by searching for string matches between corresponding EMN lexical entries and IATE lexical entries in multiple languages. In order to improve recall, we used Snowball stemming¹⁸ for the 11 supported EU languages and transformed all strings to lowercase. The search was limited to IATE concepts associated with migration (subject field 2811).

¹⁴ <http://termcoord.eu/iate/download-iate-tbx/iate-data-in-cat-tools-and-databases/> or <http://santrans.net/>

¹⁵ <http://www.maslias.eu/2013/07/iate-european-terminology-database.html?view=classic>

¹⁶ <http://termcoord.eu/resources/>

¹⁷ <http://babelnet.org/>

¹⁸ <http://snowball.tartarus.org/>

Multiple IATE concepts can match a single EMN concept. In order to decide between candidate matches, we counted the number of languages for which each match holds and used this count as a measure for match plausibility. We induced 3,028 links between EMN and IATE by considering all possible matches. Only considering the best match for each EMN concept resulted in 2,038 links (compare Table 2).

Resources	Number of links	Percentage of EMN	Precision
EMN-BabelNet	1,347	15%	69%
EMN-IATE (all matches)	3,082	35%	93%
EMN-IATE (best matches)	2,038	23%	94%

Table 2: Number of links between resources and precision of mapping.

EMN concepts were linked to BabelNet by using Babelfy (Moro et al., 2014), a named entity linking service. Invoking the Babelfy disambiguation algorithm on the written representation of the lexical entries, we extracted all the synsets with which Babelfy annotated the written representation with and considered only those annotations consisting of exactly one synset. A precision of 69% was determined by manually comparing concept definitions for a sample of 100 matches.

On the basis of the existing linking between MASC and BabelNet and the above mentioned induced links between EMN and IATE (3,028, see Table 2) as well as between EMN and BabelNet (1,347, see Table 2), by transitive closure we were able to induce 700 links between IATE and BabelNet (via EMN as pivot), 37,405 links between EMN and MASC (via BabelNet as pivot) and 7,794 between IATE and MASC (via BabelNet and EMN as pivots). The results are summarized in Table 3. To give an example, the EMN term ‘*visa*’ was linked to the matching term associated with IATE concept 3556819 and to BabelNet synset bn:00080087n, which in turn had been used to annotate 15 different tokens in MASC.

Resources	Number of links
IATE-EMN-BabelNet	700
EMN-BabelNet-MASC	37,405
IATE-EMN-BabelNet-MASC	7,794

Table 3: Number of transitive links added to resources.

We evaluated the linking precision by manually evaluating a sample of 100 generated links. Precision of the linking is defined as the number of correctly created links divided by the number of generated links. Precision was determined by manually comparing terms, definitions and sources for a sample of matches. A link was judged as correct if the concepts share

the same source or if their definitions do not contradict and there was no better matching concept. The precision of the linking is shown in Table 2. The precision of linking EMN to IATE is quite high, which is due to the fact that they are terminologies and typically only contain one sense or meaning for a certain term / lexical entry. In contrast, BabelNet contains many possible senses for each lexical entry, so that the meaning needs to be actually disambiguated automatically, which is an error-prone process. We evaluated the precision of the induced links in dependence of the number of languages for which the written representations matched. This analysis is shown in Figure 2 and Table 4. We observe that there is a clear improvement when considering links induced when the written representations for more than five languages match.

Languages Matches	Precision
1-5	82%
6-10	95%
11-15	97%
16-20	96%

Table 4: Number of EMN-IATE mappings by number of languages matching.

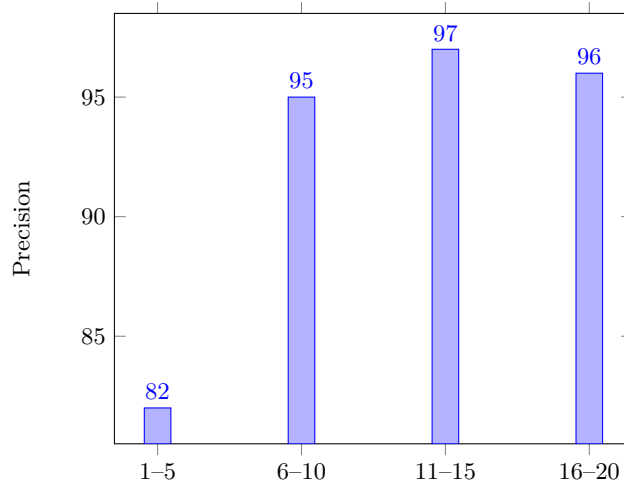


Figure 2: Precision of linking by number of languages matching for EMN-IATE mapping.

5. TBX2RDF Public Service

With the purpose of disseminating the publication of terminologies as linked data, a TBX2RDF Public Service has been released capable of converting terminologies in TBX to RDF¹⁹. The online converter consists of a form which accepts a TBX document to be uploaded or directly

¹⁹ <http://tbx2rdf.lider-project.eu/>

pasted in a box, and produces the RDF counterpart. Additional mappings can be added for specific flavours of TBX. The converter can be invoked in strict mode, in which case strict adherence to the TBX standard is ensured²⁰, and lenient mode, where some tolerance is applied. Additional information is shown when the TBX document does not conform to the standard, or when unexpected input is found. This demonstrative application has been key for gathering feedback on the quality of the conversion and the usefulness of the project itself.

In addition, the TBX2RDF Public Service is offered as a HTTP REST service²¹, supporting its integration with existing applications. The service can be tested online²² and it is accessible through its endpoint, offering the three following main functionalities:

- **Translate:** This is the basic conversion service, which admits as parameters the input TBX document, the desired namespace assigned to the new RDF resources, the option that forces the parser to have strict behaviour (optional) and an alternative set of mappings (optional). The service returns either the RDF document or an error message with a description of the problems encountered, if any.
- **ReverseTranslate:** This functionality is not yet fully implemented in the service. The goal is to admit the input RDF document as input together with a set of optional mappings and return the corresponding TBX document.
- **Enrich:** This functionality is not yet fully implemented in the service. The goal is to admit as input the URL of a terminology published as linked data and to return links to other terminologies as result.

6. Conclusion

In this paper, we have presented a new approach to publishing and linking terminologies using Linked Data principles. We have briefly described the advantages of applying linked data principles to terminologies and presented a model for representing terminologies in RDF. This model has been applied to the transformation of two terminologies, IATE and EMN, into Linked Data. We have also presented an approach to link terminologies to each other automatically. A public service for converting terminologies in TBX format to RDF has been implemented as part of this work and is freely available for anyone wanting to convert their terminologies into linked data. Future work involves developing better algorithms for linking as well as extending the current converter from TBX to RDF by a roundtrip functionality as well as by a service that can enrich existing terminologies with links to other terminologies.

In addition, following the creation (i.e., conversion) and harmonisation (i.e., linking) of open terminologies like IATE and EMN, we advance our work in a practical application of

²⁰ Conformance of the XML document to the DTD can be validated through the TBX Checker <http://www.tbxconvert.gevterm.net/>

²¹ <http://tbx2rdf.lider-project.eu/converter/doc>

²² <http://tbx2rdf.lider-project.eu/converter/tbx2rdf.html>

RDF-represented terminologies in industry/business-related scenarios. We have been experimenting with Tilde Terminology²³) already. Finally, in collaboration with the H2020-funded FREME innovation action²⁴, the next step is the application of linked data terminologies within real world business cases. The FREME project builds an open innovative commercial-grade framework of e-services for semantic and multilingual enrichment of digital content. The FREME project is developing enrichment services by building on existing mature semantic and multilingual technologies and cloud-based infrastructures previously developed by partners and used in business value adding components. The integration of the TBX2RDF service as a further component is currently planned.

7. Acknowledgements

This work is supported by the European projects LIDER (FP7 610782) and FREME (Horizon2020 644771) as well as by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R and a Juan de la Cierva grant) and the German Research Foundation (Cluster of Excellence Cognitive Interaction Technology ‘CITEC’, EXC 277, at Bielefeld University).

8. References

- Cabrio, E., Apro시오, A. P. & Villata, S. (2014). These are your rights: A natural language processing approach to automated RDF licenses generation. In *The Semantic Web: Trends and Challenges*, Springer, pp. 255–269.
- Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers.
- ISO (2008). Systems to manage terminology, knowledge and content – TermBase eXchange (TBX).
- McCrae, J., Spohr, D. & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, pp. 245–259.
- Moro, A., Cecconi, F. & Navigli, R. (2014). Multilingual word sense disambiguation and entity linking for everybody. In *Proceedings of the 13th International Conference on Semantic Web*.
- Rodriguez-Doncel, V., Villata, S. & Gomez-Perez, A. (2014). A dataset of RDF licenses. In Hoekstra, R., editor, *Proceedings of the 27th International Conference on Legal Knowledge and Information System*, pp. 187–189.

²³ <http://www.tilde.com/term>

²⁴ <http://www.freme-project.eu>

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

