

**Electronic lexicography in the 21st century:
thinking outside the paper**

Proceedings of the eLex 2013 conference

Edited by

Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek,

Margit Langemets, Maria Tuulik

<http://eki.ee/elex2013/>

17-19 October 2013

Tallinn, Estonia

**Electronic lexicography in the 21st century:
thinking outside the paper**

**Proceedings of the eLex 2013 conference, 17-19 October 2013,
Tallinn, Estonia**

Edited by Iztok Kosem, Jelena Kallas, Polona Gantar, Simon Krek, Margit Langemets,
Maria Tuulik

Published by Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia)
Eesti Keele Instituut (Tallinn, Estonia)

© Trojina, Institute for Applied Slovene Studies & Eesti Keele Instituut

Ljubljana/Tallinn, October 2013

CIP – Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

81'374:004.9(082)(0.034.2)

ELEX Conference (2013 ; Tallinn)

Electronic lexicography in the 21st century [Elektronski vir] : thinking
outside the paper : proceedings of eLex 2013 Conference, 17-19 October 2013,
Tallinn, Estonia / editors Iztok Kosem ... [et al.]. – El. zbornik. – Ljubljana :
Trojina, Institute for Applied Slovene Studies ; Tallinn : Eesti Keele Instituut,
2013

Način dostopa (URL): <http://eki.ee/elex2013/conf-proceedings/>

ISBN 978-961-93594-0-2 (Trojina, html)

1. Gl. stv. nasl. 2. Kosem, Iztok

53116002

Acknowledgements

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

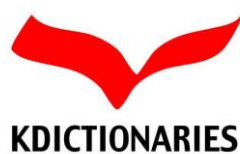
Academic partners



Main sponsors



Supporting sponsors



CONFERENCE COMMITTEES

Organising Committee

Jelena Kallas, co-chair
Iztok Kosem, co-chair
Polona Gantar
Madis Jürviste
Karmen Kosem
Simon Krek
Margit Langemets
Maria Tuulik

Scientific Committee

Andrea Abel	Iztok Kosem
Špela Arhar Holdt	Simon Krek
Lars Borin	Lothar Lemnitzer
Aljoscha Burchardt	Robert Lew
Nicoletta Calzolari	Rosamund Moon
Frantisek Čermak	Carolin Müller-Spitzer
Gilles-Maurice de Schryver	Hilary Nesi
Patrick Drouin	Vincent Ooi
Darja Fišer	Magali Paquot
Polona Gantar	Michael Rundell
Alexander Geyken	Sven Tarp
Sylviane Granger	Arvi Tavast
Gregory Grefenstette	Carole Tiberius
Patrick Hanks	Yukio Tono
Ulrich Heid	Lars Trap Jensen
Ilan Kernerman	Agnes Tutin
Adam Kilgarriff	Serge Verlinde
Annette Klosa	

TABLE OF CONTENTS

Contexts of dictionary use <i>Carolin MÜLLER-SPITZER</i>	1
Online dictionary skills <i>Robert LEW</i>	16
Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing <i>Iztok KOSEM, Polona GANTAR, Simon KREK</i>	32
A lexicographic appraisal of an automatic approach for detecting new word senses <i>Paul COOK, Jey Han LAU, Michael RUNDELL, Diana MCCARTHY, Timothy BALDWIN</i>	49
Augmenting online dictionary entries with corpus data for Search Engine Optimisation <i>Holger HVELPLUND, Adam KILGARRIFF, Vincent LANNOY, Patrick WHITE</i>	66
European Lexicography Infrastructure Components <i>Gerhard BUDIN, Karlheinz MOERTH, Matej ĎURČO</i>	76
Language Web for Frisian <i>Hindrik SIJENS, Anne DYKSTRA</i>	93
Can we determine the semantics of collocations without using semantics? <i>Pol MORENO, Gabriela FERRARO, Leo WANNER</i>	106
Online Platform for Extracting, Managing, and Utilising Multilingual Terminology <i>Marcis PINNIS, Tatiana GORNOSTAY, Raivis SKADINŠ, Andrejs VASILJEVS</i>	122
Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons <i>Núria GALA, Thomas FRANÇOIS, Cédric FAIRON</i>	132
Going Online with a German Collocations Dictionary <i>Tobias ROTH</i>	152
TERMIS: A corpus-driven approach to compiling an e-dictionary of terminology <i>Nataša LOGAR, Iztok KOSEM</i>	164
The dynamics outside the paper: user contributions to online dictionaries <i>Andrea ABEL, Christian M. MEYER</i>	179
A Jellyfish Dictionary for Arabic <i>Mohammed ATTIA, Josef VAN GENABITH</i>	195

On the Appification of Dictionaries: From a Chinese Perspective <i>Yongwei GAO</i>	213
Spiralling towards perfection: an incremental approach for mutual lexicon-tagger improvement <i>Karlheinz MOERTH, Stephan PROCHÁZKA, Omar SIAM, Thierry DECLERCK</i>	225
What should the electronic dictionary do for you – and how? <i>Oddrun GRONVIK, Christian-Emil Smith ORE</i>	243
The Woordenbank van de Nederlandse Dialecten (Wordbase of Dutch Dialects) <i>Jacques VAN KEYMEULEN, Veronique DE TIER</i>	261
Mining a parallel corpus for automatic generation of Estonian grammar exercises <i>Antoine CHALVIN, Egle EENSOO, François STUCK</i>	280
Kommunikationsverben in OWID: An Online Reference Work of German Communication Verbs with Advanced Access Structures <i>Carolin MÜLLER-SPITZER, Kristel PROOST</i>	296
Between Grammars and Dictionaries: a Swedish Constructicon <i>Emma SKÖLDBERG, Linnéa BÄCKSTRÖM, Lars BORIN, Markus FORSBERG, Benjamin LYNGFELT, Leif-Jöran OLSSON, Julia PRENTICE, Rudolf RYDSTEDT, Sofia TINGSSELL, Jonatan UPPSTRÖM</i>	310
Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Español <i>Orsolya VINCZE, Margarita ALONSO RAMOS</i>	328
Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes <i>Carla PARRA ESCARTÍN, Gyri SMORDAL LOSNEGAARD, Gunn INGER, Lyse SAMDAL, Pedro PATINO GARCÍA</i>	338
Use of support verbs in FrameNet annotations <i>Kaarlo VOIONMAA, Karin FRIBERG HEPPIN</i>	358
From DOC Files to a Modern Online Dictionary <i>Tinatin MARGALITADZE, George KERETCHASHVILI</i>	370
Online Style Guide for Slovene as a Language Resources Hub <i>Simon KREK, Helena DOBROVOLJC, Kaja DOBROVOLJC, Damjan POPIČ</i>	379
Exploring the Relationship between Language Change and Dictionary: Compilation in the Age of the Collaborative Dictionary <i>Sharon CREESE</i>	392

Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki <i>Bruno BON, Krzysztof NOWAK</i>	407
The modern electronic dictionary that always provides an answer <i>Daiga DEKSNE, Inguna SKADINA, Andrejs VASILJEVS</i>	421
Graphical representation of the web of knowledge. Analyzing the local hierarchies and the global network of connections in a specialized encyclopedia <i>Daniele BESOMI</i>	435
How preferred are preferred terms? <i>Gintare GRIGONYTE, Simon CLEMATIDE, Fabio RINALDI</i>	452
Mapping a Traditional Dialectal Dictionary with Linked Open Data <i>Eveline WANDL-VOGT, Thierry DECLERCK</i>	460
Writing assistants and automatic lexical error correction: word combinatorics <i>Leo WANNER, Serge VERLINDE, Margarita ALONSO RAMOS</i>	472
Advanced graph-based searches in an Internet dictionary portal <i>Peter MEYER</i>	488
The lexical editing system of Karp <i>Lars BORIN, Markus FORSBERG, Leif-Jöran OLSSON, Olof OLSSON, Jonatan UPPSTRÖM</i>	503

Contexts of dictionary use

Carolyn Müller-Spitzer

Institut für Deutsche Sprache (IDS), R 5, 6-13, D-68161 Mannheim

E-Mail: mueller-spitzer@ids-mannheim.de

Abstract

To design effective electronic dictionaries, reliable empirical information on how dictionaries are actually being used is of great value for lexicographers. To my knowledge, no existing empirical research addresses the context of dictionary use, or the extra-lexicographic situations in which a dictionary consultation is embedded. This is mainly due to the fact that data about these contexts is difficult to obtain. To take a first step in closing this research gap, I incorporated an open-ended question (“In which contexts or situations would you use a dictionary?”) into the online survey (N = 684) and asked the participants to answer this question by providing as much information as possible. Instead of presenting well-known facts about standardized types of usage situation, this paper will focus on the more offbeat circumstances of dictionary use and aims of users, as they are reflected in the responses. Overall, the results indicate that there is a community whose work is closely linked with dictionaries and, accordingly, they deal very routinely with this type of text. Dictionaries are also seen as a linguistic treasure trove for games or crossword puzzles, and as a standard which can be referred to as an authority. While it is important to emphasize that the results are only preliminary, they do indicate the potential of empirical research in this area.

Keywords: research into dictionary use; contexts of dictionary use; extra-lexicographic situation

1. Introduction

Dictionaries are utility tools, i.e. they are made to be used. The “user presupposition” (Wiegand et al., 2010: 680) should be the central point in every lexicographic process, and in the field of research into dictionary use, there are repeated calls for this not to be forgotten (cf. Householder, 1967; Wiegand, 1998: 259–260, 563; Bogaards, 2003: 26, 33; Tarp, 2009: 33–43). This fundamental property – serving as an appropriate tool for specific users in certain usage situations – still characterizes a good dictionary. However, the close relationship between dictionaries and their users has been weakened, at least in part.¹

“The first dictionaries ever produced may seem primitive according to the present standard, but their authors at least had the privilege of spontaneously understanding the social value of their work, i.e. the close relation between specific types of social needs and the solutions given by means of dictionaries. With the passing of the centuries and millenniums, this close relation was forgotten. [...] The social needs originally giving rise to lexicography were relegated to a secondary plane and frequently ignored.”

(Tarp, 2009: 19).

¹ The present results appear in more detailed form in Müller-Spitzer (forthcoming).

Knowledge about the needs of the user, and the situations in which the need to use a dictionary may arise, is therefore a very important issue for lexicography.

This article is structured as follows: in Section 2, the research question is introduced, and in Section 3, an analysis of the data obtained relating to contexts of dictionary use is presented, with 3.1 focusing on contexts arranged according to the categories of text production, text reception and translation, and 3.2 on users' aims and further aspects of dictionary use. Overall, the aim of this article is to give an illustrative insight into how users themselves reflect on their own use of dictionaries, particularly with regard to contexts of dictionary use.

2. Research question

To design effective electronic dictionaries, reliable empirical information on how dictionaries are actually being used is of great value for lexicographers. Research into the use of dictionaries has been focused primarily on standardized usage situations of (again) standardized user groups for which a well-functioning grid is developed, such as L1/L2/L3-speaker, text production vs. text reception or translation (cf. e.g., Atkins, 1998). In this context, Lew (2012: 16) argues that dictionaries are “most effective if they are instantly and unobtrusively available during the activities in which humans engage”. To my knowledge, no existing empirical research addresses the context of dictionary use, or, in other words, the external conditions or situations in which a dictionary consultation is embedded, also known as social situations (Tarp, 2008: 44), extra-lexicographic situations (Tarp, 2012: 114; Fuertes-Olivera, 2012: 399, 402), non-lexicographic situations (Lew, 2012: 344), “usage opportunities” (Wiegand et al., 2010: 684), in German *Benutzungsgelegenheiten* (Wiegand, 1998: 523) or contexts of use (Tono, 2001: 56).

However, it is not surprising that in this context few empirical studies exist, because these data are difficult to obtain:

“But how can theoretical lexicography find the relevant situations? In principle, it could go out and study all the hypothetical social situations in which people are involved. But that would be like trying to fill the leaking jar of the Danaids. Instead, initially lexicography needs to use a deductive procedure and focus on the needs that dictionaries have sought to satisfy until now, and on the situations in which these needs may arise.”

(Tarp, 2008: 44; cf. also Wiegand, 1998: 572).

For me, it seems to be very important to gain new empirical data relating to dictionary users in order to avoid a purely theoretical approach (cf. Simonsen, 2011, 76, who criticizes Tarp for his “intuitions and desktop research”). On the other hand, any attempt to collect real empirical data involves difficulties. With most unobtrusive methods in the context of dictionary use (i.e. particularly the analysis of log-files), it is hard to capture data about the real-life context of a dictionary consultation: firstly,

because these are personal data which in most countries cannot be collected without the explicit consent of the people; and secondly, because methods such as log-file analysis do not provide data about the circumstances of use (cf. Wiegand, 1998: 574; cf. also Verlinde & Binon, 2010: 1149; for a study that combines online questionnaires with log-file analysis see Hult, 2012). Log-file analysis mainly shows which headwords are the most frequently searched for, and which types of information are most frequently accessed. In some countries, collecting data about the URLs visited before and after the dictionary consultation is also permitted. However, what cannot be seen in log-file analysis are the contexts which lead to a dictionary consultation, e.g., for what reason text production is taking place.

However, interviews, questionnaires and laboratory studies are to a certain extent artificial situations which cannot always be generalized to everyday life (the problem of ‘external validity’). Therefore, the question arises as to whether it is a hopeless undertaking from the outset to try to collect new empirical data about contexts of dictionary use. I presume that this is not the case but that it is important to use every opportunity to obtain empirical data with all the restrictions that go with it, even if it is only possible to come closer to the goal of gaining such data step by step. The current study is a first step towards this goal (for demographic information about the participants cf. Tables 1 and 2).

In our online questionnaire study (see www.using-dictionaries.info and Müller-Spitzer et al., 2012: 429–31) we asked the participants to answer an open-ended question about the situations in which they would use a dictionary. The aim was to collect data in an exploratory way. For this, an open-ended question seemed to be the appropriate solution:

“The appeal of this type of data is that it can provide a somewhat rich description of respondent reality at a relatively low cost to the researcher. In comparison to interviews or focus groups, open-ended survey questions can offer greater anonymity to respondents and often elicit more honest responses [...]. They can also capture diversity in responses and provide alternative explanations to those that closed-ended survey questions are able to capture [...]. Open-ended questions are used in organizational research to explore, explain, and/or reconfirm existing ideas.”

(Jackson & Trochim, 2002: 307–308).

Instead of presenting well-known facts about standardized types of usage situation (text production, text reception etc.), in this paper, I will focus on the more offbeat circumstances of dictionary use, such as: from what context exactly dictionaries are used; for what reason exactly a dictionary is consulted in a text-production situation and whether there are differences between expert and non-expert users. Moreover, I am interested in the description of specific user aims (cf. Wiegand et al., 2010: 680; Wiegand, 1998: 293–298), such as: whether dictionaries are used for research; whether dictionaries are used as linguistic treasure troves for language games, and so

on. As well as these concrete questions, it is interesting to see the detail in which users are willing to describe their use of dictionaries. As the question asked was very general regarding contexts of dictionary use, it is important to emphasize that the data obtained represent a starting point for detailed research rather than an end point.

	First survey (N = 684)	
	Yes	No
Linguist	54.82%	45.18%
Translator	41.96%	58.04%
Student of linguistics	41.08%	58.92%
English/German teacher (with English/German as mother tongue)	11.55%	88.45%
EFL/DAF teacher	16.52%	83.48%
English/German learner	13.89%	86.11%

Table 1: Demographics: academic and professional background.

	First survey (N =684)
Language version of the questionnaire	English: 46.35% German: 53.65%
Sex	Female: 63.29% Male: 36.71%
Age	Younger than 21: 4.30% 21–25: 17.19% 31–30: 19.59% 31–35: 11.41% 36–45: 18.67% 36–55: 14.67% Older than 55: 14.22%
Command of English/German	Mother tongue: 64.33% Very good: 27.78% Good: 6.14% Fair: 1.46% Poor: 0.29% None: 0.00%

Table 2: Demographics: personal background.

3. Responses to the open-ended question: In which contexts or situations would you use a dictionary?

The open-ended question on contexts of dictionary use included in the online study was: “In which contexts or situations would you use a dictionary?” Participants were

asked “to answer this question by providing as much information as possible”. To gain data about real extra lexicographic situations, i.e. the contexts in which linguistic difficulties arise with no bearing on currently existing dictionaries, it would have been better to ask a question like: “In which contexts or situations do language-related problems occur in your daily life?” or “In which situations would you like to gain more knowledge of linguistic phenomena?” However, in the context of this questionnaire this would have been too general a question.

I did not expect to gain large amounts of data from the open-ended question, although the chance of obtaining more detailed and better responses to open-ended questions is higher in web surveys than in paper surveys, especially when the response field is large. This also applied to my participants: many of the nearly 700 participants (who completed the questionnaire) gave very detailed information. However, as usual, some participants dropped out of the questionnaire at the open-ended question (drop-out rate: 67 of 906 [who began the questionnaire], 7.4%). On average, the participants wrote 37 words (SD = 35.99). The minimum is unsurprisingly 0 words, the maximum 448 words. Fifty percent of participants wrote 15 to 47 words. To illustrate the range of length and level of detail of these answers, a few examples of ‘typical’ short and long answers are given in the following.

Some examples of short answers:

- “Looking up etymology.”
- “For reading articles online, for writing and translating online, for doublechecking dubious Scrabble offerings played on a gameboard in another room, etc.”
- “Consultation for work/pleasure (e.g. crossword)/to answer specific query.”

One example of a long, detailed answer:

- “To translate a word into another language. To check the meaning of a word, either in my own or in a foreign language. To find out the difference in the meanings of words in the same language, especially a foreign language I do not know very well. To find out the correct context, or the correct adpositions or cases to use with the word (for example, is it better to say “corresponds to” or “corresponds with” etc). To find out the correct spelling of a wordform – that includes finding out what that word would be in a specific case, e.g. a past form of a French verb. To find out the etymology of a word or different words. The above cases generally occur when writing a document or a letter, both for private and work purposes, be it on computer, on paper or drafting it in my mind. Usually I would use the most accessible dictionary, be it on the internet (when I am working on a computer), a paper dictionary or a portable electronic one. If no dictionary is readily available, I might write the words down and check them in a dictionary later, sometimes much later. Another time to use a dictionary is when I am reading a text I do not fully understand or am trying to find a relevant part of the text

– for example when looking for information on a Japanese web page or reading a book or article. In that case I would have a dictionary at hand, if I knew it to be a difficult text. A third case would be when I have a difference in agreement with somebody about the meaning or usage of a word or simple curiosity – for example when looking up the etymology of words to see if they have historically related meanings. Then I would use a dictionary to look it up myself or to show the entry to the other person.”

It is obvious that those participants who wrote a lot have a keen interest in the subject of the research, a fact that must be borne in mind when analyzing the results.

“[...] respondents who are more interested in the topic of an open-ended question are more likely to answer than those who are not interested. [...] Therefore, frequency counts may overrepresent the interested or disgruntled and leave a proportion of the sample with different impressions of reality underrepresented in the results.”

(Jackson & Trochim, 2002: 311).

3.1 Contexts of dictionary use relating to text production, text reception and translation

3.1.1 Data analysis

The concrete extra-lexicographic situations which lead for example to dictionary use in a text production situation are of particular interest, as pointed out in Section 2. The aim is therefore to find out more than: Do you consult a dictionary, when you are a) writing a text, b) reading a text or c) translating a text? The goal is to ascertain, for example, (a) the group ‘xy’ of users who consult a dictionary in particular when they are listening privately to foreign-language music or watching foreign-language films, or (b) users of the group ‘yz’ who consult dictionaries in particular when they are writing foreign language texts in the context of a specific subject area at work. Such insights could then lead to a more accurate picture about the situations (private/professional; written texts/spoken language/music/film, etc.) in which dictionary use is embedded.

Therefore, the first stage in the analysis was to assign the responses or parts of them to situations that relate to text production, translation or text reception. Parts of responses which were not classifiable in this way were assigned to the category “other”. The idea behind this procedure was to structure the data first in order to conduct a detailed analysis on the subsets, e.g., of what is said about the contexts in which text production takes place.

Methodologically, in the data analysis I have concentrated on one of the central techniques for analyzing data gained from open-ended questions, namely the method of structuring (cf., Dieckmann, 2010: 608–613; Mayring, 2011; for more general literature concerning the analysis of open-ended questions cf. e.g., Crabtree & Miller, 2004; Dieckmann, 2010: 531–547; Jackson & Trochim, 2002). Structuring is typically

conducted using the following steps: first, a (possibly temporary) category system is formulated; second, anchor examples are defined; and third, coding rules are established. Anchor examples are data which serve as examples for the subsequent coding process and therefore as a basis for illustrating the encoding rules. Coding rules are the rules by based on the example of this paper – a part of a response, for example, is assigned to the category of text production, while another is assigned to the category of text reception.

Here, the basic categories I assume are text production, text reception, translation and other. In the context of function theory, these are all communicative situations (cf. Tarp, 2008: 47–50; Tono, 2010: 5). Typical vocabulary, which leads to an assignment to text production, are words such as “write”, “typing”, “spell”, “correct”; for text reception, words such as “read”, “hear”, “listen to”, “watching”; and for translation, all forms of “translate” (and the corresponding German words for each, because the questionnaire was distributed in English and German). Parts of responses were assigned to the “other” category if they were either too general or they contained aspects of dictionary use other than the three basic categories. Examples are phrases such as: “When I am researching contrastive linguistics”, “solving linguistic puzzles for myself” or “during the process of designing software tools”. Therefore, the coding rules for dividing responses into the basic categories are to analyze the words used in the responses and to assign them (manually) to the four categories text production, text reception, translation and other.

In the data analyses, the corresponding parts of texts which, e.g. relate to text production are stored as extracts in a separate field. This procedure allows all parts of texts relating to text production to be analyzed separately from those which relate to translation or text reception.

3.1.2 Results of the analyses

Generally, a large number of descriptions of contexts of dictionary use can be found in the responses, which confirms what would be expected. Many participants write that they consult dictionaries constantly during their work to close lexical gaps, to ensure that they have chosen the right translation, and to check the right spelling etc. In most cases, allocating the parts of the responses to the four categories was straightforward, i.e. the extracts could be distinguished from one another relatively easily.

More than half the descriptions are related to text production situations (N = 381, 56%), followed by text reception (N = 265, 39%) and, with a very similar proportion, translation (N = 253, 38%). Forty-one percent of the responses (N = 280) are also or only assigned to the “other” category. The four categories therefore overlap, because one response may contain descriptions about text production situations and translation situations, as well as some parts which are not attributable to any of the three categories. Figure 1 shows the distribution of text production, translation and

text reception and other in the form of a Venn diagram illustrating the relationship between different types of situation.

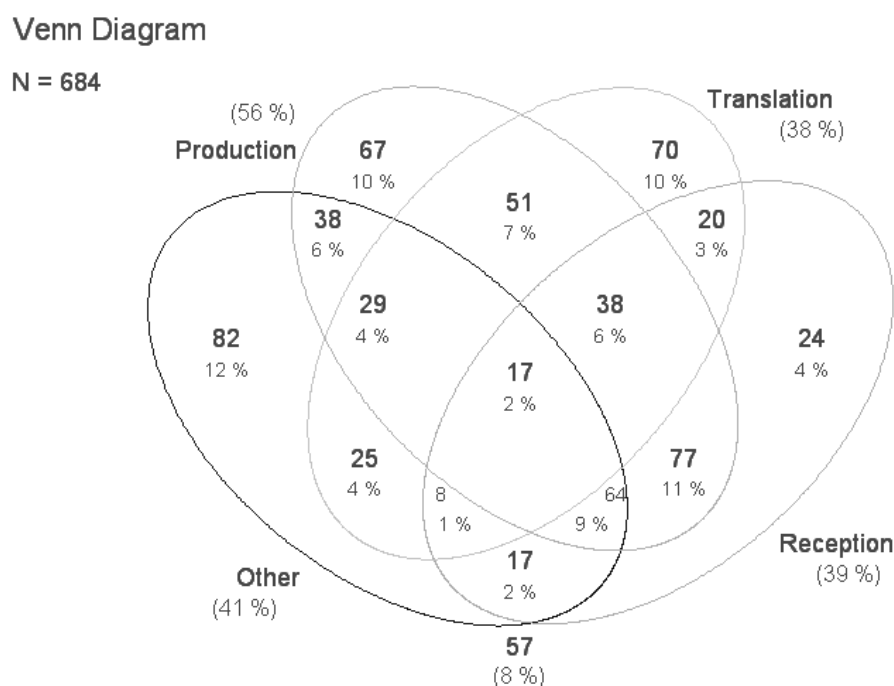


Figure 1: Venn diagram showing the distribution of text production, translation, text reception and “other”.

The diagrams show that, as already noted, dictionary consultations of situations relating to text production are described most often, followed by text reception and translation. However, 41% of the responses contain descriptions of situations which could not be assigned to any of the three categories. The level of overlap is high, i.e. many extracts are descriptions that have been assigned to more than one category. This is undoubtedly connected to the fact that some participants wrote in great detail.

Further analyses were carried out to determine whether these distributions reveal any differences between the groups, for example, that recreational users (i.e. those who use dictionaries mainly in their leisure time and predominantly for browsing) describe situations referring to text reception more frequently than experts who use dictionaries mainly for professional reasons. However, group-specific analyses revealed marginal effects in terms of the distribution of the named usage situations. It can only be stated that experts have a significantly higher value in translation ($\chi^2(7) = 61.46, p < .00$, cf. Table 2); this, however, is due to the fact that translators are part of the expert group. Therefore, this result is simply a confirmation of known facts.

The real aim of this study, however, as outlined in the introduction, is to learn more about the closer contexts of dictionary use, for example, as a result of which context texts are written and hence in which context the user need originates. The responses

contain information about this question. This will be illustrated with reference to the extracts that were assigned to text production.

For example, in many responses, indicators and clear explications are found about whether dictionary use is embedded in a personal or professional context:

- When I am writing lectures/tutorial materials at work and interested in the origin or etymology of words.
- When I am typing documents at work or sending emails internally or externally and want to check on my spelling, grammar, expression, etc.
- When I am speaking with friends online – over Facebook chat, or another messaging device – if one of my friends uses a term I am unfamiliar with, I will often “Google” it, or look it up on urbandictionary.com.

In some answers, this is also specified in more detail, i.e. some participants specifically write, e.g. “When writing Facebook entries”, “writing poetry”:

- Whenever I need to look up a word, whether [...] writing a professional document, a tweet, a Facebook message, or an email.
- Um wichtige Informationen fuer meine auslaendischen Mitbewohner zu notieren. [In order to note important information for my foreign housemates.]
- If I am writing a paper on a piece of literature that is quite old, I will look up words from that literature to make sure that my understanding of the word is the same as how the word was used at the time the literature was written.

These answers contain interesting information about the contexts of dictionary use and usage opportunities. Users’ aims are also made explicit, for example that dictionaries are used to act as someone with a high level of language skills:

- When I want to know how to pronounce something, audio pronunciation is offered by the Merriam-Webster online dictionary, especially when I want to say the word in public or in a class presentation when it is important to show that I can speak clearly and have command over the language I use.

In addition, there are descriptions of whether the work is already taking place on the computer or in another context, with the word being looked up in the online dictionary later:

- When I’m writing a paper or story, generally on my computer, and I want to check the denotation of a word that doesn’t quite seem right.
- If no dictionary is readily available, I might write the words down and check them in a dictionary later, sometimes much later.

However, sometimes important information is missing. See for example the following response:

- And if I'm talking with someone and I can't remember the right word.

Here one might wonder: When and on what sort of device does the dictionary consultation take place afterwards? Directly on a smartphone? What is then looked up exactly? Therefore, many questions remain unanswered. Beyond that, the descriptions cannot really be classified into broad categories, i.e. a clearly structured summary is not achievable. Therefore, what is difficult to evaluate from the data are the particular circumstances of contexts which lead to, e.g., a user's need for text production and therefore to a dictionary consultation. On the one hand, the question was very general, so that the responses are sometimes very general, too. On the other hand, some responses contain interesting information on the context of dictionary use, but this information cannot easily be placed in an overview. In this respect, the data, as was pointed out at the beginning, represent a starting point for further study in this field. To achieve the goal of gaining some degree of quantitatively analyzable information about contexts of dictionary use, it would therefore be advisable to use a combination of standardized and open-ended questions. Hopefully, the results of this analysis will help this eventual aim to be successfully achieved.

3.2 User aims and further aspects of dictionary use

As well as the assignment of responses to different kinds of extra-lexicographic situations, some aspects of dictionary use were often repeated in the responses and thus emerged as a category in the analysis, particularly with regard to user aims. User aim means (within the meaning of Wiegand et al., 2010: 680) the action goal which enables the user to retrieve relevant lexicographic information based on appropriate lexicographic data. Many responses contain notes on that topic, for example: "I use dictionaries for research" or "to improve my vocabulary". The analysis of these descriptions seemed to offer an interesting additional view on the data far from the basic categories of text production, text reception or translation. The emphasis is not, however, on the completeness of all named aspects, but more on the interesting and perhaps unusual categories that would not necessarily be expected.

3.2.1 Data analysis

The following categories were developed gradually during the first analysis regarding the distribution explained in 3.1. The nine categories which are relevant for this section are:

- Dictionaries used to improve vocabulary (generally, not referring to concrete text production or reception problems) (Cat. 1)
- Dictionaries used as a starting point or resource for (further) research (Cat. 2)

- Dictionaries used as mediator medium (Cat. 3)
- Dictionaries used as a resource for language games, linguistic treasure trove, for enjoyment, for personal interest, etc. (Cat. 4)

Once these categories were formed, the responses that are assigned to the appropriate category were marked.

3.2.2 Results of the analyses

Participants sometimes referred to the fact that dictionaries are used to improve and increase vocabulary independently of concrete text reception or text production problems (category 1, although explicitly only in 1% of the responses, N = 8):

- Basically, I use the dictionary in order to improve my vocabulary.

Experts in particular use dictionaries as a starting point for research (category 2). In 68 responses (10%), this aspect is explicitly mentioned. Here, there are group differences, as would be expected, especially between linguists and non-linguists ($\chi^2(1) = 23.1030, p < .00$).

Table 3 shows that 82% of those who use dictionaries as a resource for research are linguists or have a linguistic background, i.e. particular linguists are able to use dictionaries as a resource for linguistic material.

Linguist	Dictionaries used for research		Total
	no	yes	
Yes	319 52%	56 82%	375 55%
No	297 48%	12 18%	309 45%
Total	616 100%	68 100%	684 100%

Table 3: Linguist vs. non-linguist dictionary users as a resource for research

A special aspect of some responses is that dictionaries are apparently also sometimes used for linguistic discussions as mediator medium (category 3, 2%, N = 12). They are even explicitly designated as “Schlichtermedium” (conciliator medium):

- Most often, to settle questions and debates with my colleagues and/or friends about accepted pronunciations of words and word origins.
- Sometimes my friends and I dispute the usage of a word – one of us will have used it “wrong” by the other’s definition. In this case, we will turn to a dictionary for an answer.

- To settle an argument on etymology or definition when discussing words with colleagues.

Although the proportion of these responses is not high, the few examples show clearly that a very strong authority is attributed here to dictionaries. It can be assumed that such users appreciate sound lexicographic work. The user experience which is reflected here is that dictionaries provide such reliable and accurate information that they are regarded as a binding reference, even among professional colleagues.

Similarly, dictionaries also seem to be used in connection with language games such as crossword puzzles or when playing Scrabble, and also just for enjoyment or fun (category 4). In 6% (N = 39) of the responses, this aspect arises:

- For scrabble When I am bored and me and my friend have a spelling bee
- At other times I might consult the OED for information about etymology or historical use purely for personal interest or resolve a debate about word usage.
- Sometimes to see if a neologism has made it into the hallowed pgs of the OED!
- Solving linguistic puzzles for myself (having to do with usage, grammar, syntax, etymology, etc.)

4. Conclusion

It is demanding to obtain empirical data about contexts of dictionary use. In this study, I made an attempt in this direction. The willingness of the participants to give detailed information was significantly higher than expected. This is probably partly due to the fact that most of the participants have a keen interest in dictionaries. One conclusion that can be drawn from this for further research, is that this community is apparently prepared to provide information about the contexts of potential acts of dictionary use, and that this should also be used.

All in all the results show that there is a community whose work is closely linked to dictionaries and, accordingly, they deal very routinely with this type of text, and sometimes describe these usage acts in great detail. Dictionaries are also seen as a linguistic treasure trove for games or crossword puzzles and as a standard which can be referred to as an authority. What is difficult to evaluate from the data are the particular contexts of dictionary use which lead to, e.g., the user's need for text production and therefore to a dictionary consultation. Although data on this could be obtained, it is still not possible to draw a clear picture. That responses on open-ended questions are sometimes very general (like it was in the current case) is a problem which holds for answers on these kinds of questions in general:

“They can provide detailed responses in respondents’ own words, which may be a rich source of data. They avoid tipping off respondents as to what response is normative, so they may obtain more complete reports of socially undesirable behaviors. On the other hand, responses to open questions are often too vague or general to meet question objectives. Closed questions are easier to code and analyze and compare across surveys.”

(Martin, 2006: 6).

On the other hand, some responses contain interesting information on the context of dictionary use, but a synopsis of the many details in an overall image is almost impossible to achieve. In this respect, it is important to emphasize that the present results are only preliminary; but they do indicate the potential of empirical research in this area.

This will certainly be a worthwhile path to take, as knowledge about the contexts of dictionary use touches on an existential interest of lexicographers. Dictionaries are made to be used and this use is embedded in an extra-lexicographic situation. And the more that is known about these contexts, the better dictionaries can be tailored to users’ needs and made more user-friendly. Particularly when innovative dictionary projects with new kinds of interfaces are to be developed, better empirical knowledge is essential, e.g. the following quotes about the „Base lexicale du français“ show (cf. also Verlinde, 2010; Verlinde & Peeters, 2012):

“The BLF’s access structures are truly task and problem oriented and based on the idea that the dictionary user has various extra-lexicographic needs, which can lead to a limited number of occasional or more systematic consultation or usage situations. [...] We argue that the dictionary interface should reflect these consultation contexts, rather than reducing access to a small text box where the user may enter a word.”

(Verlinde, Leroyer & Binon, 2010: 8)

“The Belgian BLF project seeks a different solution to the same underlying challenge: here the users have to choose between situations before they are allowed to perform a look-up. This approach looks promising but it also draws attention to a potential catch-22 situation: on the one hand, requiring too many options and clicks of users before they can get started may scare them away. And on the other hand, a model with immediate look-up and only few options may lead to inaccurate access and lack of clarity. Whatever the situation, we need more information about user behaviour to assess which solution works more effectively.”

(Trap-Jensen, 2010: 1139)

This is particularly important at a time when people have an increasing amount of freely available language data at their disposal via the Internet. Dictionaries can only retain their high value when distinct advantages (e.g. in terms of accuracy and reliability, as well as exactly meeting users’ specific needs in concrete contexts) are provided, compared to using unstructured data for research.

My results indicate that, although these are currently difficult economic times for dictionary publishers, the participants in this study actually appreciate many of the classic characteristics of dictionaries.

5. References

- Atkins, S.B.T. (1998): *Using dictionaries. Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Niemeyer.
- Bogaards, P. (2003). Uses and users of dictionaries. In P. van Sterkenburg (ed.) *A Practical Guide to Lexikography*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 26–33.
- Crabtree, B., Miller, W. L. (eds.) (2004): *Doing Qualitative Research*, 2nd edition, London: Sage.
- Dieckmann, A. (2010): *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*, 4th edition, Hamburg: Rowohlt.
- Fuertes-Olivera, P.A. (2012). On the usability of free Internet dictionaries for teaching and learning Business English In S. Granger, M. Paquot (eds) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 399–424.
- Householder, F.W. (1967). *Problems in Lexicography*. Bloomington: Indiana University Press.
- Hult, A.K. (2012) Old and New User Study Methods Combined – Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. In: J.M. Torjusén, Vatvedt Fjeld, R. (eds.). *Proceedings of the 15th EURALEX International Congress 2012, Oslo, Norway*, pp. 922–928. Accessed at: http://www.euralex.org/elx_proceedings/Euralex2012/pp922-928%20Hult.pdf.
- Jackson, K.M., Trochim, W.M.K. (2002): Concept Mapping as an Alternative approach for the analysis of Open-Ended Survey Responses. *Organizational Research Methods*, 5, pp. 307-336. Accessed at: <http://www.socialresearchmethods.net/research/Concept%20Mapping%20as%20an%20Alternative%20Approach%20for%20the%20Analysis%20of%20Open-Ended%20Survey%20Responses.pdf>.
- 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.
- Martin, E. (2006). Survey Questionnaire Construction. *Survey Methodology. Research Report Series*, 21, pp. 1–14. Accessed at: <http://www.census.gov/srd/papers/pdf/rsm2006-13.pdf>.
- Mayring, P. (2011). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Beltz.

- 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.
- Müller-Spitzer, C. (ed.) (forthcoming): *Using online dictionaries* (Lexicographica. Series maior).
- Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography*. Berlin/New York: de Gruyter.
- Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen & S. Tarp *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, Frankfurt a.M./Berlin/Bern/Bruxelles/NewYork/Oxford/Wien: Peter Lang. pp. 17–32.
- 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.
- Tono, Y. (2001). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Tübingen: Max Niemeyer.
- Tono, Y. (2010). A Critical Review of the Theory of Lexicographical Functions. *Lexicon* 40, pp. 1–26.
- Trap-Jensen, L. (2010). One, Two, Many: Customization and User Profiles in Internet Dictionaries. In: A. Dykstra, T. Schoonheim (eds.) *Proceedings of the 14th EURALEX International Congress 2010, Leeuwarden/Ljouwert*, pp. 1133–1143. Accessed at: http://www.euralex.org/elx_proceedings/Euralex2010/105_Euralex_2010_7_TRAP-JENSEN_One,%20Two,%20Many_Customization%20and%20User%20Profiles%20in%20Internet%20Dictionaries.pdf.
- Verlinde, S., Binon, J. (2010) Monitoring Dictionary Use in the Electronic Age. In A. Dykstra, T. Schoonheim. *Proceedings of the 14th EURALEX International Congress 2010, Leeuwarden/Ljouwert*, pp. 1144–1151.
- Verlinde, S., Peeters, G. (2012). Data access revisited: The Interactive Language Toolbox In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. pp. 147–162.
- Wiegand, H.E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. Berlin/New York: de Gruyter.
- Wiegand, H.E., Beißwenger, M., Gouws, R.H., Kammerer, M., Storrer, A. & Wolski, W. (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung: mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neuen Sprachen*. Berlin/New York: de Gruyter.

Online dictionary skills

Robert Lew

Adam Mickiewicz University in Poznań
email: rlew@amu.edu.pl

Abstract

Successful dictionary use requires two ingredients: (1) high-quality, user-friendly dictionaries and (2) dictionary users who know what they are doing. The bulk of the current research effort within lexicography concentrates on making better dictionaries, with new opportunities afforded by the electronic medium. In contrast, the other ingredient—educating the user—receives comparatively little attention. The present contribution looks at dictionary reference skills in an effort to determine how traditional print dictionary skills need to evolve in order to allow users to get the most out of electronic dictionaries, in particular those offered online. The most comprehensive overview of dictionary skills to date has been conducted by Nesi (1999). Nesi's list is systematically reviewed, considering the relevance of each item in the context of online dictionaries. Novel ways of accessing lexicographic data are the most prominent quality of electronic dictionaries. The skills involved are sought by examining a selection of relevant literature on search techniques in electronic dictionaries, as well as some work done in the area of web search skills.

Keywords: electronic dictionary, online dictionary, dictionary skills, reference skills, internet skills, digital literacy, information literacy

1. Introduction

Human dictionary use involves two parties: the dictionary and the user. Therefore, successful lexicographic consultation is a two-way affair, and depends on two ingredients: how easy to use the dictionary is, and what skills related to dictionary use the user possesses.

When we review the published literature in lexicography, much of the research effort, and the bulk of the writing, focuses on ways to produce better dictionaries. With the electronic revolution upon us, we are now actively searching for new standards of quality for electronic dictionaries. But educating the user remains an equally valid concern, as convincingly shown in the context of online dictionaries by Ranalli (2013). While the literature on training in dictionary skills is not overwhelming, there are already some valid and practically useful findings (e.g. Kennedy, 1972; Herbst & Stein, 1987; Chi, 1998; Nesi, 1999; Bishop, 2000, 2001; Campoy Cubillo, 2002; Carduner, 2003; Osuchowska, 2003; Lew & Galas, 2008; Van der Merwe, 2012). However, with only isolated exceptions (Nesi, 1999; Ronald & Ozawa, 2011), authors address skills relevant in using traditional print dictionaries, with very little being said specifically about skills needed in the context of electronic dictionary use.

The present contribution shifts the focus to electronic dictionary skills, with particular emphasis on online dictionaries. I review the most comprehensive specification of dictionary skills available to date (Nesi, 1999) and consider to what

extent, and in what ways, the individual skills listed apply to dictionaries in the electronic format. It is clear that, from the user perspective, a major area of difference between electronic and print dictionaries is in how information is accessed. Accordingly, I examine two important metalexigraphic contributions treating this topic (Engelberg & Lemnitzer, 2009; Pastor & Alcina, 2010), in an attempt to identify or infer the skills which appear to be implicated in the use of the search techniques. Finally, I also look at digital skills involved in web search strategies in an effort to identify further skills which might be relevant in interacting with online dictionaries.

2. The set of dictionary skills

Hartmann (1999) claimed that the set of skills required of a dictionary user had not yet been established empirically. This is still true today, and existing specifications of dictionary skills are based largely on introspection, mostly by trying to reflect on what goes on in a dictionary consultation act. The most comprehensive listing of dictionary skills to date is one by Nesi (1999), produced with university students (in the UK) in mind. The list was intended to be relevant for both print and electronic dictionaries, though, understandably, at that point in time the coverage of issues specific to electronic dictionaries, and in particular online dictionaries, could not have been very broad by today's standards. Below, I review Nesi's list, commenting on the relevance of individual skills to modern electronic dictionaries.

Nesi organizes her set of skills into stages, which represent major hypothetical steps involved in dictionary consultation in the context of university studies. There are five such stages plus a sixth cluster of metalexigraphic skills gathered under the rubric 'Understanding lexicographical issues'. Below, I give an overview of the skills, either individually or grouped as appropriate, and comment on their relevance for digital dictionaries. For reasons of space, I will omit two of the original groups of skills of least relevance here: metalexigraphic skills focusing on understanding lexicographical issues; and the stage of recording entry information which concerns noting down the information found as a mnemonic technique and for future reference. Thus, the four stages covered below are as follows:

1. Before study (i.e. having to do with selecting a dictionary to be used in the educational studies)
2. Before dictionary consultation
3. Locating entry information
4. Interpreting entry information

Let us now go through these four stages, focusing on potential digital dictionary skills. As there are quite a few skills involved, and many are interrelated, each skill listed in this section below will receive a numbered heading to facilitate easier identification and cross-referencing. Where applicable, skills will be clustered.

2.1 Stage one: Before study

Skill 1: Knowing what types of dictionary exist, and choosing which dictionary/ies to consult and/or buy

Skill 2: Knowing what kinds of information are found in dictionaries and other types of reference works

Awareness of the range and types of dictionaries, and (more generally) reference works, used to be fairly stable knowledge in the print era. In contrast, today's reference works are evolving at such a rate that they are a real challenge to keep up with. It is increasingly hard to stay on top of what the best reference works are. Fortunately, this need not be the actual challenge: with so many alternatives available, it may be sufficient to settle for the *good enough* tool. If we believe that it is the educational system that should be responsible for teaching students (at all stages) about dictionaries, then this is made difficult by the fact that in many countries teachers tend to be left behind in the digital revolution: they find it hard to keep up with new technology, and in this they tend to fare even worse than their students (Langegard, 2011). Thus, dictionary users are pretty much left to their own devices.

2.2 Stage two: Before dictionary consultation

Skill 3: Deciding whether dictionary consultation is necessary

This decision is largely about solving the equation between the *cost* of consultation (including inconvenience, distraction, and time), and its potential *benefits*. While the trade-off persists, the parameters in the equation have shifted: consulting a digital dictionary may be less of a distraction if it is well integrated into the context of reading, writing, translating, or whatever activity the user is engaged in. On the other side, a digital resource may offer greater benefits than a printed resource. All in all, the decision to consult a dictionary is easy to make, and so is the consultation itself: studies often find digital dictionaries to be used more than their paper predecessors.

Skill 4: Deciding what to look up

Skill 5: Deciding on the appropriate form of the look-up item

In print dictionaries, important components of this pre-lookup phase are: identifying the locus of difficulty (e.g. in the text being read), deciding between a single word and a multi-word item, and then coming up with a citation form likely to have headword status in the dictionary. The more sophisticated electronic dictionaries can relieve the user of having to worry about some of the above: inflected-form search (3.1.8 below), and to some extent incremental search (3.1.1), should assist in locating the relevant entry, and multi-word expressions may be easier to find (Lew, 2012b).

Skill 6: Deciding which dictionary is most likely to satisfy the purpose of the consultation

On the one hand, the wealth of dictionaries available online (at least for English) may leave users spoilt for choice. Many online dictionaries push poor and/or out-of-date

content, but users may not be in a position to notice; instead they tend to be (mis)guided by outward appearances, unable to separate the wheat from the chaff. On the positive side, some electronic dictionaries can reshape themselves to better serve a range of different needs.

Skill 7: Contextual guessing of the meaning of the look-up item

This skill mostly applies in receptive dictionary use (reading), and is equally relevant to both print and electronic dictionaries, at least until e-dictionaries can genuinely assist in contextual sense disambiguation.

Skill 8: Identifying the word class of the look-up item

This skill is meant to facilitate the look-up by restricting it to a specific syntactic class (noun, adjective). It is relevant for those dictionaries which use part of speech as an important criterion in structuring the lexicographic data. An electronic dictionary with access to the text being read could relieve the user of having to identify the part of speech. If the word form appearing in the text is a unique inflectional form (e.g. *needed*), then this is rather trivial. Otherwise (e.g. *needs*), some parsing and tagging is required to identify the part of speech positively.

2.3 Stage three: Locating entry information

Skill 9: Understanding the structure of the dictionary

Like print dictionaries, electronic dictionaries are structured entities. However, the electronic medium accommodates a greater variety of types of structures, and this can present a serious challenge to users – even those experienced in using paper dictionaries. The broad diversity of types of electronic dictionaries is a sign of technological divergence, and can be contrasted with the structural convergence of paper dictionaries, which, over the centuries, have developed a fairly uniform set of conventions.

Skill 10: Understanding alphabetization and letter distribution

The role of alphabetical ordering is quite significantly reduced when consulting electronic dictionaries, as these dictionaries allow users to be ‘liberated from the straitjacket of ... alphabetical order’ (Atkins, 1996: 516). It is only in some superficially retro-digitized versions of paper dictionaries (cf. Lew, 2011) that alphabetical sequencing, so crucial in navigating most print dictionaries, matters. Similarly, letter distribution – that is, the relative amount of space that specific letter sections occupy – is rarely an issue.

Skill 11: Understanding grapho-phonemic correspondence (and the lack of it)

Few electronic dictionaries today offer explicit phonemic look-up options, but speech recognition seems to be the way forward, once it can overcome the difficulties involved in dealing with foreign accents and individual idiosyncrasies. Not infrequently, learners of English approximate the pronunciation of a word by making

an attempt at respelling it, and the better ‘did you mean’ systems can often guess at the word actually intended (Lew & Mitton, 2011, 2013).

Skill 12: Understanding the use of wildcards in electronic dictionary searches

This is a digital-only skill, and is covered under 3.1.2 below.

Skill 13: Choosing amongst homonyms

The contrast between homonymy and polysemy is not necessarily as relevant for modern dictionaries. The current tendency, largely inspired by learner lexicography, is to group senses by part of speech rather than historical relatedness. In any case, this skill appears to be a subset of a more general skill: being able to locate the relevant sense in the dictionary. This issue has attracted some attention in the context of electronic dictionaries: Lew & Tokarek (2010) found that active menus can help improve success (and speed) and make sense selection less of a challenge to the not-so-skilled dictionary user.

Skill 14: Finding derived forms

An electronic dictionary can assist users significantly in the task of locating derived items by providing explicit links between the related forms, or else by being able to compute derived forms in real time when equipped with ‘morphological awareness’. This is of particular importance to non-native language dictionary users, whose command of the derivational morphology of the language may be far from complete, though also a potential source of difficulty for the less skilled native writer.

Skill 15: Finding multi-word units

Being able to locate multi-word units is, according to Nesi (1999), a much-neglected skill. As noted by Lew (2012b), access to this notoriously troublesome type of item can be significantly enhanced by including full treatment (or hyperlinks to full treatment) under all relevant component lemmata, making the user’s failure to guess the keyword of an expression less critical. And, this skill becomes irrelevant in a dictionary capable of recognizing multi-word units (assuming it ‘sees’ the text being read or translated) and extracting specific information from its database.

Skill 16: Understanding the cross-referencing system in print dictionaries, and hyperlinking in electronic dictionaries

Dictionary users’ ability to take advantage of hypertext features of dictionaries is likely to improve with the growing role of the Web in today’s life and work. The skill implies awareness of which elements are linked, and what the hyperlinks point to. Principles of user-centred design should ensure that hyperlinks are made evident to the users, but the actual decision of whether to follow a hyperlink needs to be grounded in an awareness of dictionary content and structure.

2.4 Stage four: Interpreting entry information

Skill 17: Distinguishing the component parts of the entry

In the context of electronic lexicography, awareness of the microstructural make-up of a dictionary becomes a more complex skill, depending on how the different types of lexicographic data are organized and presented in a particular e-dictionary. In principle, the data presented need not include everything held in the database. Some entry components, such as phonemic transcription or (additional) examples may well be hidden from initial view. This potential ‘latency’ of lexicographic data makes it harder for the user to recognize the potential components of the entry at first sight.

Skill 18: Distinguishing relevant from irrelevant information

Recognizing the relevance of information to the task at hand is a general cognitive skill, and it is dependent on a sound understanding of one’s information needs in a particular context. These needs have to be matched against dictionary content, so users need an awareness of the types of information that a dictionary is able to offer them.

Skill 19: Finding information about the spelling of words

Modern electronic dictionaries have revolutionized ways of accessing spelling information. First, hypothetical spelling forms can be typed into the search box, and so the ordering of headwords (crucial to paper dictionaries) becomes almost an irrelevancy, if it exists at all. Second, a suggest-as-you-type facility can supply the missing portion of a word as long as a few initial characters are entered correctly (Lew, 2012a). Third, reasonable misspellings stand a chance of being corrected by the ‘did you mean’ function (Lew & Mitton, 2011, 2013). Checking spelling in an e-dictionary is thus generally easier and less of a specialized skill. On the other hand, the need for isolated consultations for word spelling is largely obviated by the spellchecking functions increasingly available in applications such as word processing software or email clients.

Skill 20: Understanding typographical conventions and the use of symbols, numbered superscripts, punctuation

Those typographical conventions that are primarily motivated by constraints of space can be discarded in electronic dictionaries, though only up to a point, as constraints on presentation space continue to apply in electronic dictionaries (Lew, in press). Still, some of the traditionally cryptic shorthand symbols may be spelled out, while for others dictionaries can supply pop-up explanations.

Skill 21: Interpreting IPA and pronunciation information

Electronic dictionaries can (and an increasing number do) supply pronunciation information by presenting spoken audio representations of items, a technological impossibility in print dictionaries. These work well for native speakers of the language; however, a language learner may not recognize the phonemic make-up of

an item from just hearing it, as perception of speech sounds depends on the phonological system of one's native language. Thus, for language learners, the ease of audio representations is deceptive. Transcription still has a place in electronic dictionaries, as it provides an explicit and unambiguous phonemic representation (and possibly a degree of phonetic detail). Of course, interpreting transcription is a fairly technical skill and is not something a *casual* user would be expected to be able to master.

Skill 22: Interpreting etymological information

Skill 23: Interpreting morphological and syntactic information

Skill 24: Interpreting the definition or translation

Skill 25: Interpreting information about collocations

Skill 26: Interpreting information about idiomatic and figurative use

Skill 27: Deriving information from examples

Skill 28: Interpreting restrictive labels

I have grouped the above skills, as they all fall under the more general umbrella skill of deriving specific linguistic and metalinguistic information from lexicographic data. These skills are less dependent on the print-versus-electronic opposition, and have more to do with ways of representing particular information. Therefore, the above skills have similar relevance in e-dictionaries, except when the electronic medium can offer more user-friendly presentation than that inherited from paper dictionaries (such as, say, a more satisfying presentation of examples). I will not discuss these detailed options here for reasons of space.

Skill 29: Referring to additional dictionary information (in front matter, appendices, hypertext links)

In general, the electronic medium offers a potential for better integration of what used to be separate major textual components of paper dictionaries. This is achieved through embedding, integrating and hyperlinking. By the same token, users should find it easier to navigate between the different sections of lexicographic data.

Skill 30: Verifying and applying look-up information

Once the information has been extracted from an entry, it needs to be applied in a comprehension, production, or translation task which prompted the look-up. This is a sophisticated skill and, again, it will not be made appreciably easier by going digital, except when the dictionary forms part of a more elaborate lexical tool such as an intelligent writing assistant.

As mentioned above, Nesi's (1999) final stage concerns the recording of entry information as a memory aid or for future reference. This will not be developed here. Instead, I will approach the issue from a different angle, focusing on what is most distinct about digital dictionaries: access to data.

3. Search techniques in online dictionaries

Access to lexicographic data is a fundamental aspect in which electronic dictionaries differ from their paper predecessors. Based on a comprehensive corpus of metalexicographic texts, De Schryver (2012: Figure 33) notes the steady replacement of 'looking up' with 'searching'. This he attributes to the growing role of electronic dictionaries.

3.1 Overview of search techniques

Pastor & Alcina (2010: 308) emphasize the relevance of search techniques to the teaching of electronic dictionary skills. They observe that:

...we have found no studies that establish a 'universal' classification or arrangement of the search techniques that can be used in a dictionary, in other words, one that is valid for training in electronic dictionary use in general, and that can be adapted to any specific dictionary.

A detailed overview of possible search techniques in electronic dictionaries is provided by Engelberg & Lemnitzer (2009: 101-102). These authors list the following options:

1. Incremental search (Inkrementelle Suche)
2. Wildcard search (Suche mit Platzhaltersymbolen)
3. Boolean search (Suche mit logischen Konnektoren)
4. Filtered search (Filterbasierte Suche)
5. Sound search (Lautformbasierte Suche)
6. Fuzzy-spelling search (Schreibungstolerante Suche)
7. Inflected form search (Flexionsformbasierte Suche)
8. Index-based search (Indexbasierte Suche)
9. External-text-based search (Textbasierte wörterbuchexterne Suche)
10. Picture-based search (Bildbasierte onomasiologische Suche)
11. Scanner-based search (Scannerbasierte Suche)

Of the above techniques, only numbers 8 and 10 apply to print dictionaries: the rest are exclusively digital.

Skilful users of online dictionaries should be able to utilize the above search techniques, and decide beforehand which of the approaches will be optimal for a specific information need. Obviously, few (if any) dictionaries will offer a complete set of the above options, so users need to be aware what the actual choices are for a given dictionary.

Below I attempt a provisional specification of skills associated with the search techniques identified by Engelberg and Lemnitzer (2009), supplemented with Pastor & Alcina's (2010) proposal.

3.1.1 Incremental search

Recently, this search technique (or, perhaps, more precisely, term-entry technique) has become quite popular in various user interfaces (e.g. Wikipedia), although it could already be found in some early electronic dictionaries. The feature involves automated term completion from an index of available terms, before the complete term is typed. This search enhancement has been variously referred to as '*type-ahead search, search-as-you-type, incremental search, inline search, or instant search*' (Lew, 2012a: 351). Autocompletion may kick in after a predetermined number of characters have been keyed (usually a reasonably low number such as two, three, four, or five). Users interacting with this feature need to anticipate that a drop-down list of options will suddenly appear, and they will need to know that they can keep on typing (usually a sensible strategy) to further narrow down the list of target terms. Some of the better-designed dictionaries (notably *Macmillan English Dictionary Online*) also include among the incremental suggestions multi-word expressions, a particularly problematic set of lexical items to locate.

3.1.2 Wildcard search

Wildcard search involves the use of wildcard and truncation symbols, most usually the question mark '?' to replace a single character and the asterisk '*' or a plus sign '+' to replace a sequence of characters. These are not the only options, however; for example, the Polish word-game dictionary <http://www.krzyzowki.info> requires the dot '.' as the single-character replacement, and the percent symbol '%' as the multiple-character truncation symbol. Skilful use of wildcards includes an optimal choice as to how many characters to specify, and how many to replace with a wildcard. This type of decision is informed through an awareness of the lexicostatistical nature of the vocabulary of the language, which allows the user to make a rough estimate of the number of items beginning with a specific sequence of letters. Such searches are often helpful in using dictionaries to solve word games (crossword puzzles and the like). For a specific dictionary, users need to know if a wildcard search is at all possible, what the wildcard characters are, and at which positions the wildcard characters are allowed: string-initial, string-internal, or string-final.

3.1.3 Boolean search

A Boolean search combines terms with the use of logical operators of conjunction (AND), disjunction (OR) and negation (NOT), possibly grouping expressions with the use of parentheses. Support for Boolean operators in search interfaces was once a popular option in web search engines, and some early electronic dictionaries (such as the PC-based *Collins English Dictionary*) included it as well. However, continued research on human-computer interaction has found that a large majority of computer users are unable to build well-formed or reasonable queries using formal logic operators. There is now a tendency in computer interfaces towards a more natural-language syntax, so that many systems now assume conjunction as the default

operator, and some online dictionaries try to accommodate natural-language queries (the *OneLook Reverse Dictionary* being one case in point). No doubt one reason for this is the poor uptake of formal logic syntax (Markey, 2007). In dictionary searches, few users would find the need for Boolean operators, and not just because they are difficult to formulate, but because rarely is a dictionary user's idea of what they are looking for readily expressible as a logical formula. Successful use of a Boolean search requires the knowledge of the form of operators (e.g. 'AND' or '&'; 'OR' versus '|'; 'NOT' or '!' or '~' or '-'), as well as their semantics. Some dictionaries may support (a subset of) regular expressions (Pastor & Alcina, 2010). Obviously, the skill of using regular expressions is largely restricted to a small percentage of dictionary users, mostly those with some programming experience.

A 'lightweight' implementation of a Boolean search is one which uses separate descriptive text fields rather than logical operators, usually 'all the words' and 'any of the words'. Such an approach is less flexible than an expression-based query, as it restricts a single query to either a conjunction or disjunction of terms, but should be easier to grasp thanks to being more intuitive, and some users may be familiar with the choices from web-based search experience (such as from using an advanced interface of an internet search engine).

3.1.4 Filtered search

Certain electronic dictionaries include various filters capable of restricting search results to a well-defined subset of the lemmas. This could be based on formal (e.g. part of speech), distributional (frequency), semantic (e.g. subject domain), or pragmatic (e.g. taboo, slang, formal, spoken, humorous) properties. The prerequisite for the ability to use such filters is the users' metalinguistic and metalexigraphic awareness of the existence and significance of these categories.

3.1.5 Sound search

Dictionary access via a phonological (or phonetic) representation has been the focus of Sobkowiak's work (1999). One purpose of using sound-based selection would be pedagogical: to make it possible to select words with specific interesting properties, such as problematic phonotactic sequences, so that they can be put to use in language-teaching practice. Another possible application is accessing items whose orthographic representation is unknown. However, for languages such as English at least, with relatively complex phoneme inventories, it is doubtful if most users, be it learners or native speakers, would be able to correctly input phonological representations by typing in or clicking on phonemic symbols. Such skills are just too demanding for most but a minority of users (such as language professionals). There may be greater promise in voice-recognition-based access, and the goal of the technology is to require a minimum of special skills.

3.1.6 Fuzzy-spelling search

Engelberg & Lemnitzer (2009) treat fuzzy-spelling search as a dedicated search option; however, modern online dictionaries tend to have this as an always-on feature in the form of a 'did you mean' function, which provides target item suggestions for possibly misspelled queries. The quality of this function in even the best online dictionaries still leaves considerable room for improvement (Lew & Mitton, 2011; 2013). No special skill should be required to use fuzzy-spelling search; it is in fact designed to compensate for insufficient skills in using standard spelling. Nevertheless, the user still needs to be able to interpret the list of suggestions normally returned by the 'did you mean' function.

3.1.7 Anagram search

The need to search for anagrams is probably largely restricted to dictionary users engaging in word games. Such users, often driven by a particular passion, usually know quite well what they are doing when using lexical tools to help them solve lexical puzzles.

3.1.8 Inflected form search

In inflected languages, many actual and potential word forms are subsumed under a single citation form used by convention as a lemma sign. Again, as in fuzzy-spelling search, the ability of a dictionary to take the user to the right entry from an inflected form should help in those cases when users have problems reducing to the citation form of a word, or if they are not aware that dictionaries conventionally nest word forms under a single form. The importance of this function is rather greater for heavily inflected languages. For example, Russian includes aspectual pairs of verbs, and in print dictionaries it is sometimes hard to guess which member of the pair one should look up.

3.1.9 Index-based search

Index-based search consists in locating a term on a list, usually arranged vertically. This mode is reminiscent of print dictionary consultation, but there are differences. Navigation of the index list may be enhanced with search-as-you-type technology. The index may contain not just article headwords, but also sublemmatic items, such as nested derivatives or multi-word items, but this may not be necessarily clear to all users, and some, out of habit, will want to proceed via the main headword. On the other hand, some internet dictionaries include clickable letter sections, so the user first needs to click on the initial letter and then further navigate the target letter section. This is somewhat parallel to a thumb index in a print dictionary, and calls for somewhat similar skills, but of course translated into the ergonomics of the computer.

3.1.10 External-text-based search

This access mode refers to cases when lexicographic assistance is requested for an item displayed on screen, embedded in an electronic text. A case in point is the Google dictionary plug-in (available for the Chrome browser), which displays a call-out with a definition upon clicking a word anywhere on a webpage. This is an economical and user-friendly option, especially if accompanied by inflected form reduction, plus, ideally, contextual awareness so that multi-word units can automatically be identified (Lew, 2012a) and some sense disambiguation is effected. The skill required to use this search mode correctly is basically restricted to an awareness of the option to click on the word most likely requiring lexicographic support.

3.1.11 Picture-based search

In dictionaries featuring synoptic pictures which combine elements of a particular complex scene or setting, such as ‘the airport’, linking the elements of the picture to their lexicographic information is possible. This may be an efficient way to use a dictionary to get to know specific lexical fields, such as preparing for an oral examination on a particular topic. It seems that skill requirements for this type of access are low, and largely limited to an awareness of the fact that labels are linked to entries. Things get considerably more difficult if elements of a picture remain unlabelled by default.

3.1.12 Scanner-based search

This look-up mode refers to optical scanning devices which convert print to electronic text (utilizing character recognition technology). Skills involved are dependent on the particular implementation of the technology, be it reading-pen or point-and-shoot.

3.1.13 Further search options in Pastor & Alcina’s (2010) model

Pastor & Alcina (2010) identify some search options beyond those proposed by Engelberg and Lemnitzer. However, the fifteen resources they examine include some lexical databases whose status as a dictionary may be debated. Consequently, the associated search techniques may be rather untypical of dictionaries in the narrower sense.

Pastor & Alcina try to systematize their description of search techniques by breaking down the search event into three components: the query (expression introduced by the user), the resource (element of the dictionary interface), and the result (what the dictionary presents back to the user).

With regard to the query, they distinguish searches for 1) an exact word, 2) a partial word, 3) an approximate expression (which subsumes inflected form and spelling similarity), 4) an anagram, and 5) a combination of two or more words (2010: 320).

These types of queries entail particular user skills related to the formulation of the search. A partial-word search involves appropriate skills to indicate truncation (cf. 3.1.2 above). I have already discussed search options based on spelling similarity as well as anagrams.

Pastor & Alcina (2010) point out that some electronic dictionaries offer multiple search entry points (which they dub ‘resources’). Clearly, in such cases users should learn to select the one that is appropriate. Most similar to traditional print dictionaries is a list of headwords, where the user would enter the search word. Such a list (called the *entry field* by Pastor & Alcina) may be extended to include multi-word expressions (as in the *Macmillan English Dictionary Online*). Some dictionaries may allow searching *content fields* (Pastor & Alcina, 2010: 332) such as definitions, examples, or even a corpus accompanying the dictionary, where available. Relevant user skills in this case would be (1) recognizing the entry points available; (2) selecting the entry point that best meets their information need; and (3) adapting their query so that it makes good sense at the particular entry point. Bank (2010), for example, notes frequent cases of users of the *Base lexicale du français* resource being misled about the entry point and searching the Leuven University website rather than the dictionary. This particular problem stemmed primarily, as Bank rightly pointed out, from the design problems of the resource, but user-friendliness and user skills are two complementary sides of the same lexicographic coin.

The third component of a search event in Pastor & Alcina’s model (2010) is the search result. Electronic dictionaries may present the complete entry, or they may only give a list of headwords. An intermediate possibility is a list of incomplete entries or entry snippets (as in *COBUILD online*).

If a single complete entry is presented, the situation parallels the familiar case of print dictionaries. However, in some cases additional ‘did you mean’ suggestions may appear. Users faced with a mere list of terms need to know that they should select the most likely option to get more complete information. This may be obvious for most, if not all; more challenging are entry snippets, where some users may get stuck at this intermediate level, never getting to see the complete entries, as they fail to realize that more complete information is only a click away.

4. Internet skills: digital literacy, information literacy

Since online dictionaries are offered on the internet, skills for using online dictionaries should not be considered in isolation from skills of using the internet more generally. There are various ways to conceptualize skills related to the use of computer-mediated information retrieval. Two common terms are *digital literacy* and *information literacy* (Bawden, 2008; Lankshear & Knobel, 2008). A search of the relevant literature reveals that these concepts tend to be described in fairly broad terms but usually include recognition of an information need, its nature, and extent.

Slightly more specific items include entering search terms and understanding site navigation. Hargittai (2005) made an attempt to reduce active internet skills to declared familiarity with internet terms.

Web users tend to resort to very simple strategies for internet-based information retrieval. A comprehensive overview by Markey (2007) reports only a minority of web users (less than 15%) making use of the AND operator. This low rate may be one reason why today's major search engines no longer support the operator explicitly. Other operators are used even more sparsely: a tiny 3% for the OR operator, and below 2% for the NOT operator. This may testify to the users' general tendency to gravitate towards natural-language queries. Further, end-users tend not to change the default settings of an information retrieval system (Markey, 2007: 1077). These findings may invite the conclusion that online dictionaries should try to reflect the unsophisticated strategies of general web use. This is a conclusion that many lexicographers find hard to accept, and an argument can be made that a minority of expert users (such as language professionals) are worth catering for as well. Ideally, an online dictionary interface will combine simplicity (for those who cannot be bothered) with sophistication (for those who can). A reasonable way to achieve this is to offer a simple default interface with an optional advanced alternative.

5. Conclusion

The shift to electronic dictionaries is bringing about a parallel change in the skills needed to make efficient use of dictionaries. Some traditional skills are becoming largely obsolete, such as those related to paper page navigation or reducing a word form to its citation form. However, new skills arise from the numerous new search techniques afforded by electronic dictionaries.

A salient component relevant in dictionary-using skills in the electronic age is the movement away from the word-based model implied by print lexicography, and a greater focus on multi-word units and larger text chunks.

An important concern is finding an appropriate context for teaching e-dictionary skills. An online platform for courses integrating dictionary skills and language awareness, preferably embedded in the curriculum, appears promising (Ranalli, 2013).

6. References

- Atkins, B.T.S. (1996). Bilingual dictionaries - past, present and future, in Gellerstam, M., Jarborg, J., Malmgren, S.-G., Noren, K., Rogström, L. & Pappmehl, C.R. (eds.), *EURALEX '96 Proceedings*. Göteborg: Department of Swedish, Göteborg University, pp. 515-546.
- Bank, C. (2010). Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen. M.A., Universität Hildesheim.

- Bawden, D. (2008). Origins and concepts of digital literacy, in Lankshear, C., Knobel, M. (eds.), *Digital literacies: concepts, policies and practices* Peter Lang, pp. 17-32.
- Bishop, G. (2000). Developing learner strategies in the use of dictionaries as a productive language learning tool, *Language Learning Journal* 22, pp. 58-62.
- Bishop, G. (2001). Using quality and accuracy ratings to quantify the value added of a dictionary skills training course, *Language Learning Journal* 24, pp. 62-69.
- Campoy Cubillo, M.C. (2002). Dictionary use and dictionary needs of ESP students: An experimental approach, *International Journal of Lexicography* 15(3), pp. 206-228.
- Carduner, J. (2003). Productive dictionary skills training: What do language learners find useful?, *Language Learning Journal* 28, pp. 70-76.
- Chi, M.-L.A. (1998). Teaching dictionary skills in the classroom, in Fontenelle, T., Hilgsmann, P., Michiels, A., Moulin, A. & Theissen, S. (eds.), *EURALEX '98 Actes/Proceedings*. Liege: Université Départements d'Anglais et de Néerlandais, pp. 565-577.
- De Schryver, G.-M. (2012). Trends in twenty-five years of academic lexicography, *International Journal of Lexicography* 25(4), pp. 464-506.
- Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*, 4th edition. Tübingen: Stauffenburg Verlag.
- Hargittai, E. (2005). Survey measures of web-oriented digital literacy, *Social Science Computer Review* 23(3), pp. 371-379.
- Hartmann, R.R.K. (1999). Lexical reference books - what are the issues?, *International Journal of Lexicography* 12(1), pp. 5-12.
- Herbst, T., Stein, G. (1987). Dictionary-using skills: A plea for a new orientation in language teaching, in Cowie, A.P. (ed.), *The dictionary and the language learner. Papers from the EURALEX Seminar at the University of Leeds, 1-3 Apr. 1985, (Lexicographica Series Maior 17)*. Tübingen: Niemeyer, pp. 115-127.
- Kennedy, L.D. (1972). The teaching of dictionary skills in the upper grades, *Elementary English* 49(1), pp. 71-73.
- Langegard, A.-M. (2011). How are digital dictionaries used by young Norwegian learners of EFL? A case study of attitudes and practices. M.A., Oslo University.
- Lankshear, C., Knobel, M. (eds.). (2008), *Digital literacies: concepts, policies and practices*: Peter Lang.
- Lew, R. (2011). Online dictionaries of English, in Fuertes-Olivera, P.A., Bergenholtz, H. (eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, pp. 230-250.
- Lew, R. (2012a). How can we make electronic dictionaries more effective?, in Granger, S., Paquot, M. (eds.), *Electronic lexicography*. Oxford: Oxford University Press, pp. 343-361.
- Lew, R. (2012b). The role of syntactic class, frequency, and word order in looking up English multi-word expressions, *Lexikos* 22, pp. 243-260.
- Lew, R. (in press). Space restrictions in paper and electronic dictionaries and their

- implications for the design of production dictionaries, in Bański, P., Wójtowicz, B. (eds.), *Issues in Modern Lexicography*. München: Lincom Europa.
- Lew, R., Galas, K. (2008). Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English, in Bernal, E., DeCesaris, J. (eds.), *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 1273-1285.
- Lew, R., Mitton, R. (2011). Not the word I wanted? How online English learners' dictionaries deal with misspelled words, in Kosem, I., Kosem, K. (eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011, Bled, 10-12 November 2011*. Ljubljana: Trojina – Institute for Applied Slovene Studies, pp. 165-174.
- Lew, R., Mitton, R. (2013). Online English learners' dictionaries and misspellings: One year on, *International Journal of Lexicography* 26(2), pp. 219-233.
- Lew, R., Tokarek, P. (2010). Entry menus in bilingual electronic dictionaries, in Granger, S., Paquot, M. (eds.), *eLexicography in the 21st century: New challenges, new applications*. Louvain-la-Neuve: Cahiers du CENTAL, pp. 193-202.
- Markey, K. (2007). Twenty-five years of end-user searching, Part 1: Research findings, *Journal of the American Society for Information Science and Technology* 58(8), pp. 1071-1081.
- Nesi, H. (1999). The specification of dictionary reference skills in higher education, in Hartmann, R.R.K. (ed.), *Dictionaries in language learning. Recommendations, national reports and thematic reports from the Thematic Network Project in the Area of Languages, sub-project 9: dictionaries*. Berlin: Freie Universität Berlin, pp. 53-67.
- Osuchowska, D. (2003). The do's & don'ts of teaching dictionary reference skills at the college/university level, *Papers From the Third Chełm Symposium Held in April 2003*. Chełm: NKJO-CHEŁM PUBLISHERS.
- Pastor, V., Alcina, A. (2010). Search techniques in electronic dictionaries: A classification for translators, *International Journal of Lexicography* 23(3), pp. 307-354.
- Ranalli, J. (2013). Online strategy instruction of integrated dictionary skills and language awareness, *Language Learning & Technology* 17(2), pp. 75-99.
- Ronald, J., Ozawa, S. (2011). Electronic dictionary use: Identifying and addressing user difficulties, in Akasu, K., Uchida, S. (eds.), *ASIALEX2011 Proceedings Lexicography: Theoretical and practical perspectives*. Kyoto: Asian Association for Lexicography, pp. 436-446.
- Sobkowiak, W. (1999). *Pronunciation in EFL machine-readable dictionaries*. Poznań: Motivex.
- Van der Merwe, M. (2012). A study of the use of the HAT Afrikaanse Skoolwoordeboek by primary school children, *Lexikos* 22, pp. 352-366.

Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing

Iztok Kosem¹, Polona Gantar², Simon Krek³

¹Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia

²Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia

³Jožef Stefan Institute, Ljubljana, Slovenia

E-mail: iztok.kosem@trojina.si, apolonija.gantar@guest.arnes.si, simon.krek@guest.arnes.si

Abstract

A new approach to lexicographic work, in which the lexicographer is seen more as a validator of the choices made by computer, was recently envisaged by Rundell and Kilgarriff (2011). In this paper, we describe an experiment using such an approach during the creation of the Slovene Lexical Database (Gantar & Krek, 2011). The corpus data, i.e. grammatical relations, collocations, examples, and grammatical labels, were automatically extracted from the 1.18-billion-word Gigafida corpus of Slovene. An evaluation of the extracted data consisted of making a comparison between a manual entry and a (semi)-automatic entry, and identifying potential improvements in the extraction algorithm and in the presentation of data. An important finding was that the automatic approach was far more effective than the manual approach, without any significant loss of information. Based on our experience, we would propose a slightly revised version of the approach envisaged by Rundell and Kilgarriff in which the validation of data is left to lower-level linguists or crowd-sourcing, whereas high-level tasks such as meaning description remain the domain of lexicographers. Such an approach indeed reduces the scope of lexicographers' work; however, it also results in the ability of making content available to the users more quickly.

Keywords: automatic extraction, crowd-sourcing, Slovene Lexical Database, validation

1. Introduction

The last decade has been very eventful for lexicography, mainly due to technological progress. This allowed the building of larger and larger corpora, providing lexicographers access to increasingly larger databases of language. In addition, the introduction of the electronic medium and the online format in particular, which has truly established itself as the main medium for dictionary content in most parts of the world, has meant that dictionary content can be available to users faster than ever before.

However, technological progress has also brought about new challenges for lexicographers: there is (much) more data to analyze, and less time to do so due to (more) demanding users. Various tools such as Word Sketch (Kilgarriff and Tugwell, 2002) and TickBox Lexicography (Kilgarriff et al., 2010) have been designed as part of corpus query systems to help lexicographers tackle this problem, but their design and purpose still requires lexicographers to select and transfer relevant corpus information to the dictionary writing system.

These new challenges for lexicographers have prompted researchers to rethink the definition of what lexicographer's work should entail. Recently, a new approach to lexicographic work, in which the lexicographer is seen more as a validator of choices made by a computer, was envisaged by Rundell and Kilgarriff (2011). As they argue “it is more efficient to edit out the computer’s errors than to go through the whole data-selection process from the beginning”. This approach redefines not only the lexicographer’s tasks but also the role of a corpus in the lexicographic process.

In this paper, we describe an experiment using such an approach during the creation of a new lexical database for Slovene. Firstly, we present the lexical database, describing its contents and structure. Next, we focus on the method of automatic data extraction from the corpus, outlining the elements needed for developing the algorithm for data extraction, and describing the output. Then, we focus on evaluation of the automatic method, by comparing it with the “manual” method used in the early stages of building the lexical database, examining its accuracy, and pointing out the parts that can still be improved. A section is dedicated to a planned implementation of automatic methods in the compilation of a proposed new dictionary of contemporary Slovene, where crowd-sourcing would also be utilized as a clean-up stage between automatic extraction of data and lexicographic editing. We conclude by considering future improvements of the method, as well as discussing which other approaches could be made more automatic and combined with the method presented here.

2. Slovene Lexical Database

The Slovene Lexical Database (SLD) is one of the results of the Communication in Slovene¹ project, a project that has developed language data resources, natural language processing tools and resources, and language description resources for Slovene. The SLD has a twofold goal: it is intended as the basis for the future compilation of different dictionaries of Slovene, both monolingual and bilingual, and as such its concept is biased towards lexicography. Secondly, it will be used for the enhancement of natural language processing tools for Slovene.

Reflecting its two-fold purpose, the SLD contains two different types of information. On the one hand, there is lexico-grammatical information – intended for human end users – such as sense descriptions in semantic frames, representing the starting point for whole sentence definitions (Sinclair, 1987), collocations attributed to particular senses of the lemma, and examples from the corpus. On the other hand, there is

¹ The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational program Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013. Project web page: <http://eng.slovenscina.eu/>.

information designed for natural language processing tools. This information is encoded in a more complex way and, in addition to its immediate use in NLP tools, requires an expert to process or interpret it. Among this information is the formal encoding of syntactic patterns on the phrasal and clause level as well as the formal encoding of semantic arguments and their types.

The database is conceptualized as a network of interrelated lexico-grammatical information on six hierarchical levels with the semantic level functioning as the organizing level for the subordinate ones. The six levels are:

- a) Lemma, or the headword, representing the top hierarchical level and functioning as the umbrella for all lexical units placed under it.
- b) Senses and subsenses, labelled with semantic indicators, whose primary function is to form a sense menu intended for easy navigation within a polysemic entry structure. Another kind of information recorded on the sense level is semantic frames which are conceptually close to frames in the FrameNet project (Fillmore & Atkins, 1992; Baker, Fillmore & Cronin, 2003) and to prototypical syntagmatic patterns in the Corpus Pattern Analysis system (Hanks, 2013).

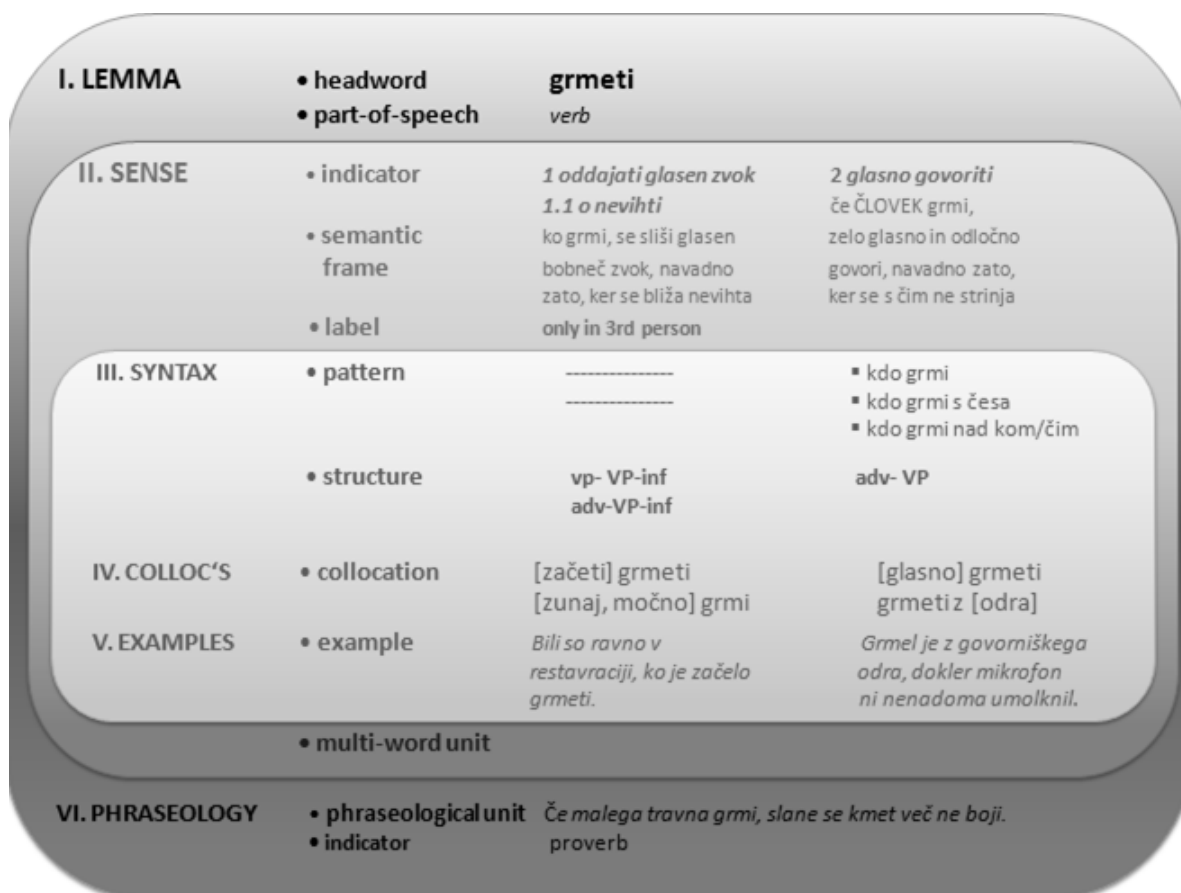


Figure 1. Structure of the Slovene lexical database

- a) Multi-word expressions, which are registered only for noun or adjective headwords. Multi-word expression must demonstrate a non-compositional idiosyncratic sense.
- b) Syntactic structures, representing a formalization of typical patterns on the clause and phrasal level and primarily intended for natural language processing tools.
- c) Collocations and examples. On the collocation level, patterns and structures are verified by recording typical collocates of the headword realized in the anticipated syntactic positions. Collocations and its related parent levels (patterns, structures and frames with semantic types) are attested with corpus examples.

3. Compiling entries using automatic extraction of data

The decision to introduce automatic extraction of data from the corpus was made early in the process of compiling an entry, as it became obvious that there were several bottlenecks. We used the Sketch Engine (Kilgarriff et al., 2004), a leading lexicographic tool for corpus analysis, with (lexicographic) functions such as Word Sketch and TickBox Lexicography; however, the time spent on selecting under each syntactic structure the relevant collocates and their examples, and copy-pasting them into a dictionary-writing system was considered excessive.

The time-consuming nature of these tasks also had a negative effect on lexicographers' distribution of time (and effort) to different tasks. For example, for headwords with many (sub)senses and syntactic patterns, lexicographers could on average dedicate less time to identifying different (sub)senses and devising semantic frames and indicators for each (sub)sense.

3.1 Methodology

The procedure of automatic extraction provided lexical information, related to grammatical structures recorded in the lexical database, from the 1.18-billion-word Gigafida corpus of Slovene (Logar Berginc et al., 2012). The information was extracted in an XML format and imported into the iLex dictionary-writing software (Erlandsen, 2004). The relevant lexical information comprised collocations and related corpus examples. The procedure required the following:

- i. a selection of lemmas for extraction,
- ii. finely-grained sketch grammar, designed specifically for the purposes of automatic extraction,
- iii. GDEX (Good Dictionary EXamples; Kilgarriff et al., 2008) configuration(s),

- iv. an API script to extract data from word sketch information in the Sketch Engine, and
- v. settings for extraction (e.g. minimum collocation frequency, minimum collocation salience).

3.2 Selecting lemmas

We wanted to focus on a group of lemmas that would enable an evaluation without the problem of large quantities of data, and that would be more homogeneous in nature as to facilitate gradual improvement of GDEX configurations and settings for extraction. Thus, lemmas had to fulfil three criteria:

- a) Frequent enough to offer a good-sized word sketch. Namely, initial testing showed that word sketches for less frequent lemmas (less than 600 hits in Gigafida) did not provide enough relevant data. Consequently, we divided lemmas of each word class into five different frequency groups, and then focussed on frequency ranges that provided the best word sketches for a manageable number of lemmas.
- b) Monosemous or having up to two synsets/senses in sloWNet, a Slovene version of Wordnet (Fišer, 2009), or, exceptionally, in the Dictionary of Standard Slovenian (SSKJ).
- c) Found in sloWnet, preferably, but not in SSKJ, as we wanted to focus on new words and/or senses.

The final selection included 515 nouns, 260 verbs, 275 adjectives and 117 adverbs and was dominated by lemmas with frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words). There were a few lemmas with frequency below or above this range for the purposes of additional testing, especially for testing the effectiveness of the API script in extracting data for all grammatical relations in the sketch grammar.

3.3 Sketch grammar

The sketch grammar (Krek and Kilgarriff, 2006), designed specifically for automatic extraction, utilized the directives *CONSTRUCTION, *COLLOC and *SEPARATEPAGE; elements that represented new additions to the Sketch Engine at that time. The first of these three directives enables the identification of grammatical relations without collocations, which is particularly useful for extraction of verb patterns. The second directive is used to identify elements that are categorized as syntactic combinations in the lexical database, such as preposition-noun-preposition. The third directive is intended for creating a separate word sketch page for relations with three elements (directive *TRINARY), which enables the introduction of relations with prepositions that can have more specific definitions (for example they

can include the case of the preposition).²

directive	number of gramrels
TRINARY	36
DUAL	25
UNARY	2
CONSTRUCTION ³	13
CONSTRUCTION+UNARY	6
COLLOC	3
SYMMETRIC	2
no directive	18
total	105

Table 1: Gramrels by directives

The new sketch grammar included all the structures registered in the lexical database, and therefore contains significantly more gramrels (grammatical relations) than the sketch grammar used for preparing data for manually compiled entries. There are 103 gramrels in total; categorization is shown in Table 1.

All the directives with three elements (*TRINARY) were used with a separate page output. The combination CONSTRUCTION+UNARY was used to alert the lexicographers, in a separate column called Constructions, to gramrels occurring very frequently in the corpus (this is the main function of the UNARY directive). Using this directive, we can also automatically generate alerts such as *pogosto zanikano* (often in negative), *pogosto v 3. os. ednine* (often in 3rd person singular), etc., that are recorded in <oznaka> (label) tag in the database and are candidates for labels in the dictionary.

Each gramrel in the sketch grammar contains the information about the name of the structure in the lexical database, for example:

```
*DUAL
=S_v_rodil-s/S_s-koga-česa
```

The structure used to extract combinations of a noun in any case with a noun in genitive (e.g. *delovanje motorja*, ‘working of an engine’ (gen.)) is recorded in the lexical database as SBZO sbz2, if the headword is the head noun, or as sbZO SBZ2, if the headword is a noun in the genitive case. The relevant information is added to each gramrel:

² This was not possible in earlier sketch grammars as it would result in a very high number of relations/columns in the word sketch.

³ For more on the CONSTRUCTION directive, see Rychlý (2010) and Krek (2012).


```
# LBS-XX #####
# /1/ <struktura>SBZO sbz2</struktura>
# /2/ <struktura>sbzo SBZ2</struktura>
#####
```

The sketch grammar presented above is intended solely for the purposes of automatic extraction of data from the corpus, as it produces word sketches that are difficult to process by a human user due to a high number of relations and their complex naming system.

3.4 GDEX configurations

Corpus examples are an important part of the lexical database, as they attest word senses, definitions, collocations, patterns, domain and genre-related characteristics, pragmatics, etc. According to Atkins and Rundell (2008: 458), a good corpus example should meet at least three criteria: naturalness and typicality, informativeness and understandability. However, as corpora are becoming larger and larger, it means there is more data to analyze, which is making the search for good examples more and more difficult and time-consuming.

GDEX is a tool that assists lexicographers in finding good corpus examples by ranking them according to their quality. Ranking is done on the basis of parameters such as example length, whole sentence form, syntax, and presence/absence of rare words, etc., which are measurable and in some way connected with the aforementioned criteria for a good example.

The first version of GDEX for Slovene (Kosem et al., 2011) was developed to meet the needs of lexicographers compiling manual entries in the lexical database. The existing version of GDEX for Slovene was not suitable for the purposes of automatic extraction due to differences in the relationship between computer and lexicographer. In the normal, “manual” procedure the lexicographer uses corpus tools to analyze corpus data, selects them and transfers them into dictionary-writing software. The role of GDEX was to provide at least three good examples among the ten offered in the TickBox Lexicography.

In the automatic procedure, on the other hand, the data is automatically exported from the corpus into dictionary-writing software, where they are examined, selected and edited by the lexicographer. The main aim was to reduce manual inserting of data in the database, and to reduce the need for manual removal of irrelevant or incorrect information; therefore, the aim was to design a GDEX configuration where the **top three** examples would meet the criteria of a good example.

The experience from designing the first GDEX for Slovene indicated that GDEX results could be improved by devising a separate configuration for each word class. Thus, four different GDEX configurations were prepared, for nouns, verbs, adjectives, and adverbs, respectively. All configurations contained classifiers, listed in Table 2,

but differed in settings. Initial configurations, which did not contain all the listed classifiers, were devised from the first GDEX of Slovene, with values of classifiers set by analyzing existing examples in the lexical database that were manually selected by lexicographers.

- whole sentence
- contains token with frequency of less than 3
- sentence longer than 7 tokens
- sentence shorter than 60 tokens
- lemma is repeated
- contains email address or URL
- optimum length (between X and Y tokens)
- contains rare lemmas
- contains token, longer than 12 characters
- number of punctuation marks (excluding commas)
- number of commas
- tokens starting with a capital letter
- tokens containing mixed symbols (e.g. letters and numbers)
- number of personal names
- number of pronouns
- position of lemma
- stop list of words at the beginning
- stop list of phrases at the beginning
- second collocate (collocate of a collocation)
- Levenshtein distance

Table 2: GDEX classifiers for automatic extraction

After initial configuration for each word class was devised, it was tested in the Sketch Engine by evaluating examples for a sample of lemmas from the selection that would be used in the automatic extraction. Then, values for classifiers were modified according to observations during evaluation, and a new configuration was devised. The evaluation then compared the results given by both configurations, and further modifications were made. The procedure was repeated until the GDEX configurations that provided the most satisfactory results were obtained. An important consequence of this method was the formation of several new classifiers, which were not found in the first GDEX for Slovene. Particularly noteworthy additions are stop lists of words and phrases at the beginning of examples and second collocate (collocate of a collocation). The latter classifier brought significant improvement to the results of automatic extraction because it indirectly detects colligational typicality of a collocation. For example, for the collocation *klavrn* + *podoba* ('poor image'), the classifier awards points to examples with the second collocate *kazati* ('show'), and consequently, the configuration containing this classifier offers examples containing typical structures of this collocation: *kazati klavrno podobo česa* ('show poor image of sth').

3.5 Preparing the API script

The API script for automatic extraction was written in Python and required certain updates to the Sketch Engine tool. Before the API script could be run, word sketch had to be created using the sketch grammar for automatic extraction. The following parameters had to be set when running the script:

- corpus
- lemma (or a list of lemmas in a file)
- gramrel (or a list of gramrels in a file)
- GDEX configuration
- number of examples per collocate
- number of collocates per grammatical relation
- minimum frequency of a collocate
- minimum frequency of a grammatical relation
- minimum salience of a collocate
- minimum salience of a grammatical relation.

An XML template for extracted data had to be prepared, and its structure matched with the DTD of the lexical database to enable importing of automatically extracted data into the dictionary-writing program. In order to make the exported data easier to view, we added attributes to <kolokacija> and <zgled> in the DTD, namely, an ID for a collocate, so that the connection between a collocate and its examples was maintained; the index number of a token in the <zgled> element, which also enables an identification of an example in the corpus; and a number for each example of a collocate, reflecting the GDEX ranking.

3.5.1 Setting the parameter values

Initial tests in automatic extraction used the following settings: 10 collocates per relation, 6 examples per collocate, minimum salience of a relation or collocate = 0, minimum frequency of a collocate = 0, and minimum frequency of a relation = 25; however, the evaluation showed that the same settings cannot be used for all the relations and collocates, since the output contained many irrelevant relations and associated collocates, or missed relevant relations and collocates. Also, the number of examples had to be reduced as editing took too long.

Initial settings were improved by obtaining the statistical data for grammatical relations and collocates, available in word sketches, of all the lemmas for automatic extraction; then, the values for each relation within lemmas of a word class were analyzed to obtain the optimal minimum frequency and salience of the relation. Also relevant was information on the percentage of the lemma occurrences in a particular relation.

The statistical analysis was combined with manual analysis of word sketches, and the finding was that if a relation covered a low percentage of occurrences of a lemma, it was often not a candidate for automatic extraction for that lemma. An additional benefit of manual analysis of word sketches was that it led to the identification of a few shortcomings in the sketch grammar (e.g. incorrectly defined or classified gramrel), which were then corrected before the final automatic extraction. Minimum frequency and salience values for collocates were determined by examining the collocates under each gramrel for each of the word classes, and identifying the lowest values where the collocation still yielded relevant results.

The analysis of data extracted using initial settings showed that the number of collocates per grammatical relation was a very important parameter. Namely, if the first ten collocates (default settings) did not exceed the minimum frequency or salience, the relation was not extracted, even if it is very frequent. As a result, the minimum number of collocates per relation was increased to 25, and the selection of relevant collocates was 'left' to minimum frequency and salience settings. The number of examples per collocate was reduced to three, as the evaluation showed that in most cases at least one of the top three examples offered by GDEX was good (in fact, often all three were good).

Another issue encountered was that in some cases an entire relation, which was frequent for a particular lemma, was not extracted because none of its collocates was above the frequency and/or salience threshold. However, this issue was mainly observed with low frequency lemmas and was solved by dividing lemmas into frequency groups, and preparing separate settings for each group.

3.6 Evaluation

In order to be able to evaluate whether using automatically extracted data is time-effective, we first finalized the entries for headwords with automatically extracted data. Then, we compared the time needed to manually devise an entry in the lexical database (i.e. selecting the relevant corpus data, mainly on the basis of analysing word sketches, transferring it into the dictionary-writing system, and adding other information), with the time needed to devise an entry using the automatic method. The results clearly favoured the approach using the automatic method: on average, using the manual method, it takes a lexicographer just over four hours to devise an entry (0.23 entries per hour), whereas using the automatic method, a lexicographer devises an entry in two hours (0.5 entries per hour). Consequently, the automatic method more than halves the time required to devise dictionary entries.

Another aim of evaluation was to identify the (lexicographic) work required to create final entries from the automatically extracted data, and to assess the reliability of the automatic method. The automatic method renders some routine tasks unnecessary, such as copying the data to a dictionary-writing system, but under the condition that

the lexicographer does not often need to consult the corpus to add missing information. The evaluation showed that the automatic method was very reliable, and extracted examples always attested for all (sub)senses of the headword. In comparison with the manual method, the entries showed differences in terms of sense division and definitions, which was expected as they were devised by different lexicographers, but the main finding was that none of the information needed to devise the entries was lost using the automatic method.

Tasks still allocated to lexicographers are of two types: analytical and editorial. Analytical tasks comprise sense division, preparing sense indicators and definitions, identification of compounds, phrases and pragmatic characteristics of meanings, and adding style and domain labels. Editorial tasks include distributing the extracted information according to the information added by lexicographers (e.g. collocates under the relevant sense), copying grammatical relations and collocates if they are typical for more than one (sub)sense, and deleting irrelevant relations, collocates and corpus examples.

The evaluation indicated that editorial tasks can sometimes still take a considerable amount of time when devising an entry. Although some can be eliminated or shortened by improving the automatic extraction method or by automating some of the steps (e.g. grouping collocates using the Thesaurus function in the Sketch Engine), these tasks are likely to remain an integral part of lexicographic work. Nonetheless, as the tasks are relatively less demanding in nature, and some are in fact very routine, we wanted to test whether they can be successfully completed by non-lexicographers (people with good knowledge of a language but without lexicographic experience), using the crowd-sourcing process.

3.6.1 Crowd-sourcing

One of the main challenges of trying to introduce crowd-sourcing into the lexicographic process was the design of procedures that would enable quick and successful completion of editorial tasks without the need for extensive learning of the concept and nature of work on the lexical database. We identified three activities that were potentially suitable for crowd-sourcing:

- a) evaluating examples to identify false collocations,
- b) evaluating examples to identify incorrect examples (i.e. the ones where the collocation does not match the grammatical relation it belongs to), and
- c) distributing collocations and their examples under (sub)senses.

The first two activities can be conducted on automatically extracted data and should follow one another, whereas the third activity requires that the analytical work is completed first.

ZAČETNA STRAN OCENJEVANJE BESEDNIH KOMBINACIJ LESTVICA UPORABNIKOV INFO

Ocenjevanje slovnične ustreznosti besednih kombinacij

V tej nalogi vas prosimo, da ocenite, ali kombinacija besed v zgledu ustreza navedeni slovnični strukturi. S pravilnimi odgovori boste iz spletnega slovarja odstranili zglede, v katerih besedne kombinacije ne ustrezajo slovničnim strukturam, pod katere so bile uvrščene na podlagi avtomatskega postopka. Pozorni morate biti predvsem na pripis besedne vrste, sklona in stavčne vloge pri kateri od obarvanih besed v zgledu.

Ali kombinacija besed v zgledu ustreza navedeni slovnični strukturi?

Beseda
franšiza - **samostalnik**

Slovnična struktura
glagol + **za** + **samostalnik v tožilniku**

Zgled
Vsak poslovni sistem - ne glede na to, ali **gre za franšizo** ali ne - ima svoj cilj oziroma poslanstvo, ki vam lahko ustreza ali pa ne.

DA NE Ne vem

30%

Figure 2: Evaluating examples (Task 2) in an online tool

The crowd-sourcing experiment comprised two tasks (covering activities a and b above) that were prepared in an online tool designed specifically for crowd-sourcing and was first used for checking translations in slowNet (Tavčar et al., 2012).

In Task 1, we wanted to identify false collocations through their corpus examples. In many cases, false collocations can be identified with a great degree of certainty without even looking at corpus examples; however, we have established that it is much easier, and more reliable, for non-lexicographers to identify such collocations indirectly, i.e. by evaluating corpus examples. In Task 2, which follows Task 1, the focus is on removing incorrect examples for the remaining collocations (see Figure 2), i.e. examples that do not show the collocation correctly (e.g. do not contain the collocate in the case defined in the relation). Task 2 is more demanding than Task 1, and we provided help for the evaluators in the form of colours for different elements of a grammatical relation.

Both tasks are designed in a way that the question is asked and the data shown, and then the evaluator is offered three possible answers: YES, NO, and DON'T KNOW. For example, the question at Task 1 is: *Would you expect to find the example below in a dictionary under the entry X?* We intentionally wanted to avoid questions such as *How good do you think this example is?* that would require the evaluators to grade the example on a scale.

When preparing the data for crowd-sourcing, we decided not to include all the grammatical relations, as some were too complex for evaluation (e.g. verb

constructions, who + verb + to whom) and some often provided poor results and thus needed an improvement of their definition in the sketch grammar. For each task, we needed to provide a so-called “gold standard”, a set of collocates and their examples with the answer already provided. The examples from the gold standard are then used randomly during the task to help determine the reliability of the evaluator.

The crowd-sourcing experiment is still in its early stages but initial tests have shown high reliability of crowd-sourcing data, also confirming that the tasks are designed appropriately.

4. Putting it all together in a dictionary project

The Slovene Lexical Database has, from the very beginning, been seen as a project that would provide and test new methods, and which could be used in the making of a new dictionary of Slovene. It is worth noting that the last comprehensive dictionary of Slovene (SSKJ) was published in 1991, and since that dictionary took more than 20 years to make, many of its entries were already outdated or lacked information on new meanings and usage by the time the dictionary was published. The new version of SSKJ is expected to be published in 2014; however, since it will combine old data with new information, it is bound to suffer several of the shortcomings of its predecessor. In addition, the second version of SSKJ is likely to be initially available in print format only, which is surprising given that the research shows that Slovene dictionary users, especially younger generations, rarely or almost never use printed dictionaries.

The Slovene language is in need of a completely new description that would reflect the way words and their meanings are perceived in the modern world. In addition, such a description would have to be updated regularly to meet the needs of its users; consequently, it has to exist in an online format. Such a description needs to be made available quickly, and Krek et al. (2013) prepared a proposal for a dictionary of contemporary Slovene (SSSJ) that would provide exactly that, using the methods described in this paper. The proposed dictionary envisages the use of a process of making dictionary entries in five phases:

- a.** Red phase: completely automatic and involves the extraction of grammatical relations, collocates and examples from the corpus.
- b.** Orange phase: consists of crowdsourcing activities, where incorrect or irrelevant data from the red phase are identified and excluded from the database (and the dictionary).
- c.** Yellow phase: the most important phase, in which lexicographers carry out all analytical tasks (e.g. sense division, identifying compounds) on the extracted data, adding missing information if needed. This phase also includes crowdsourcing for routine tasks of distributing collocates and examples under

relevant (sub)senses.

- d. Blue phase: in which specialists such as terminologists and etymologists are consulted.
- e. Green phase: the final editorial check is performed.

Considering the reliability demonstrated by the automatic method, SSSJ would not be offered to users after all entries are completed, but immediately after the automatic extraction of data for all entries, i.e. in the red phase. Then, entries would be updated after the completion of subsequent phases. To alert users to any changes and potential incompleteness of an entry, each entry would contain the information on the phase of the entry and the date of the last update (see Figure 3).

During the making of SSSJ, priority would be given to topical and core vocabulary, and to terminology that is becoming part of general language (even if only for a certain period). Topical vocabulary would be detected by monitoring webpages of news portals, newspapers and other resources. Moreover, new words and meanings would be added regularly, either based on corpus monitoring or on user feedback.

The screenshot shows a dictionary entry for the Slovene word "globalen". On the left, a sidebar contains navigation options: "Pomen", "Oblike", "Sinonimi", "Izvor", "Govor", "Vizualizacija", "Multimedija", and "Statistika". A red arrow points to a date indicator "1. 4. 2013" above a progress bar with five dots, the first of which is green. The main entry area is titled "SLOVAR SODOBNEGA SLOV" and features the word "globalen" (pridevnik) with a pronunciation icon and the frequency "P 3000". The entry is organized into three main sections:

- 1. svetovni; mednarodni**: Includes the sub-section "1.1 splošno veljaven; razširjen" with the definition "če postanejo neke dejavnosti ali lastnosti globalne, jih upošteva vedno več" and an example: "Merila, kakšna ženska je lepa, postajajo vse bolj globalna."
- 2. zemeljski; planetarni**: Includes the definition "globalne spremembe v okolju vplivajo na celoten zemeljski planet" and an example: "Eden najbolj preprostih in praktično izvedljivih načinov za zmanjšanje globalne"

Figure 3: Date and stage information in the proposed dictionary of contemporary Slovene

The methods to be used in making the proposed dictionary are not new, if taken individually, as similar methods have been used in dictionary projects around the world. For example, automatic extraction has been used in the making of automatic collocation dictionaries (Kilgarriff et al., 2013); crowdsourcing, albeit in a different form, has been used by the Oxford English Dictionary, Macmillan English Dictionary, and Wordnik, etc. However, the proposal introduces a new concept of compiling a dictionary using automatically extracted data as a point of departure. Lexicographic analysis is still corpus-based (or driven); however, the initial selection of corpus data to be analyzed is left to the computer. The lexicographer then examines, validates, and completes the information and shapes it into the final dictionary entry. The benefits of using this approach for making a dictionary are particularly significant for languages where a dictionary needs to be made from scratch, and needs to be available to users almost immediately.

5. Conclusion

Lexicography is not far from making the vision of Rundell and Kilgarriff a reality. Automatization can be implemented in many aspects of lexicographers' work, saving considerable amounts of time and money. Nonetheless, some tasks, especially anything connected with meaning, remain in the domain of lexicographers, at least for now.

Our experience from preparing the Slovene Lexical Database supports these claims, but also shows that the implementation of automatic procedures calls for a different division of human work, and the introduction of a new participant to the lexicographic project. In this new division of work, lexicographers focus on more difficult, analytical tasks, whereas non-lexicographers (via crowdsourcing) are used for less demanding, more routine tasks. Such a division of work speeds up the dictionary-making process and should be particularly useful in the age of e-lexicography, when users demand immediate access to up-to-date lexicographic information.

In summary, we propose a slight revision of the approach proposed by Rundell and Kilgarriff; in our adaptation, there are three elements: a computer, a non-lexicographer and a lexicographer. The computer provides data, the non-lexicographer cleans it for the lexicographer (separating the wheat from the chaff), as well as redistributing it, and the lexicographer shapes it into the final product.

6. References

- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Baker, C.F., Fillmore, C.J. & Cronin, B. (2003). The Structure of the FrameNet Database. *International Journal of Lexicography* 16(3), pp. 281-296.
- Erlandsen, J. (2004). iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004). Brno, 6. – 7. september 2004*. Available at: <http://nlp.fi.muni.cz/dws2004/pres/#15>.
- Fillmore, C.J., Atkins, S.B.T. (1992). Towards a Frame-based organization of the lexicon: the semantics of RISK and its neighbors. In A. Lehrer, E. Kittay (eds.) *Frames, Fields, and Contrasts: New Essays in Semantics and Lexical Organization*. Hillsdale: Lawrence Erlbaum, pp. 75-102.
- Fišer, D. (2009). SloWNet – slovenski semantični leksikon. In M. Stabej (eds.) *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 145–149.
- Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) *Natural language processing, multilinguality: sixth international conference, Modra, Slovaška, 20-21 Oktober 2011*, pp. 72-80.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
- Kilgarriff, A., Husak, M., Jakubicek, M. (forthcoming) *eLex 2013 Proceedings, 17-19 October 2013, Tallinn, Estonia*.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, 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., Kovář, V., Rychlý, P. (2010). Tickbox Lexicography. In S. Granger, M. Paquot. *eLexicography in the 21st century: New challenges, new applications*. Brussels: Presses universitaires de Louvain, pp. 411-418.
- Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the 11th Euralex International Congress*. Lorient: Université de Bretagne-Sud, pp. 105-116.
- Kilgarriff, A., Tugwell, D. (2002). Sketching words. In H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Euralex, pp. 125-137.
- 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, Bled, 10-12 November 2011*. Ljubljana:

- Trojina, Institute for Applied Slovene Studies, pp. 151-159.
- Krek, S. (2012). *New Slovene sketch grammar for automatic extraction of lexical data*. Presented at SKEW3 workshop, 21-22 March 2012, Brno, Czech Republic. Available at:
http://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw
- Krek, S., Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec, J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Ljubljana, Slovenia. Available at:
http://nl.ijs.si/is-ltco6/proc/12_Krek.pdf.
- Krek, S., Kosem, I. Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*, v1.1. Available at:
http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Rychlý, P. (2010). *Extensions (completed, and planned) to formalism*. Presented at Sketch Grammar Workshop, 3-4 February 2010, Faculty of Social Sciences, Ljubljana, Slovenia. Available at:
http://projekt.slovenscina.eu/Media/BesedneSkice/Predstavitve/Pavel/extensions_cql_skegr.pdf.
- Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.). *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: Benjamins, pp. 257-281.
- Sinclair, J. (ed.) (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.
- SSKJ: *Slovar slovenskega knjižnega jezika* (1991) Ljubljana: ZRC SAZU. Online version available at: <http://bos.zrc-sazu.si/sskj.html>.
- Tavčar, A., Fišer, D., Erjavec T. (2012). SloWCrowd: orodje za popraviljanje wordneta z izkoriščanjem moči množic. V: T. Erjavec in J. Žganec Gros (ur.) *Zbornik Osmo konference Jezikovne tehnologije. Proceedings of the Eighth Language Technologies Conference*. 8. do 12. oktober 2012 / October 8th - 12th, 2012 Ljubljana, Slovenija, pp. 197-202.

A lexicographic appraisal of an automatic approach for detecting new word-senses

Paul Cook,^{*} Jey Han Lau,^{*} Michael Rundell,[†] Diana McCarthy,^{*} and Timothy Baldwin^{*}

- ♣ Department of Computing and Information Systems, The University of Melbourne
 - ♦ Lexicography MasterClass and Macmillan Dictionaries
- ♠ Department of Theoretical and Applied Linguistics, University of Cambridge
Email: paulcook@unimelb.edu.au, jeyhan.lau@gmail.com,
michael.rundell@lexmasterclass.com, diana@dianamccarthy.co.uk, tb@ldwin.net

Abstract

Over the last 20 or so years, lexicographical tasks, such as finding collocations and selecting examples, have been automated to some degree, both supplementing lexicographers' intuitions with empirical data, and reducing the "drudgery" of lexicography to allow lexicographers to focus on tasks which cannot easily be automated. Automated determination of word senses and identification of usages of a given sense, however, have proven difficult due to their covert nature. In this paper, we present a method, based on an automatic word sense induction system, for identifying novel word senses in a more recent Focus Corpus with respect to an older Reference Corpus. We evaluate this method in the context of updating a dictionary, and find that it could be a useful lexicographical tool for identifying new senses, and also dictionary entries whose definitions or examples should be updated.

Keywords: computational lexicography, neologisms, word senses, word sense induction

1. Updating dictionaries

Lexicography is expensive. Despite the falling cost of corpus resources, the process of compiling and editing dictionary text remains labour-intensive. This applies not only to developing new resources from scratch, but also to the (more usual) job of updating existing dictionaries. One promising strategy for publishers is to automate some of the editorial tasks, and significant progress has been made in this area over the last ten years (Kilgarriff and Rychlý, 2010; Rundell and Kilgarriff, 2011; Rundell, 2012). In brief, corpus-analysis software can aid in: (1) determination of the syntactic, collocational, and text-type preferences of a given word or meaning; (2) selection of a shortlist of suitable example sentences; and (3) (at a later stage) streamlining of the process of editing and finalising dictionary text. The current approach to dictionary development has the software presenting data to the lexicographer in a useful predigested form. But recent advances offer the prospect of a model where "the software selects what it believes to be relevant data and actually populates the appropriate fields in the dictionary database" (Rundell and Kilgarriff, 2011, page 278), leaving the human expert to validate (or refine, or reject) decisions made by the computer.

The various components of this model have all been trialled on real dictionary projects, providing the conditions for incremental improvements in performance. The GDEX software, for instance, which automatically finds appropriate dictionary examples in a corpus, was used initially on a project at Macmillan, when there was a requirement for a large number of new example sentences for specific collocational pairs (Kilgarriff et al., 2008). The results were uneven but broadly positive, with the editorial team completing the task more quickly than if they had taken a purely “manual” route. Versions of GDEX have since been used in other ventures. The heuristics and weightings have been optimised for a number of languages (e.g., Kosem et al., 2011), and the software is now a standard feature in the editorial toolkit of a number of dictionary developers.

In other areas, progress towards automation has been slower. But the direction of travel is clear: we are gradually putting together a suite of robust applications which collectively streamline the job of compiling and editing dictionary text. If the effect of all this is to transfer some lexicographic tasks from humans to machines, the goal is to produce better dictionaries at a lower cost. A striking outcome of the work done so far in this area is that automation not only delivers efficiency savings but also leads to improvements in quality. Automating a process forces us to go back to first principles and be explicit about what the task involves. What, for instance, are the features of a “good” dictionary example, or at what point can we say with confidence that a particular syntactic pattern is “typical” of a word? All of which is contributing to the goal of producing dictionaries that are more systematic, more internally-consistent, and less reliant on the subjective judgment of individual lexicographers.

Improving the language-description process presupposes having some language that needs describing. Methodologies for extracting candidate headword lists from corpora are already well-established. Meanwhile, the requirement for tracking language change (more pressing than ever now that most dictionaries are online and their users expect them to be up-to-date) is also being addressed, and the task of identifying emerging new words is benefitting from computational approaches (Rundell and Kilgarriff, 2011, pages 263–267). But (notwithstanding the media’s obsession with shiny new headwords), there is more to updating a dictionary than adding neologisms. Two other salient aspects of keeping a dictionary up-to-date are finding novel senses of existing words, and ensuring that dictionary entries reflect contemporary conditions and technologies.

From the 1980s, as computer technology moved out of its specialist ghetto to become part of most people’s everyday experience, words like *mouse*, *icon*, *virus* and *window* acquired new senses. (The word *computer* itself, for that matter, began life in the 17th century as a job title for someone whose work involved calculation.) Earlier dictionaries do not include these meanings, so they had to be added. More recent examples include words like *cloud* and *tablet*, *hybrid* (a type of car), *sick* (used in contemporary slang as a term of approval), and *toxic* (when referring to financial

assets or debts). None of these meanings existed when the Macmillan Dictionary was first published (in print form) in 2002, and all have been added to the online edition (Macmillan English Dictionary Online, hereafter MEDO).¹ An equally important, but more elusive, goal is to ensure that definitions and examples reflect contemporary realities. In recent updates to MEDO, for example, changes have been made to the definitions of *meeting* (participants do not have to be in the same location), *marriage* (not just between a man and woman), and indeed *dictionary* (no longer simply “a book which ...”). MEDO has also targeted example sentences with dated contexts, like this one exemplifying one of the meanings of the verb *to slot*:

(1) *She slotted another tape into the cassette player.*

Traditionally, these are labour-intensive operations. In an ideal world, a well-funded editorial team would carefully review every entry, consulting contemporary corpus data, and identify anything that needed changing or updating. This is increasingly impracticable. Budget constraints weigh heavily on most non-commercial institutions, while commercial lexicography is in the process of replacing a simple and reliable business model (selling books) with something more complex and (for the time being) less profitable.

So, for the sake of both systematicity and feasibility within limited budgets, it makes sense to see how far we can automate the tasks of finding novel senses and identifying other areas of the text that might need updating.

In this paper, we examine a previously-proposed technique for automatically identifying word senses that are new to one corpus with respect to another (Lau et al., 2012), based on an automatic word sense induction system. We propose a further extension to that system which can incorporate human intuitions about topics for which we expect to see many new word-senses. We describe our previous evaluations of the core system, and its ability to identify new word-senses. We then present a new evaluation of our proposed method in the context of updating a dictionary, in collaboration with a professional lexicographer (the third-named author of this paper). Our findings suggest that this method could indeed be a useful new addition to the lexicographer’s toolkit.

2. Automatic novel sense detection

Word sense induction (WSI) is the task of automatically grouping the usages of a given word in a corpus according to sense, such that all usages exhibiting a particular sense are in the same group, and each group includes usages corresponding to only one sense (Navigli, 2009). The category “word sense” is not of course uncontroversial. There is no general agreement about what constitutes a discrete

¹ <http://www.macmillandictionary.com/>

meaning of a word, and dictionaries often exhibit considerable variation in their treatment of the same polysemous word. But although word meanings are unstable entities, often with shifting boundaries, dictionary conventions traditionally require that lemmas are divided up into numbered senses, and a good lexicographers' style guide will provide criteria for doing this.² Here, we describe a WSI technique we developed and its application to the task of identifying novel word senses.

The WSI methodology we use is based on a model we previously proposed (Lau et al., 2012). The core machinery of this method is driven by probabilistic topic models (Latent Dirichlet Allocation, LDA: Blei et al., 2003), where latent or unseen topics are viewed as the driving force for generating the words in text documents. In this model a document is viewed as a probability distribution over topics, and each topic is represented as a probability distribution over words. The probability distributions for documents and topics are automatically “learned” from the corpus. Crucially, the “topics” in a topic model do not necessarily correspond to topics in the sense of the subject of a text. Applying topic models to induce the word senses of a lemma of interest, these “topics” are interpreted as the induced senses.

In traditional topic models, the number of topics to be learnt is a parameter that must be set manually in advance. In WSI, this parameter translates to the number of senses to be induced for a lemma. To develop a model without this requirement, and which can learn varying numbers of senses for different lemmas as appropriate, we used a Hierarchical Dirichlet Process (HDP, Teh et al., 2006), a variant of LDA that also learns an appropriate number of topics/senses.

Following our previous work, for each usage of a target lemma we extract a three-sentence context, where the second sentence contains the usage of the lemma, and the first and third sentences are the preceding and succeeding sentences, respectively. These three-sentence snippets are viewed as the “documents” in the topic model. We represent each document as the bag-of-words it contains, as is common for topic models.³ We also include additional positional word information to represent the local context of the target lemma. Specifically, we introduce an additional word feature for each of the three words to the left and right of the target lemma. An example of the features is given in Table 1. To illustrate the senses induced by our model and the usages that correspond to the senses, we present Tables 2 and 3 respectively, for the example lemma *cheat*.

² For a full discussion of word senses, see Hanks (2013, pages 65–83).

³ We use the term bag-of-words to refer to the multiset of items occurring in some context, as it is commonly used in natural language processing. As described in Sections 3 and 4.1, we lemmatise our corpora. Our “bag-of-words” representation is therefore in fact a bag-of-lemmas.

Target lemma	dog
Context sentence	Most breeds of dogs are at most a few hundred years old
Bag-of-word features	most, breed, of, be, at, most, a, few, hundred, year, old
Positional word features	most_#-3, breed_#-2, of_#-1, be_#+1, at_#+2, most_#+3

Table 1: An example of the topic model features.

Sense Number	Top-10 Terms
1	heat think want ... love feel tell guy include find
2	cheat student cheating test game school to teacher exam study
3	husband wife cheat wife_#1 tiger husband_#-1 on ... woman marriage
4	cheat woman relationship cheating partner reason man woman_#-1 to spouse
5	cheat game play player cheating poker to card cheated money
6	cheat exchange china chinese foreign cheat_#-2 cheat_#2 china_#-1 to team
7	tina bette kirk walk accuse mon pok symkyn nick star
8	fat jones ashley pen body taste weight expectation parent able
9	euro goal luck fair france irish single 2000 point complain

Table 2: The top 10 terms for each of the senses induced for the lemma cheat.

Sense number	Usage
4	<p>While I was single I slept with several married men. I had relationship with them. Now that I am married I feel horrible for having done so. I am always afraid my husband is going to <u>cheat</u> on me.</p> <p>It appears to me that there are people who are just disloyal. A man who <u>cheats</u> on his wife will <u>cheat</u> other partners whether that partner is a business partner or a wife – disloyalty transfer.</p> <p>I find it ignorant when men <u>cheat</u> on their wife, and when they found out the wife was sleeping around, they get mad. That makes no sense.</p>
5	<p>Lastly, the foremost argument in my personal opinion is that the profit margin of the online poker room is so large, that they simply would not need to <u>cheat</u> their own players. They are practically doing it already. Fairly.</p> <p>Do you feel you have been <u>cheated</u> when playing online poker? Well, guess what. You have been! The question is: do you want to continue being <u>cheated</u>?</p> <p>“There is not a card player who would not <u>cheat</u>, if he knows how.” - Walter Irving Scott, the phantom of the card table.</p>

Table 3: Corresponding usages for induced senses 4 and 5 of the lemma cheat.

To identify novel senses, we compare a Focus Corpus with a Reference Corpus. In the application we consider here (updating a dictionary), the Focus Corpus would consist of newer texts; the Reference Corpus, on the other hand, would be older material, and common usages in this corpus would be expected to be reflected in the dictionary. (Details of the Reference and Focus Corpora used in this study are given in Section 4.1.) We combine the Focus and Reference Corpora to produce a supercorpus. For a given lemma of interest we then apply our WSI methodology to all of its usages in this supercorpus. (In this study we consider all lemmas meeting some frequency and keywordness cutoffs, also described in Section 4.1.) The WSI step automatically labels each usage of the lemma with its induced sense. We then calculate the “novelty” of an induced sense in the Focus Corpus as the ratio of its relative frequency in the Focus and Reference Corpora, akin to a simple approach to keywords (Kilgarriff, 2009), but applied to induced senses. We rank the lemmas according to the novelty of their highest-scoring induced sense. The highest-scoring induced sense for a given lemma is referred to as its novel sense.

New senses often arise for prominent cultural concepts (Ayto, 2006). In this paper, we introduce a new variant to our method for identifying novel senses that incorporates this observation. We first manually form a list of terms related to a particular topic (computing and the Internet for the analysis presented in Sections 4 and 5). For each induced sense we then determine its relevance to this topic based on its probability distribution over words from the topic modeller. We independently rank each induced sense by its relevance and its novelty score, and then rank each induced sense by the sum of its rank under each of these two rankings. This approach identifies induced senses which are both novel and related to a particular topic, and is referred to as “rank sum”.

3. Previous evaluation

In this section we describe previously-presented evaluations of the WSI component of our method on several benchmarked WSI tasks, and an evaluation of the accuracy of our method for detecting whether a given word exhibits a novel sense in a more recent Focus Corpus compared to an older Reference Corpus and, furthermore, whether it can detect specific instances of a novel sense within the Focus Corpus. In Sections 4 and 5 we present a new evaluation of our method for identifying novel senses in the context of updating a dictionary.

Our WSI technique was first presented in Lau et al. (2012), and was initially evaluated using two datasets (Agirre and Soroa, 2007; Manandhar et al., 2010) to compare the system to the state-of-the-art in WSI. These datasets were produced within the auspices of a series of international events (SemEval, formerly SENSEVAL) for the objective comparison of computational systems that provide semantic analysis. Both datasets require the systems to induce senses for a sample of lemmas from some

training data and then label some unseen data with these senses. From the evaluation, our system outperformed the state-of-the-art systems, given the same conditions for tuning parameters. Moreover, on the more recent 2010 dataset our model, which uses HDP to automatically learn the optimal number of topics (senses), outperformed a more basic LDA model even when the latter was manually told how many topics to learn.

More recently we evaluated our WSI technique by participating in two SemEval 2013 WSI tasks. “Word Sense Induction for Graded and Non-Graded Senses” (Jurgens and Klapaftis, 2013) was similar to the previous WSI evaluations considered, but additionally required systems to identify not just the single most appropriate induced sense for a given test usage, but rather all applicable senses, and the extent to which they apply. In this evaluation a number of different metrics were considered, with our method outperforming all other participating systems in terms of one metric, and achieving strong results overall (Lau et al., 2013a). “Evaluating Word Sense Induction & Disambiguation within an End-User Application” (Navigli and Vannella, 2013) considered whether WSI can be applied to diversify search engine results. In this task our system performed best out of all participating systems, further demonstrating the effectiveness of our WSI approach (Lau et al., 2013b).

To evaluate the application of our WSI method for novel sense detection, our earlier work (Lau et al., 2012) provided the first, and to date only, available dataset, albeit a relatively small one. The production of such a dataset is difficult because word senses are covert and manually labelling occurrences in a corpus is a very time-consuming and laborious process. We focused on a small sample of lemmas which were identified as having senses arising in the period between the early nineties and 2007. This period was selected simply because of the availability of a Reference Corpus, the British National Corpus (BNC, Burnard, 1995), and a more recent Focus Corpus, the ukWaC (Ferraresi et al., 2008), produced automatically from data from the Web in 2007.⁴ Since these corpora are of different sizes, they were made more comparable by using only the written portion of the BNC and extracting a similar-sized random sample of documents from the ukWaC and using TreeTagger (Schmid, 1994) to tokenise and lemmatise both corpora.

We used the Concise Oxford English Dictionary editions which best reflected contemporary usage for the two respective time periods: Thompson (1995, COD95) and Soanes and Stevenson (2008, COD08). Working on the assumption that new senses often arise for culturally salient concepts (Ayto, 2006), we directed our search towards entries relevant to computing and with sufficient frequency (more than 1000) in the BNC. The lexical selection was supported with a manual inspection of 100 random occurrences from the respective corpora and also a manual inspection of the

⁴ Note that the new evaluation presented in this paper uses different Reference and Focus Corpora than our earlier work.

collocates of the candidate lexemes using word sketches (Kilgarriff and Tugwell, 2002).⁵

The above procedure yielded five genuine lemmas with a novel sense arising in the respective period.⁶ We then selected five distractor lemmas with the same part of speech as a target and of similar frequency within the BNC, but where there was no evidence of a new sense given the respective entries in COD95 and COD08. The automatic WSI method was applied to the similarly-sized set of the documents from the BNC and the ukWaC and the output used for ranking the lexical items by their novelty score. The lemmas with a high novelty score had significantly higher ranks compared to the distractors; meanwhile, a baseline which only considered the frequency difference across the two corpora did not produce a significant difference in ranking. We additionally used the manually tagged samples to demonstrate that not only could the approach successfully rank lemmas on the basis of novelty, but also it could be used to identify the novel occurrences in the Focus Corpus. Promising results were obtained overall simply by identifying the specific novel sense with the topic that was automatically ranked highest for novelty and using that to identify occurrences. Furthermore, because the induced senses are modelled as lists of salient words, topic models afford a readily interpretable representation for word sense, highlighting the potential for such automatic methods to produce output that can inform the lexicographic process.

4. Lexicographical evaluation

In this section we describe an evaluation of our proposed method for identifying novel word senses in the context of updating a dictionary, based on manual analysis by a lexicographer.

4.1 Corpora and pre-processing

Our previous evaluation of the ability of our WSI method to identify novel senses (presented in Section 3) used the BNC and ukWaC, corpora which consist of very different genres. For this analysis we consider more-comparable corpora. We use the English Gigaword Fourth Edition (Parker et al., 2009), henceforth referred to as GIGAWORD, which consists of newswire articles from six services including the New York Times Newswire Service; the Los Angeles Times/Washington Post Newswire Service; and the Agence France-Press, English Service for the years 1994–2008.⁷ For our Reference and Focus Corpora we use the sub-corpora of Gigaword for the years 1995 and 2008, respectively, the earliest and latest years in the corpus for

⁵ <http://www.sketchengine.co.uk/>

⁶ The five lemmas were *domain* (n), *export* (v), *mirror* (n), *worm* (n), and *poster* (n).

⁷ There is a fifth edition of this corpus which additionally includes data for 2009 and 2010, but we unfortunately do not have a license for this edition of the corpus.

which data from all services are available. This provides Reference and Focus Corpora which are comparable, in that they both consist of newswire data from the same sources for a given year, although there are of course topical differences between the corpora for the two years. Moreover, these corpora are diverse, consisting of data from six sources, although all data are from newswires.

Gigaword consists of several document types with the by far most frequent being “story”, which corresponds to a typical newswire story. We only consider these documents. Gigaword is known to contain a substantial number of Spanish documents. To reduce the amount of non-English content in our corpora, we filter all documents not identified as English using `langid.py` (Lui and Baldwin, 2012), a statistical language identification tool. Newswire text contains duplicate and near-duplicate documents, corresponding to, for example, an update to a previous story. We apply exact deduplication, and near-deduplication using `Onion` (Pomikalek, 2011), to remove such documents. Finally, we part-of-speech tag and lemmatise the resulting corpora with `TreeTagger` (Schmid, 1994), in line with our earlier experiments over the BNC and ukWaC.

The Reference (1995) and Focus (2008) Corpora consist of 193M and 202M words, and 471k and 536k documents, respectively. We count the words in each corpus, and compute keywords using the method recommended by Kilgarriff (2009). We identify all nouns with frequency greater than 1000 in each corpus, frequency less than that of the 100th most-frequent noun in each corpus, and keywordness between 0.5 and 2. This gives 3185 nouns over which we run our proposed method for identifying novel word senses.

For the “relevance” component of the rank sum method for identifying novel word-senses we manually identify words related to computing and the Internet, topics that increased in prominence between the time periods of our Reference and Focus Corpora. We compute keywords for our Focus Corpus relative to our Reference Corpus, again using the method of Kilgarriff (2009). This method includes a parameter, α , which roughly controls the frequency range of the resulting keywords. We identify the top-1000 lower-case keywords with length at least three for α set to 1, 10, and 100 to consider keywords with a range of frequencies. The first and second authors of this paper independently annotated the keyword list to identify those that they judged to be primarily related to computing and the Internet in the newswire domain. Thirty-three keywords were selected by both annotators, and these words were used as the domain-specific words in computing relevance.

4.2 Lemma selection

We ran our method for identifying novel word senses on all 3185 nouns matching our frequency and keywordness criteria from the previous subsection. We considered both the novelty and rank sum approaches. The top-10 items for each method were selected for further analysis.

It is possible that our proposed method fails to identify many new word-senses, i.e., that amongst the lemmas not identified by our system there are many new senses. In an effort to evaluate this we also analysed ten randomly selected lemmas. The thirty lemmas analysed are shown in Table 4.

Novelty	Rank Sum	Random
airstrikes	advertiser	arena
candy	cell	audit
cleric	click	beauty
junta	copyright	follow-up
militiaman	fingerprint	fraction
nutrition	instinct	likelihood
plastic	search	lyric
prostitution	text	stockpile
truce	video	taxis
vest	web	tension

Table 4: The 30 lemmas selected for analysis and the method through which they were selected (presented in alphabetical order in each column). For items shown in **bold-face** the analysis revealed a noteworthy change in usage.

4.3 Analysis process

For each lemma we produced a summary consisting of the following information:

- The words associated with the topic corresponding to the candidate novel sense (provided by the topic modeller);
- The ten highest confidence novel sense usages from each corpus;
- The number and proportion of usages corresponding to the novel sense in each corpus;
- A random sample of ten usages from each corpus.

These summaries were then given to a professional lexicographer to analyse. Crucially, the lexicographer (the third-named author of this paper) did not know whether a given lemma was included because it scored highly for the novelty or rank sum method, or because it was one of the randomly selected items. The analysis was carried out with respect to the following questions.

- Would the candidate novel sense be included in various types of dictionaries (e.g. a general pedagogical dictionary, a large “native-speaker” dictionary, an online dictionary)?
- Has the candidate novel sense already been included in dictionaries, but only in those for specialised domains?
- Is the candidate novel sense interesting for some other reason?

Throughout the analysis two “reference” dictionaries were consulted:

MEDO Macmillan English Dictionary Online: a medium-sized, monolingual, mainly pedagogical dictionary with approximately 50,000 headwords;⁸

ODE Oxford Dictionary of English: a standard monolingual “desktop” dictionary aimed at native speakers with about 80,000–90,000 headwords.⁹

Two other dictionaries aimed at a similar market to MEDO were also sometimes referred to: the Cambridge Advanced Learners Dictionary (CALD),¹⁰ and the Longman Dictionary of Contemporary English (LDOCE).¹¹

5. Analysis

Table 4 shows the lemmas analysed, displaying which were found to have a notable difference in usage between the Reference and Focus Corpora. Overall there are more “interesting” findings for the lemmas obtained through novelty than the randomly selected lemmas. Moreover, for the rank sum method, all lemmas correspond to an interesting difference in usage in the Focus Corpus. This suggests that our proposed method could be a useful tool for identifying changes in usage.

5.1 Uninteresting findings

For the lemmas not shown in boldface in Table 4, a notable difference in usage was not observed between the Reference and Focus Corpora. In all of these cases the data provide no evidence of a novel sense in the Focus Corpus, and the sense instantiated in the data is adequately covered in the two “reference” dictionaries considered, and in other general dictionaries. For the “random” lemmas this is not surprising, and we will not discuss them further here.

In the case of each item in this category identified by novelty (i.e., *airstrikes*, *candy*, *junta*, *plastic*, *prostitution*) there are marked contextual differences between the new and old corpora, and random and selected sets of usages. Here the proposed method has identified a novel configuration of frequent collocates — a sudden spike which typically reflects a (briefly) salient news story. Thus at *junta*, the collocates list (including *myanmar*, *aid*, *cyclone*, *relief*) relate to a cyclone which hit Myanmar/Burma in 2008,¹² causing huge loss of life. Similarly, the data for *candy* in the Focus Corpus are skewed by a news story about Chinese candy being contaminated by melamine. What tends to happen in these cases is that the other data

⁸ <http://www.macmillandictionary.com/>

⁹ <http://oxforddictionaries.com/>

¹⁰ <http://dictionary.cambridge.org/>

¹¹ <http://www.ldoceonline.com/>

¹² http://en.wikipedia.org/wiki/Cyclone_Nargis

(selected data from the Reference Corpus and all random data) exhibit the same sense but a wider range of contexts. Topical differences are known to be challenging for methods for identifying differences in word sense between corpora (Peirsman et al., 2010), and indeed similar observations in our earlier work led to the development of the rank sum method to address this. That none of the top-10 lemmas for the rank sum approach are in this category suggests that it has been successful in this regard.

5.2 Dictionary account needs tweaking

In the following cases, the data provide evidence which suggests that some existing dictionary accounts (sometimes in MEDO, sometimes in the other dictionaries referred to in Section 4.3) may need to be tweaked or broadened. Most of these cases, however, do not indicate the emergence of a genuine new word-sense.

advertiser Examples from the Focus Corpus refer overwhelmingly to web advertising—but many of those from the Reference Corpus do too. Web advertising was already established in 1995, and MEDO’s entry reflects this (though that is not the case in some other dictionaries). Several of the corpus examples for *advertiser* include references to *publishers*, and many dictionaries are still lagging in their definitions of what “publishing” entails (typically focussing on the traditional media of books, music, journals, and the like). So the co-occurrence of *advertiser* and *publisher* in the data serves as a useful reminder that one or both of these entries may need updating to reflect the words’ contemporary use.

cell In both the Focus and Reference Corpora, the examples refer to *cell phones* (the usual term in American English, though not in British English, where *mobile (phone)* is preferred). All the dictionaries examined record this use of *cell*. However, the Focus Corpus includes at least two references to *cell sites*, and this appears to be a valid term, defined in Wikipedia as: “a cellular telephone site where antennas and electronic communications equipment are placed”. *Cell site* does not appear in any general English dictionary, but it is at least worth considering whether it should.

cleric The data from the Focus Corpus overwhelmingly refer to *Muslim clerics* (who are typically characterized as *radical* and/or *fundamentalist*), and this marks a clear shift from what we find in the Reference Corpus, where *cleric* tends to suggest an innocuous Church of England figure of the type found in a Trollope novel. Although the entries in the two “reference dictionaries” both take account of this change, the definitions and/or example sentences in some dictionaries do not: LDOCE, for example, defines *cleric* simply as “a member of the clergy”.

copyright The Focus Corpus data often mention *copyright* in the context of new media (games or software, for example), whereas older data refer to more traditional contexts (songs, books etc.). There is no change in the essential meaning (“protection of ones’ intellectual property”), but some dictionaries may need to update definitions and/or example sentences in order to account for the broader scope of this term.

militiaman The data suggest that some updating is required at dictionary entries for *militia*. (*Militiaman* itself is adequately defined as “a member of a militia”.) The Focus Corpus contexts point to the now dominant use of *militia* to refer to an unofficial

armed group, typically with links to terrorism or insurgency (*Shiite militias*, etc.). The current definition in MEDO (“a group of ordinary people who are trained as soldiers to fight in an emergency”) invokes an older, more neutral use, referring to a citizen army, and most other dictionaries have the same emphasis.

truce The selected examples (Focus and Reference) reflect the standard use of *truce* (and the contexts – mostly to do with Palestine and Israel – show depressingly little change over the period). But the randomly selected usages include at least two cases where the context is not war, but business or politics. This may indicate a separate sense: more a cessation of argument or opposition than of fighting and hostilities. The current definition in MEDO could be said to cover all these scenarios: “an agreement between two people or groups involved in a war, fight, or disagreement to stop it for a period of time”. But the example sentences all refer to war-type contexts and ODE’s entry has a similar focus.

vest All the Focus Corpus examples (but only one or two from the Reference Corpus) refer to “suicide vests” or “explosive vests” – evidently a salient context in contemporary texts. The closest sense in MEDO defines a vest as “a piece of clothing with no sleeves or collar worn over other clothes, for example for protection”, and follows with an example: *a bulletproof vest*. This does not fully reflect current usage, so the entry may need tweaking.

video The Reference and Focus Corpora have very different emphases, with the newer data referring exclusively to online videos (with collocates such as *circulate*), whereas the older data refer to movies or TV programmes stored on VHS devices (the prevailing technology in the 1990s). Here, the entire MEDO entry is out of date (it refers to material “recorded on videotape”) and had in fact already been flagged for attention in the next update. ODE has already updated its entry to take account of changing technologies, and its definition reads: “a recording of moving visual images made digitally or on videotape”. This is not a novel sense as such, but the dictionary record definitely needs updating.

web In the sense of “the Web”, this is a fairly recent but by no means novel meaning. One interesting point is that the Reference Corpus data include several citations for the expression *world wide web*, which is now very dated. Most dictionaries have a neutral entry for this term, and in many (including ODE and MEDO), definitions of *web* or *the Web* simply say “the World Wide Web”, cross-referring to another entry. In 2013, this is the wrong way around – rather like defining *bus* as “an omnibus” (as would have happened in dictionaries 100 years ago). So here again the data serve as a useful reminder to make adjustments to an entry which could easily have been ignored.

Two of the “random” lemmas – *follow-up* and *fraction* – were also assigned to this category when the data were analysed from a lexicographic viewpoint. It is not so surprising that a randomly chosen lemma would appear in different contexts, given the different dates of the two corpora. Since we have already established that the automated method tends to find more noteworthy cases than random ones, we do not discuss the “random” lemmas further here.

5.3 Novel senses

For these lemmas, the data indicate a genuine novel sense.

click The use of *click* meaning “an instance of a user clicking on something” was already established in 1995. MEDO includes this meaning, with the example: *You can order anything with a single click*. However, examples like the following suggest a newer use:

(2) *Total paid clicks in the fourth quarter rose 30 percent from the same 2006 period.*

(3) *For instance, comScore estimated Google’s fourth-quarter clicks increased 25 percent.*

This reflects the Web business model, where each click on an advertising link represents a specific value for the publisher. The current MEDO entry does not adequately cover this newer use, which (though more specialised) is nevertheless valid.

fingerprint Several examples from the Focus Corpus data refer to *digital fingerprint* (which is not found in the Reference Corpus data). Most likely this simply refers to a digital record of a fingerprint. But the term *digital fingerprint* is also used in data security contexts with a different meaning. This second meaning does not appear in any of the four general dictionaries we consulted (see Section 4.3). But it is recorded in the more specialised *businessdictionary.com*, where it is defined as: “Coded string of binary digits (generated by a mathematical algorithm) that uniquely identifies a data file”. This is followed by the more familiar second sense: “Analog fingerprint of a person converted (digitised) into a binary file”. There may be a case for a similar two-sense entry in general dictionaries.

search The Focus Corpus provides evidence for a novel sense of *search*, and this is absent from the Reference Corpus. The novel sense refers to the business of search (on the Web), and is an uncountable noun (distinct from “doing a Google search for something”). This use was added to MEDO in an update carried out in early 2013, as follows: “3 [uncountable] the process of searching for information on the Internet, or the business and technology that supports this”: *Founded in 1995, Yahoo was quick to get into search*. This use is not currently accounted for in most dictionaries.

text All the data from the Focus Corpus relate to *text messaging*, which was still rare in 1995 and does not appear in the data from the Reference Corpus. (The BNC has no examples of *text messaging* either.) There has clearly been a huge shift in the frequency profile of the word *text* over this period. The proposed automated method has successfully identified this newer usage, though in this case it is something that all the checked dictionaries take account of.

One of the “random” lemmas, *audit*, was also found to exhibit a genuine novel sense in the data considered here. The Focus Corpus usages refer mainly to the contexts of aviation and slaughterhouses, and indicate an inspection aimed at ensuring safety and compliance with regulations. This appears to be fairly recent (the Reference Corpus data – both random and selected – focus on the older “financial audit” sense, the work done by *auditors*). What we see here is probably a fairly recent sense, though there is some evidence for it in the (1992) BNC, e.g. for *environmental audit* (43 hits), and most dictionaries already cover it.

5.4 Other cases

instinct The data from the Focus Corpus relate to a smartphone with this proprietary name released in 2008 (hence collocates like *iphone* and *samsung*), and the word always appears with initial uppercase (*Instinct*). This usage would not typically be recorded in the dictionaries consulted in this analysis, and it could potentially be identified by more simple means (such as a keyword analysis in which case is preserved). However, information about case is not available to the automated method, and so from this limited perspective, the system has successfully identified that *instinct* has a new usage in the Focus Corpus.

nutrition All selected examples of *nutrition* in the Focus Corpus are of the following type:

- (4) *NUTRITION Per serving (based on 8): 179 calories, 2 g protein, 42 g carbohydrates, 1 g fat, 0 g saturated fat, 0 mg cholesterol, 67 mg sodium, 5 g dietary fiber*

This relates to a standard format for nutritional information on food labelling. Although this cannot be considered a novel sense, it is a usage which is far more common in the Focus Corpus than in the Reference Corpus, and the proposed method has identified it.

6. Conclusions

We presented an automatic method for identifying new word-senses in a Focus Corpus of more recent texts with respect to an older Reference Corpus. An evaluation of our method in the context of updating a dictionary suggests that this method has promise as a tool for helping lexicographers to identify new word-senses. Moreover, this method was shown to have the potential to aid in identifying dictionary entries that require updating, for example, because definitions or example sentences are out of date. Crucially, although these tasks are important for keeping dictionaries current, they are also very expensive, and there have been few previous efforts to automate them.

At the heart of our proposed method is a word sense induction system, which groups together similar usages of a given word in a corpus. In future work we intend to consider whether this system can be applied to other dictionary writing tasks, for example, identifying good dictionary examples for a particular word sense, or semi-automatic dictionary drafting (Kilgarriff and Rychlý, 2010).

To encourage further research on topic modelling approaches (such as the one used by our system) in computational lexicography, and the use of our proposed method in lexicographical projects, we have made our word sense induction system publicly available under a license which permits its use for commercial purposes.¹³

¹³ <https://github.com/jhlau/hdp-wsi>

7. References

- Agirre, E. and Soroa, A. (2007). SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic.
- Ayto, J. (2006). *Movers and Shakers: A Chronology of Words that Shaped our Age*. Oxford University Press, Oxford, UK.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Burnard, L. (1995). *User Guide for the British National Corpus*. Oxford University Computing Service, Oxford, UK.
- Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, pages 47–54, Marrakech, Morocco.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, USA.
- Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, USA.
- Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress (EURALEX 2008)*, Barcelona, Spain.
- Kilgarriff, A. and Rychlý, P. (2010). Semi-automatic dictionary drafting. In de Schryver, G.-M., editor, *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*, pages 299–312. Menha Publishers, Kampala, Uganda.
- Kilgarriff, A. and Tugwell, D. (2002). Sketching words. In Corréard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125–137. Euralex, Grenoble, France.
- Kosem, I., Husak, M., and McCarthy, D. (2011). GDEX for Slovene. In *Proceedings of eLex 2011*, pages 151–159, Bled, Slovenia.
- Lau, J. H., Cook, P., and Baldwin, T. (2013a). unimelb: Topic modelling-based word sense induction. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311, Atlanta, USA.
- Lau, J. H., Cook, P., and Baldwin, T. (2013b). unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the*

- Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 217–221, Atlanta, USA.
- Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.
- Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.
- Manandhar, S., Klapaftis, I., Dligach, D., and Pradhan, S. (2010). SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.
- Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).
- Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.
- Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). *English Gigaword Fourth Edition*. Linguistic Data Consortium, Philadelphia, USA.
- Peirsman, Y., Geeraerts, D., and Speelman, D. (2010). The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(4):469–491.
- Pomikálek, J. (2011). *Removing Boilerplate and Duplicate Content from Web Corpora*. PhD thesis, Masaryk University.
- Rundell, M. (2012). The road to automated lexicography: an editor’s viewpoint. In Granger, S. and Paquot, M., editors, *Electronic Lexicography*, pages 15–30. Oxford University Press, Oxford, UK.
- Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In Meunier, F., Cock, S. D., Gilquin, G., and Paquot, M., editors, *A Taste for Corpora. In honour of Sylviane Granger*, pages 257–282. John Benjamins, Amsterdam, Netherlands.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.
- Soanes, C. and Stevenson, A., editors (2008). *The Concise Oxford English Dictionary*. Oxford University Press, eleventh (revised) edition. Oxford Reference Online.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.
- Thompson, D., editor (1995). *The Concise Oxford Dictionary of Current English*. Oxford University Press, Oxford, UK, ninth edition.

Augmenting online dictionary entries with corpus data for Search Engine Optimisation

**Holger Hvelplund,¹ Adam Kilgarriff,²
Vincent Lannoy,¹ Patrick White³**

¹IDM, Paris, France

²Lexical Computing Ltd., Brighton, England

³Oxford University Press

E-mail: hvelplund@idm.fr, adam@lexmasterclass.com,
lannoy@idm.fr, patrick.white@oup.com

Abstract

Search Engine Optimisation is a challenge for dictionary publishers. As soon as a dictionary appears online, one part of its success will be measured by its web traffic. Central to the volume of web traffic is where it appears on search engine results pages when a user searches for a word. There are many strategies for improving search engine rankings: the one explored here is automatically augmenting dictionary entries with corpus-derived collocates and related words, as identified by the Sketch Engine's word sketches and distributional thesaurus. We took the online version of the Oxford Advanced Learner's Dictionary and augmented a set of entries, to find whether they then saw an increase in web traffic. They did.

Keywords: corpus, collocation, SEO, Search Engine Optimisation, online dictionary

1. Introduction

A challenge faced by online dictionaries with no parallels in paper dictionaries is Search Engine Optimisation (SEO): coming top (or somewhere near the top) of search engine listings when a user 'Googles' (e.g., searches in a search engine) for a word. SEO is a new art form of great importance to any enterprise using the Web. For an online dictionary to reach a large audience, it must effectively do its SEO.

Lannoy (2010) demonstrates how a resource such as WordNet can support SEO by contributing relevant, hyperlinked text to online dictionary entries. This paper develops that work in two ways: first, by using collocations and related words discovered through a state-of-the-art corpus query system to augment entries; and secondly, through an experiment on the online version of a leading, branded dictionary, where we test the hypothesis that the additions really do bring more traffic to the website.

The dictionary in question is the Oxford Advanced Learner's Dictionary (<http://oald8.oxfordlearnersdictionaries.com>).

2. Corpus data

The corpus methods used were ‘word sketches’ and a distributional thesaurus as generated (for a large number of languages, though in this case English) within the Sketch Engine corpus query tool (Kilgarriff et al., 2004, <http://www.sketchengine.co.uk>). Word sketches are one-page summaries of a word's grammatical and collocational behaviour. They have been used by lexicographers since 1998. A distributional thesaurus shows, for the target word, the words that share most collocates with it, in the sense that *tea* and *coffee* both ‘share’ the collocate *drink* (in the grammatical relation “object of”).

For each word, the dictionary entry can be augmented with the collocates¹ from the word's word sketch, and the ‘related words’ from its thesaurus entry.

This information is valuable both to the dictionary user, since it tells them more about the usage of the word, and for SEO.

3. Benefits for SEO

All else being equal, pages with more text and more links are preferred by search engines, in the sense that search engine robots have more material to crawl. However, the text and links must be relevant: the search engines go to great lengths to counteract the efforts of spammers to put spam pages at the top of search results and have sophisticated algorithms for identifying junk text and junk links. As the collocates and related words are specific to the headword, and are relevant for the user, we believe they are, and will remain, acceptable to the search engines.

Each collocate and related word can be made into a link to its entry in the dictionary. This is useful to the user, as they can then click to see the entry for that word, and also beneficial for SEO. The links, to other pages on the dictionary's website, will be site-internal: site-internal links have lower weighting, within the search engines' ranking algorithms, than incoming links from external sources, but they do still carry weight.

4. Infrastructure

OALD online is managed by IDM, in DPS4. IDM created a local installation of the Sketch Engine and set up IDM DPS Processing script to use the Sketch Engine API to gather collocates and related words from the Sketch Engine. To allow flexible re-use in one or several dictionaries the script saves auto-generated content entries in a DPS

¹ In our terminology, a *collocation* comprises the node word and the *collocate*, standing in a specific grammatical relation to each other. Thus the words from the word sketch which are added to the node word's entry are its *collocates*.

project. The DPS process responsible for delivery of content for the online dictionary adapts and merges the new data into the manually produced and editorially checked OALD entries.

For fine-tuning and adapting the auto-generated content to editorial requirements, the method described here has proven to be flexible and extensible.

5. Experiment

To run the experiment, it was necessary to answer the following questions:

1. Which entries would we augment?
2. Which collocates and related words would we add, and how many of them?
3. How would we present the new information to the user?
4. How would we measure results of the experiment?

Throughout, it was essential to pay heed to the OUP brand: OUP is authoritative, and does not make mistakes or present nonsensical material.

5.1 Which headwords?

The headwords we used for the experiment were a random sample of 231 low-frequency words, presented below:

abalone abjure abstruse adroit aerobatics aggrandizement agoraphobia ague amanuensis
 ammonite antonym apostate apprise arachnid arrears askance askew auburn aura
 autoimmune avocation azure backgammon ballpoint barbell bargaining barista bashful
 beanie berserk besotted bespoke beta betrothal bidet bigamy bitumen bling blinker
 bonkers bonsai booger brainiac brainwave burlesque calumny cardamom cashew
 centigrade centipede cephalopod ceramic chamois charged chicanery chiropodist chirpy
 chivalrous cliffhanger clunk colander concatenation consonance contextualize cordially
 countable covetous credulous curtsy decision-making denotation diphthong dirge
 disestablish doldrums doodle dork douche downtime dumpling dystopia edification
 effrontery egress emoticon enamoured esophagus extrovert fascia feces fellatio fricative
 frostbite futon gerund get-together geyser glutton google gruel guava hale highbrow
 hold-up homonym homophone hovercraft hypotenuse iconoclast igloo incensed inchoate
 incorrigible infatuated ingenuous ingress interjection intransitive introvert iterate
 jingoism khaki kin lackadaisical laminate languor lassitude legit leitmotif levity lexis
 liquorice located loquacious lychee lye mankind marsupial masseuse media meerkat
 merry-go-round mezzanine mnemonic mocha muffler mugging mutton myrrh naught
 neigh newbie niqab obdurate obeisance obliging obsequious occult okra omnivore
 ostentation panoply parallelogram paramour paroxysm peeve peevish perdition perfidy
 pestle phishing plasma platinum pre-empt prevaricate proboscis prosody prude psychotic
 puerile pugnacious quietude quintessence recon retrograde ruckus satiate satiety scissors

scotch segue sepulchre smartphone snazzy snitch snorkel snowdrift sorority spendthrift
 stapler stole sty sudoku sunglasses suntan supercilious sycophant synecdoche taciturn
 tarmac tautology thither thyroid tidings tights trendsetter triage troubleshoot truant
 turmeric typhoid uncountable unflappable verbose vexation wallflower well-being
 wizened wrestling wrought xylophone

5.2 Which collocates and related words?

The items to add were the highest-scoring collocates from the word sketch and the highest-scoring related words from the distributional thesaurus. The score, for both collocates and related words, was the standard measure in use in the Sketch Engine.² Ensuring the quality of these items involved a number of iterations and checks.

Initially we used the UKWaC corpus (Baroni et al., 2012), comprising 1.3 billion words. However, for many of the low-frequency headwords in our sample there was not enough data: a collocate based on less than five hits is not trustworthy, and many of the words did not have collocates meeting that threshold. Therefore, we switched to enTenTen12 (Jakubicek et al., 2013), with 11.2 billion words.

In the Sketch Engine, each collocation has three parts: the headword, the collocate and the grammatical relation holding between them (e.g., *object*, *modifier*). After some discussion we decided to include the grammatical relation as well as the collocate in the augmented entry. We also removed duplicates where the same collocate occurred with more than one grammatical relation. (These cases were sometimes linguistically valid, for example *brush*, at headword *hair*, can be both the verb that the headword is object of (“she brushed her hair”) and a modified noun (“the hair brush”); however, the duplicates were often the outcome of part-of-speech tagging errors, and in any case, the duplication would not be helpful for the dictionary user.)

We considered it important not to overload the user with excessive information. We, therefore, set a limit of 20 collocates in a given grammatical relation and 20 related words. We did not present related words if there was only one to present.

It was necessary for all words presented to be entries in OALD themselves. Subsequently, all added words were links to the word’s OALD entry.

To add a collocate, the frequency of the collocation was at least five. This criteria was set after some discussion of the precision-recall trade-off: a higher threshold would give fewer lexicographically dubious collocates, but would mean there were fewer entries which were augmented, therefore reducing the scale of the experiment.

² The measure for collocates is logdice, based on the Dice coefficient. Measures are defined in the Sketch Engine documentation at <http://trac.sketchengine.co.uk/wiki>.

In the experiment, all collocates and related words were checked by an OUP lexicographer. The work took 8 to 10 hours for the initial 250 entries. (For 19 entries, no collocates or related words passed all filters, resulting in 231 for which entries were augmented.) Of 3367 links automatically added, 98 (3%) were removed.

While this procedure would be too expensive to augment all entries, for an experiment it was of great value as it exposed a number of areas of difficulty. One of these was web spam, a significant problem in enTenTen12 (Kilgarriff and Suchomel 2013). The exercise has focussed efforts on developing very large corpora with no, or very little, web spam. Another problem was a failure to identify, and set aside, proper names which were also lexical words.

We have a number of further ideas for improving the automatic filtering. We hope to gain access to a corpus which is smaller, but spam-free and processed with different tools. We would subsequently only include collocates if the collocation occurred at least once in the second corpus, and related words if they occurred above a threshold.

5.3 Presentation

The presentation of the augmented dictionary entry is shown below, for a concrete noun (*myrrh*), a verb (*iterate*), an adjective (*peevish*) and an abstract noun (*languor*). These examples also comprise entries with many or few added words.

The data were ready and the experimental run begun on July 4th 2013. Usage statistics were gathered using Google Analytics. At time of writing, the experiment is still underway and the results presented are provisional. In addition, the augmented entries account for only 0.5% of OALD web traffic, so sample size at this point is modest.

myrrh NOUN
 mɜː(r) BrE ; mɜːr NAme
 [UNCOUNTABLE]



a sticky substance with a sweet smell that comes from trees and is used to make perfume and incense

Beta: Collocates	▪ frankincense	▪ musk
MODIFIER	▪ aloe	PREPOSITIONAL OBJECT OF
▪ powdered	▪ patchouli	▪ tincture
MODIFIES	▪ sandalwood	PREPOSITIONAL OBJECT WITH
▪ unguent	▪ balsam	▪ perfume
AND/OR	▪ incense	

Beta: Related Entries	▪ eucalyptus	▪ geranium
▪ frankincense	▪ cardamom	▪ rosemary
▪ patchouli	▪ hyssop	▪ thyme
▪ sandalwood	▪ marjoram	▪ saffron
▪ bergamot	▪ nutmeg	▪ coriander
▪ peppermint	▪ musk	▪ anise
▪ chamomile	▪ jasmine	▪ lavender

Fig. 1: Augmented entry for *myrrh*

iterate VERB

'ɪtəreɪt  BrE ; 'ɪtəreɪt  NAmE



[INTRANSITIVE]

to repeat a mathematical or computing process or set of instructions again and again, each time applying it to the result of the previous stage

Beta: Collocates	▪ loop	▪ innovate
SUBJECT	AND/OR	▪ reiterate

Fig. 2: Augmented entry for *iterate*

peevish ADJECTIVE



'pi:viʃ  BrE ; 'pi:viʃ  NAmE

easily annoyed by unimportant things; bad-tempered

▶ **SYNONYM** IRRITABLE

▪ *Sebastian was a sickly, peevish child.*

▶ **peevishly**

'pi:viʃli  BrE ; 'pi:viʃli  NAmE

ADVERB



▪ *'It's your own fault,' she said peevishly.*

Beta: Collocates	▪ fretful
AND/OR	▪ irritable

Beta: Related Entries	▪ morose	▪ sullen
▪ sulky	▪ grouchy	▪ petulant
▪ churlish	▪ resentful	▪ fretful
▪ uncommunicative	▪ testy	▪ despondent
▪ dissatisfied	▪ taciturn	▪ uncooperative
▪ querulous	▪ quarrelsome	▪ irascible
▪ glum	▪ touchy	▪ crabby

Fig. 3: Augmented entry for *peevish*

languor NOUN



'læŋgə(r)  BrE ; 'læŋgə(r)  NAmE

[UNCOUNTABLE, SINGULAR] (LITERARY)

the pleasant state of feeling lazy and without energy

▪ *A delicious languor was stealing over him.*



▶ **languorous**

'læŋgərəs  BrE ; 'læŋgərəs  NAmE

ADJECTIVE

▪ *a languorous pace of life*

▶ **languorously**

 BrE ;  NAmE

ADVERB

Beta: Related Entries	▪ lassitude	▪ debility
-----------------------	-------------	------------

Fig. 4: Augmented entry for *languor*

5.4 Results for users

As the experiment had only been running for two months at time of writing this paper, and only on a small sample of entries, it is too early to have gathered feedback from users; this paper therefore simply emphasises SEO benefits. However, we have received three unsolicited reviews, from Poland:

I have opened the dictionary today and saw the additions for the first time. I think it is a great idea and very useful! Both Collocates and Related Entries can help my students and myself in learning and teaching English. They are very intuitive and easy to use. I do hope you will develop this BETA version and we will be able to use more of it soon. Congratulations on great improvement!

From Italy:

I've just come across the beta version panel and I think it is a great idea. I do like it and I wish I could find it as much as possible

And from Spain:

I really appreciate the usefulness of the “Relative Entries” addition. I think they are a good complement that helps very much in learning vocabulary. With them it is a pleasure to relate words that in another way are difficult to find for a foreign student. I would like that, little by little, you could increase the number of entries.

5.5 Results for SEO

To establish whether the augmentations have made a difference, it is necessary to compare web traffic for the same entries, pre- and post-augmentation. Moreover, since web behaviour displays annual cyclical patterns, it is best to compare data for the same dates in different years. Web traffic is measured using two variables: pageviews (the number of times a page was viewed), and visits (where a single visit may involve a number of pageviews, as the user navigates to and fro).³ In Table 1 we present figures for the 231 test entries for the same time periods (4 July – 3 Sept) in 2012 (pre-augmentation) and 2013 (post-augmentation).

	2012	2013	% change
Pageviews index	100	177	77%
Visits index	100	196	96%

Table 1: Test entries web traffic 2012 and 2013.

³ These constructs are defined in detail in Google Analytics documentation, where the relation between the indexes in the table and the actual numbers is also presented.

OALD web traffic has been increasing overall between 2012 and 2013, and this must be considered when determining if the augmentations have made a difference. The figures for OALD overall are presented in Table 2.

	2012	2013	% change
Pageviews index	100	142	42%
Visits index	100	166	66%

Table 2: All entries web traffic 2012 and 2013.

Thus, pageviews increased by 35% (77% minus 42%) more for the test entries than for OALD overall; visits increased by 30% more.

To establish whether the change in pageviews was significant, we established, for each of the 231 words in the sample, whether the 2013 figure was more than 42% higher than the 2012 figure. In 141 cases it was. If we were to accept the null hypothesis that the augmentation had had no impact, this number would have had a mean of 115.5 (231/2), and a standard deviation of 7.6. The observed figure of 141 is 25.5 (or 3.36 standard deviations) from the mean. We apply a two-tailed test and conclude with 99.9% confidence that the null hypothesis is false. Augmentations increase web traffic.

The change can also be observed in a graph. For the ten entries having the most pageviews in 2013, Fig. 5 shows search traffic for the period of January to July 2013. The red line shows the point where the augmentations were made. Four trend lines are shown in the graph:

- The blue line shows all visits to the ten entries.
- The orange line shows visits from search engines to the ten entries.
- The green line shows all visits from direct traffic (that is, not from search engines) to the ten entries.
- The purple line shows referral traffic (visitors who come from direct links on other websites rather than directly or from search engines).

6. Corpus size

As noted above, for the sample of words selected, there was often not enough data in 1.3 billion words. However these samples concerned fairly infrequent words. A one-billion-word corpus would be adequate for the approximately 20,000 commonest words of a language.

Another perspective is that, for the world's major languages, where there is ample data on the Web, we are in a position to prepare these very large corpora. Lexical

Computing Ltd. has recently built corpora of over 5 billion words for Arabic, English, French, Japanese, Portuguese, Russian and Spanish.

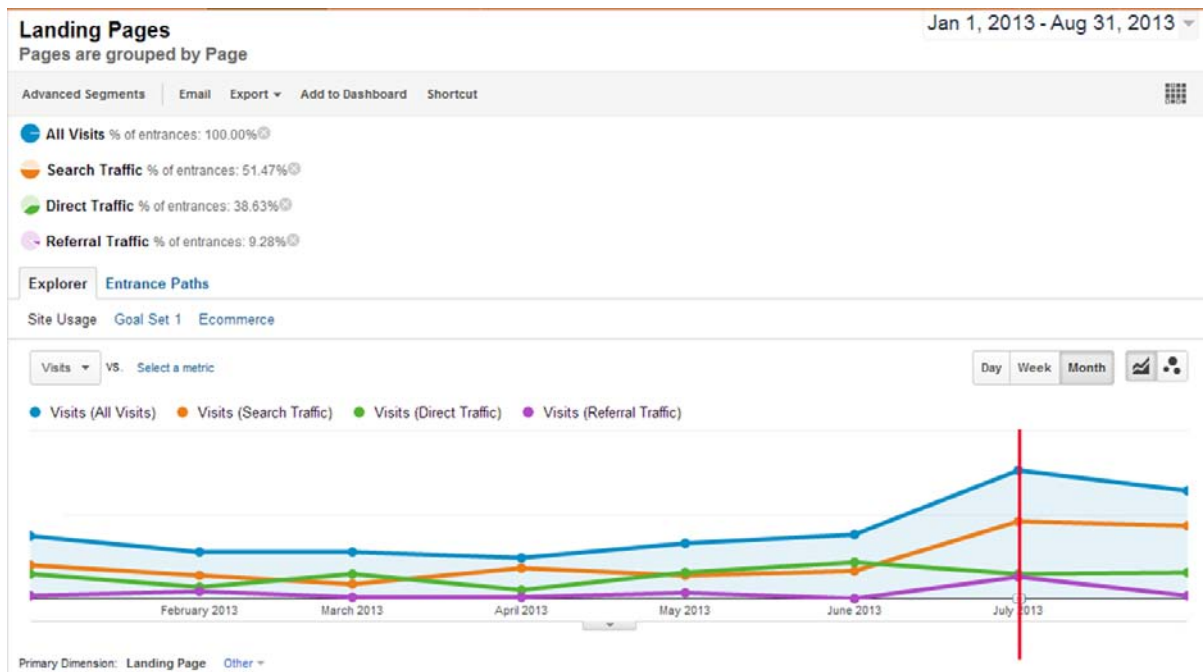


Figure 5: Pageviews for ten test entries, January to July 2013

7. Conclusion

Dictionary publishers in the age of the Web need their dictionary to fare well in search engine rankings. They therefore need to engage with Search Engine Optimisation. While there are many ways to achieve this, one that fits well with a corpus philosophy and which improves entries for human uses as well as for SEO, is to add collocates and related words (all hyperlinked to their own entries) to the entry. We ran an experiment to test the hypothesis that this method would increase web traffic. The experiment, for English, used the online version of the Oxford Advanced Learner’s Dictionary and augmented entries automatically with collocates and related words found using the Sketch Engine in the 11.3-billion-word enTenTen12 corpus. The experiment was run for a sample of 231 entries. Web traffic for these entries increased by 77% from the previous year, as compared to an increase of 42% for OALD in general.

Automatically augmenting dictionary entries with corpus-derived collocates and related words is an effective way of boosting web traffic with useful and relevant information to human users.

8. References

- Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3): 209-226
- Jakubíček, M., A. Kilgarriff, V. Kovář, P. Rychlý, V. Suchomel (2013). The TenTen Corpus Family. *Proc. Int. Conf. on Corpus Linguistics*, Lancaster, UK.
- Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell D. (2004). The Sketch Engine *Proc. Euralex*. Lorient, France.
- Kilgarriff, A. & Suchomel, V. (2013). Web Spam. *Proc. 8th Web as Corpus Workshop (WAC-8)*, Lancaster, UK.
- Lannoy, V. (2010) The IDM Free Online Platform for Dictionary Publishers. *Proc. Euralex*, Leeuwarden, Netherlands.

European Lexicography Infrastructure Components

Gerhard Budin^{1,2}, Karlheinz Moerth², Matej Ďurčo¹

¹Centre for Translation Studies, University of Vienna,
Gymnasiumstrasse 50, A-1090 Vienna

²Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Vienna

E-mail: gerhard.budin@univie.ac.at, karlheinz.moerth@oeaw.ac.at,
matej.durco@univie.ac.at

Abstract

Industrial dictionary production has long since started to make use of modern ICT, and while in the world of Academia one can still find many projects working with slip boxes and simple text processors, academic dictionary writing has also begun to move towards digital methods. Although there is plenty of software available, the situation for smaller groups of researchers and individual linguists looks rather bleak. Tools are there, what is – however – needed by many researchers is readily available, standards-based, interoperable, and sustainable infrastructure. In our paper we will describe particular infrastructure components that can be used in building lexicographic infrastructure and describe the work of a group of researchers of several Austrian academic institutions who are currently putting together existing pieces of software to build an integrated modular toolbox for academic dictionary writing that would enable researchers to create, maintain and publish digital dictionaries. In the introduction of the paper, we will also try to give an outline of the institutional settings in which these activities are being carried out which is important in view of the fact that all of the described components are designed as Austrian contributions to the European infrastructures CLARIN-ERIC and DARIAH.

Keywords: research infrastructures; eLexicography; standards, tools

1. Introduction

Industrial dictionary production has long since started to make use of modern Information and Communications Technology (ICT), and while in the world of Academia and smaller lexicographic projects one can still find researchers working with slip boxes and simple text processors, dictionary writing in general has also begun to move, step by step, towards digital methods. Although large amounts of software were developed for use in big publishing houses, the situation for smaller groups of researchers and individual linguists looks rather bleak, as many solutions come at forbiddingly high prices. Infrastructure is there; what is needed by researchers is more common infrastructure: readily available, standards-based, interoperable, and sustainable infrastructure. This report concerns Austrian developments that may help to remedy this problem.

2. Digital research infrastructures

There exist many definitions of research infrastructures. A recent one has been formulated by the European Commission in their *Legal framework for a European Research Infrastructure Consortium (ERIC)*:

“research infrastructure” means facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and covers major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information; enabling Information and Communications Technology-based infrastructures such as Grid computing, software and communication, or any other entity of a unique nature essential to achieve excellence in research.

While many institutions in the humanities are concerned with building up basic technical facilities and services, others have already begun to think about next generation research infrastructures: infrastructures that are supposed to foster international cooperation as the key to the “excellence of research” by means of knowledge and technology exchange. Key words in these discussions are the ‘Grid’, the ‘Cloud’ and ‘big data’.

2.1 ESFRI

In the European Union, the institutional foundation of activities in the field of digital research infrastructures started with ESFRI, the European Strategy Forum on Research Infrastructures, which was founded eleven years ago, in 2002. ESFRI is a group of national delegates and a representative of the Commission, who work together and pool resources to provide Europe with the most up-to-date research infrastructures. It is a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.

ESFRI’s task is not funding of projects, or realising infrastructures. It is rather an instrument to chart the landscape, to gather relevant information and to direct relevant developments. ESFRI has published a number of reports which describe the situation with regard to research infrastructures in the various scientific fields. In its Roadmap, an ongoing endeavour, it identifies potential new pan-European research infrastructures that are likely to be realised in the next 10 to 20 years. The number of candidate projects has been growing over the years. Roadmap 2006 listed 35 projects; the 2008 Update comprised 44. In 2010, the various scientific disciplines were organised into six major groups (Social Sciences and Humanities, Environmental Sciences, Energy, Biological and Medical Sciences, Materials and Analytical Facilities and Physical Sciences and Engineering) which comprise 48 projects (ESFRI 2010). The next update of the Roadmap is planned for 2015.

One example of a large-scale digital RI that countless researchers in the humanities

use (usually without even being aware of its existence) is GÉANT, the pan-European research and education network. GÉANT is a high-speed network interconnecting Europe's National Research and Education Networks (NRENs). It was launched to facilitate cooperation and to enable scientists to share knowledge and resources. An indispensable service many researchers access when travelling across Europe and its universities is *eduroam*, the international roaming service for users in higher education.

An example of best practice and standards of infrastructure components is the Text Encoding Initiative (TEI), which also caters for text-oriented researchers. Most of what the TEI offers belongs in the category of community-based standards. However, the TEI is more than that, as it also provides tools (e.g. standardised schemas, ROMA, OxGarage etc.) and very effective and well used communication channels, such as the TEI mailing list.

The number of projects bearing the term *infrastructure* in their name, or explicitly aiming to build infrastructures, has risen steadily in recent years. Those of interest with respect to the SSH disciplines include EUDAT (European Data Infrastructure), CENDARI (Collaborative European Digital Archive Infrastructure) and EHRI (European Holocaust Research Infrastructure).

2.2 Digital Humanities

The fields and disciplines with which we are concerned are at the top of the ESFRI list (ESFRI 2010). The two initiatives mentioned there are CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructure for the Arts and Humanities).

2.2.1 CLARIN

After a preparatory phase of several years, CLARIN was finally granted ERIC status by the European Commission in 2012.

CLARIN aims to provide easy and sustainable access to digital language data and advanced tools to discover, annotate, analyse or combine these data, irrespective of their physical location or format. The data involved are made up of a wide range of different types of language resources: representations of written and spoken language, some are text, others are offered as sound or video files. The target audience of CLARIN-ERIC are scholars in the humanities and social sciences. Currently, CLARIN-ERIC is in the process of establishing a networked federation of European data repositories, service centres and centres of expertise. They are planning to implement simple sign-on access for all members of the academic community in all participating countries. They are working on the interoperability of tools and data across the network, in order to allow researchers to combine distributed and heterogeneous data and to perform complex operations on these. This infrastructure is still under construction and will be so for quite some time. However, a number of

participating centres have already started to offer services providing data, tools and expertise. Currently there are nine certified CLARIN centres¹:

- ASV Leipzig, Bayerisches Archiv für Sprachsignale
- Berlin-Brandenburg Academy of Sciences and Humanities
- Eberhard Karls Universität Tübingen
- Hamburger Zentrum für Sprachkorpora
- IMS, Universität Stuttgart
- Institut für Deutsche Sprache
- MPI for Psycholinguistics
- Universität des Saarlandes

Others are preparing to obtain the official status of CLARIN centre:

- Centre of Estonian Language Resources (CELR)
- clarin.dk
- Austrian Centre for Digital Humanities
- DANS (Data Archiving and Networked Services)
- Huygens Instituut
- INL (Instituut voor Nederlandse Lexicologie)
- LINDAT-Clarin
- MI (Meertens Instituut)

2.2.2 DARIAH

The other focal large scale infrastructure initiative is DARIAH (Digital Research Infrastructure for the Arts and Humanities). As the name suggests it targets a very large community. Its declared goals are to enhance and support digitally-enabled research across the arts and humanities, to develop, maintain and operate an infrastructure in support of ICT-based research practices and to support researchers using ICT-enabled methods to analyse and interpret digital resources (DARIAH-EU Coordination Office 2013). The group of participating institutions and researchers is also aiming to set up an ERIC. DARIAH applied for ERIC legal status in autumn 2012.

In contrast to CLARIN, which organises its activities around physical service centres in the member countries, DARIAH has been operating through a network of four

¹ <http://www.clarin.eu/node/2971>

virtual competency centres (VCC):

- e-Infrastructure
- Scholarly Content Management
- Research and Education
- Advocacy

So far, each of the VCCs has been headed by two member countries and is formed of mixed groups of stakeholders. The VCCs have their own internal structure and specific workflows which are determined by the necessities of the particular tasks.

2.3 Infrastructure components

Infrastructure can be conceptualised in different ways, though this is beyond the remit of this paper. In a somewhat simplified manner, they can be seen as complex systems formed of a wide range of diverse technical (hardware, software, data) and organisational parts. Not all researchers require the same infrastructure components (ICs), and various disciplines have naturally varying requirements.

Language resources (LRs) are substantial in many fields today. Not only required by content producers and others active in cultural heritage, the work of an increasing body of researchers in SSH disciplines relies on availability of LRs. LRs can be described as a triad of tools, data and interoperability mechanisms. Tools comprise a combination of hardware and software, servers and services being put at the disposal of researchers. Data such as corpora, dictionaries, term-banks etc. constitute the contents, and interoperability mechanisms can be considered the glue that keeps tools and data together; they are the standards and norms that make LRs reusable. Neatly defined and well-documented interfaces are the basis for efficient service-based architectures that function in a distributed and heterogeneous digital biotope. In addition, we must not forget handbooks, documentation of all steps in the lifecycle of digital projects, and best practice guidelines in general to ensure reusability of newly-developed infrastructure components.

One particular type of language resource is dictionaries, which are an indispensable part of the scholarly tool inventory in many fields of the arts and humanities, in particular in all language-related disciplines. Libraries without dictionaries are unthinkable, and professionals, students, teachers, researchers and scholars equally use dictionaries, regardless of their field. Dictionaries have always been one of the most basic and integral elements of arts and humanities infrastructures.

2.4 The Austrian involvement

Research groups in Austria have been involved in both CLARIN and DARIAH for quite some time. In particular, two institutions played an important role in the

establishment of CLARIN and DARIAH in Austria: the University of Vienna and the Austrian Academy of Sciences. The tight institutional connection of CLARIN-AT and DARIAH-AT allows synergism between the two groups.

2.4.1 CLARIN-AT

As mentioned before, the CLARIN technical infrastructure is being built around physical centres; institutions that have sufficient resources and expertise to make long-term commitment more likely. In some countries, several candidates for such centres exist and will undergo an evaluation process before becoming official CLARIN centres. Others have only just begun the process of establishing such centres. Austria is currently establishing a national CLARIN centre, the Austrian Centre for Digital Humanities (ACDH). ACDH will provide the community with several services. One of these will be an OAI-PMH endpoint that will give Austrian researchers the opportunity to feed their metadata into the CLARIN network. The Open Archives Initiative Protocol for Metadata Harvesting (Lagoze et al. 2002) is a standard, offering a comparatively simple mechanism to expose structured metadata in the Internet that has been adopted by the CLARIN community.

Given the wide community with different requirements with regard to metadata (e.g. OLAC, Dublin Core, TEI Headers etc.), CLARIN did not try to impose any one particular metadata scheme for describing the resources, but rather introduced a generic overarching architecture: CMDI (Component Metadata Infrastructure) (Broeder et al. 2012) which is able to accommodate various metadata schemes. Austrian researchers were also active in the development of CMDI.

The availability of research data has become an important issue in recent years. While more and more relevant data are being produced, many institutions conducting research programmes are not in a position to ensure long-term availability of data. Very often, databases move with researchers, corpora are left behind at departments and are no longer traceable once projects have ended. Although funding agencies are getting increasingly aware of the issue and are trying to impose stricter policies, many institutions neither have the required infrastructure nor the funds for long-term preservation of research data generated in these projects.

ACDH is planning to function as a host for such data, while also attempting to access already relinquished and forgotten data. It will offer researchers access to a dedicated repository for linguistically relevant research data.

2.4.2 DARIAH-AT

Austria is heading (together with Germany) the DARIAH Virtual Competency Centre 1. VCC1 is in charge of digital infrastructures; in a manner of speaking, taking care of the infrastructure of the infrastructure. In the context of the overarching project, this implies very particular core services such as authentication and authorisation, persistent identifiers and infrastructure components.

At the moment, DARIAH-AT's top priorities are digital infrastructures for the creation, maintenance and publication of digital language resources, in particular lexicographical data and large text collections. This is motivated by the general interests of the main partners currently involved in the construction work, which are departments concerned with linguistic, lexicographic and terminological research questions.

3. Lexicography infrastructure

The following paragraphs will provide detail about the infrastructure components that have come into existence as part of Austria's CLARIN and DARIAH engagements.

3.1 Dictionary-in-a-box

'Dictionary-in-a-box' is designed as an integrated modular toolbox offering lexicographers, working as individuals as well as in groups, all the necessary software to create, maintain and publish digital dictionaries. This suite is designed as a comprehensive virtual research environment geared towards the needs of researchers collecting lexicographic data. The target group is quite diversified, intended to include linguists from various fields, professionals in need of a simple lexicographic infrastructure, terminologists, etc. The suite will consist of the freely available dictionary editor Viennese Lexicographic Editor (VLE), styles, schemas, and server scripts that can be easily distributed and handled.

3.2 Dictionary editor

There exists a great deal of software for editing lexicographic data. Indeed, the list of well-established dictionary editing applications is quite long (for a short list see Budin and Moerth 2011). Some of these products provide a wide range of functionalities which can be applied to the whole lifecycle of the dictionary creating process: collecting, editing, refining and enhancing lexicographic data. Some packages are fully integrated systems; others are built in a modular way. Some are being used for particular purposes such as endangered languages, while some offer specialised multi-media support. Technically, dictionary writing software is often built around RDBM systems, very often making use of some client-server or multi-tier architecture.

The above mentioned VLE is a fairly new piece of software that came into existence as a by-product of an entirely different development activity. It was developed as part of an interactive online learning system for university students. It was first used in a collaborative glossary editing project carried out as part of university language courses at the University of Vienna. Over time, the tool proved to be sufficiently flexible and adaptable, and was put to work for other purposes in other projects. The interface is built around an XML editor that provides a number of functionalities

typically required in editing linguistic data.

The motives to embark on this project were manifold. Some of the already existing systems were primarily intended for use in big publishing houses, pricing of licences accordingly high and the software consequently out of reach for small projects producing dictionary data. As the software evolved as a by-product of several smaller projects, production costs were manageable. In addition to the economic limitations, our projects were in need of full support for varying XML formats. The application was supposed to process standard-based lexicographic and terminological data such as LMF, TBX, and TEI. We were in need of simple scripting capabilities, a configurable interface allowing access to corpora and offering support for sophisticated validation mechanisms.

One of the particular features of VLE is a special module easing the integration of corpus examples into dictionaries. The main goal when programming this module was optimised access to digital corpora. It was intended to enable lexicographers to gather relevant sample sentences from external resources such as structured corpora (or the Internet) and to integrate these into dictionary entries in a reasonably comfortable manner. The focus in this work was direct access to the data. VLE's corpus interface enables lexicographers to launch corpus queries, and offers functionalities for selectively inserting data into existing dictionary entries without using the clipboard to copy-and-paste, which inevitably results in a lot of inefficient typing or clicking.

So far, VLE has been used to edit LMF, TEI, TBX and RDF data. The program provides a number of useful functions to automate editing procedures. Some of these cater to the needs motivated by the underlying XML structure. The editor is capable of highlighting XML elements and performing automatic text completion. The program can continually check the structural integrity (well-formedness) of input on the fly. Technologically, it draws not only on the XML core specification, but also on several cognate technologies. XSLT and XPath play an important role both for visualising and modifying existing datasets. Lexicographers can insert elements on the basis of predefined XML Schemas. Most of the functions can be applied both to single and multiple records.

Validation is a key issue in all XML based document editing. It is the process of checking the data on a level beyond the basic structural XML requirements (well-formedness). When validating the structure of a document, it is checked against a set of definitions of permissible elements and information as to where these elements may appear in the document. Currently, VLE expects document type definitions in the form of an XML Schema which is, like XML, a W3C recommendation. On the to-do-list of the programmers, there is also the implementation of an option to validate against RELAX NG, an ISO standard which has found much support in the TEI and OpenDocument communities.

In VLE, editing of dictionary data can be performed in two ways: the editor works either in XML mode (Figure 1), which may be considered as the expert mode, or in an editor form with predefined entry controls. The second option enables working in an interface made up of controls that are arranged like traditional database input fields. While working on an entry, it is possible to switch between the two modes. The second option, i.e. making use of edit controls for particular XML elements, is useful especially when working in the same field across a number of dictionary entries. Navigating is admittedly more cumbersome in the expert mode than in the edit controls. However, more complex structures, in particular elements nested inside one another, often make it necessary to switch into XML mode.

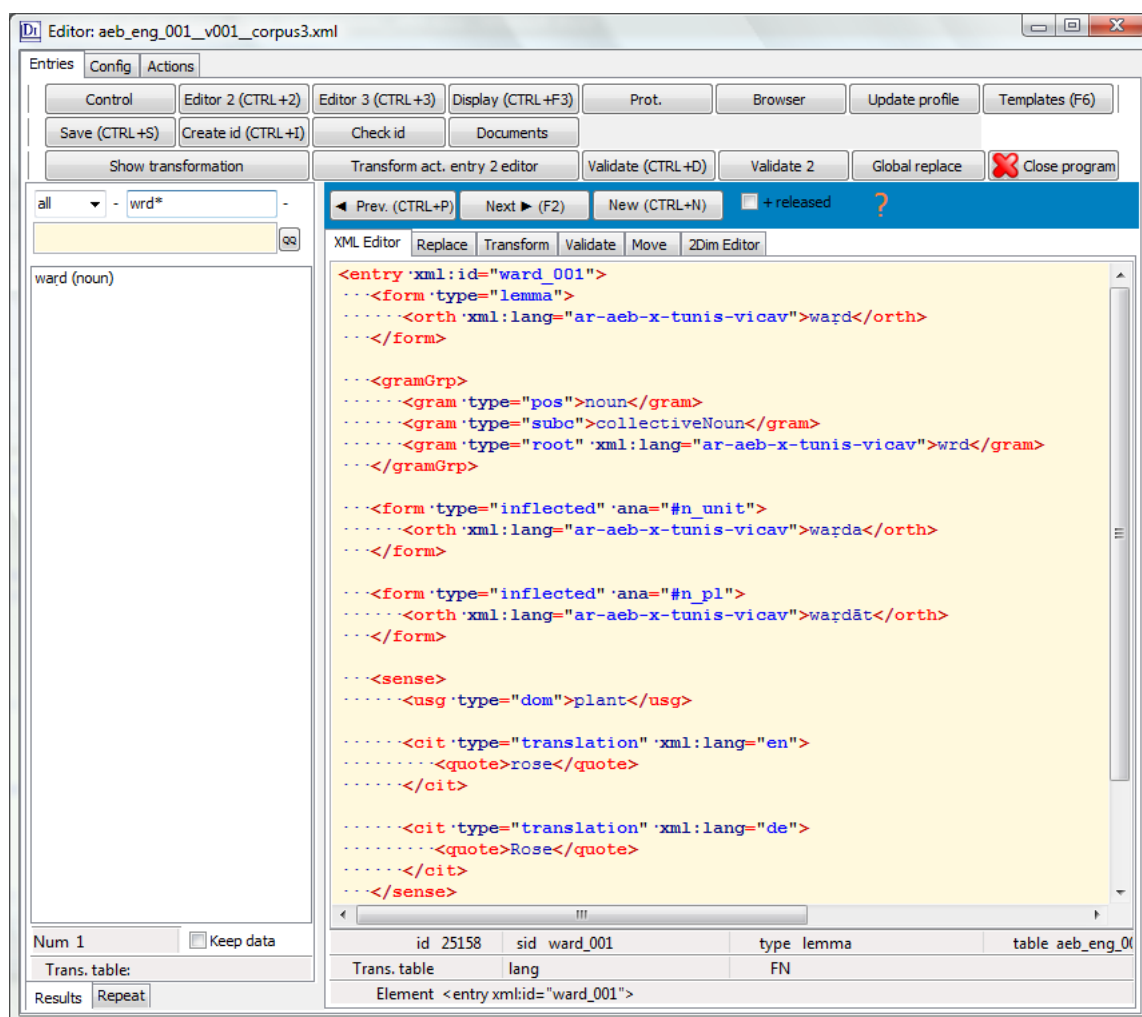


Figure 1: XML view

The tool visualises data by means of freely configurable XSLT stylesheets (Figure 2). While this functionality is quite commonplace in many XML based applications today, VLE proves to be particularly versatile. It is possible to apply different styles by switching between different views of the same set of data. Automatically generated

links in the output data (usually HTML) enable navigation from these visualisations back into the editor control.

The program has a number of features that are intended to ease the lexicographer's workload. One of these features is a configurable keyboard layout which is designed to support the comfortable input of Unicode characters usually not available in standard key assignments. The software can be configured to automatically choose the appropriate keyboard assignment when moving from one element to another. This functionality is based on the `@xml:lang` attribute and spares the user from manually switching between keyboard layouts. For example, when working on contents of an element having an `@xml:lang="ru"` attribute, VLE automatically activates the Russian keyboard layout; on entering an element with the attribute `@xml:lang="de"`, it switches back to German. The program is able to automatically create unique and meaningful identifiers for entries and example sentences on the basis of the contents of the respective items.

The current VLE version is a stand-alone application that requires Windows operating system. One of the project's midterm goals is the development of a fully-fledged browser-based interface. While the list of requirements for such an interface is clearly defined, the implementation would require time and resources which are currently being sought.

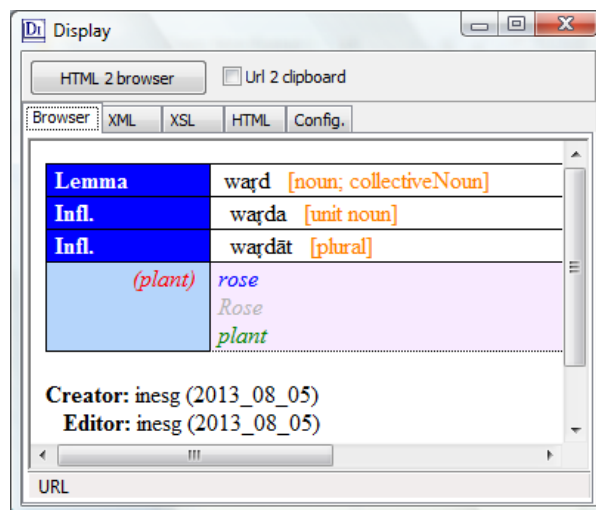


Figure 3: HTML view

3.3 Dictionary server

Usually, VLE does not work on locally stored data. Data are stored on a remote server that can easily be set up and configured. The system is organised as a client-server architecture. The communication between the dictionary client, i.e. VLE, and the server has been implemented as a REST (Representational State Transfer) web

service, which facilitates access to the server and consequently the data with tools other than our own.

The server builds entirely on open and freely available software that can be readily distributed. In the first implementation level, it makes use of the MySQL database, which is connected to clients through a REST-style web service. Querying works on the basis of SRU/CQL (Search/Retrieval via URL + Contextual Query Language). This search protocol was developed by the Library of Congress as successor of the Z39.50 protocol and is being tested and worked on by CLARIN's Federated Content Search (FCS) working group.

The distributed architecture has a number of obvious advantages. Being able to work on the data wherever one has access to the internet is unquestionably a useful feature. Lexicographers can work from anywhere, without having to carry their data around. But, most importantly, this setup also allows for collaborative work on the dictionary data.

VLE allows several editors to work simultaneously on the same dictionary, making use of a simple locking mechanism. When one lexicographer opens an entry, the entry can still be read by other editors, but cannot be edited. An additional feature of the server module currently being developed is an efficient versioning mechanism. We anticipate that this functionality, which might be of particular interest in collaborative settings, will be available by early 2014.

3.4 DictGate

DARIAH-AT is planning to set up a server that will allow (groups of) researchers to host lexicographic data. This infrastructure is intended both for producing and publishing lexicographic data. Thus, the dictionary gate is designed as a two-lane carriageway that will allow both data entry and retrieval. Users will be able to use the central server to produce data and to set up web-based interfaces that make use of the DictGate's web services.

Primary target groups are not commercial entities but researchers working on smaller lexicographic projects that are in need of solutions that can be applied without much logistical and technical overhead. Institutionally, the service will be based at the Austrian Academy of Sciences which has a long-standing and quite diversified tradition in dictionary production.

3.5 Lexicographic data

With respect to data, the DictGate working group pursues several lines. A first stock will be provided by lexicographic data that are being created in the context of the VICAV (Vienna Corpus of Arabic Varieties) project. The contributors of this project are currently setting up a platform to host and exchange a wide range of digital

language resources for Arabic studies. Among these data (language profiles, bibliographies, corpora, ...) there are also smaller digital dictionaries of Arabic varieties. Besides of Damascus Arabic, dictionaries for the varieties of Morocco (Rabat) and Egypt (Cairo) are being compiled. A dictionary of Tunis Arabic will be elaborated as part of the project *Lexical dynamics in the Greater Tunis area: a corpus based approach*, which was approved by the Austrian Science Fund in March 2013 and will run for three years. These four dictionaries are being compiled with a special focus on comparative research questions and are structured in a manner that will enable performing queries on the four dictionaries to retrieve integrated datasets. These language resources (tools and data) are intended both for research purposes and academic language instruction.

There are several other research groups that plan to publish their data through the DictGate platform. A first product will be a Persian–English Dictionary of Single Word Verbs and a Russian–German dictionary which is currently being developed. One of the long-term dictionary projects of the Austrian Academy of Sciences, the *Dictionary of Bavarian dialects in Austria*, is also involved and will contribute data to the platform. While the current focus is on linguistics, the project generally targets a wider humanities audience. Among the resources to be made available there are also historical dictionaries that are of interest for disciplines other than linguistics.

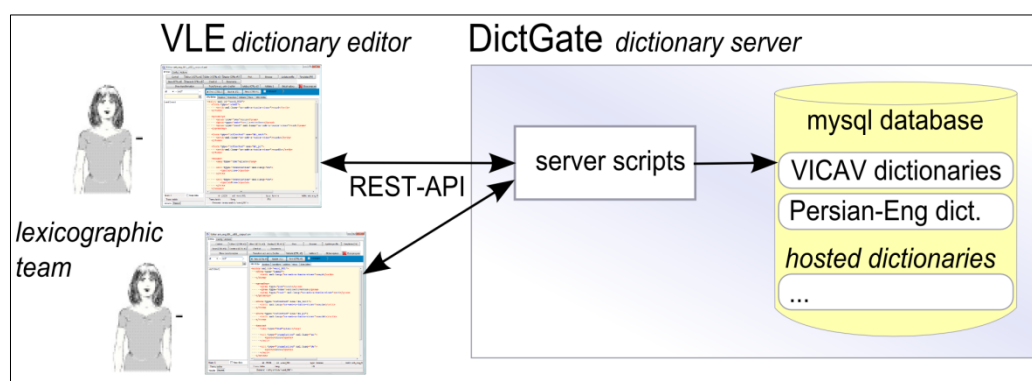


Figure 4: Client-server setup of the lexicographic infrastructure

We will also strive to access data without institutional backing to host or maintain them, to work towards efficient service-based lexicographical infrastructures that also offer data that can be used for NLP applications.

3.6 Access Policy

Basically, the focus of all these activities is on open-access resources. So far, no binding decision has been made as to the licence under which DictGate offerings will be available. However, there is a strong case for a Creative Commons licence, CC-BY being the favoured option. Discussions with interested researchers and other stakeholders have shown that permission to create derivative works is usually

regarded an important prerequisite in order to ensure reuse of data.

Free access will not be a preclusive condition for the incorporation of data in the DictGate platform. However, open access will be strongly encouraged as funding organisations increasingly demand open access to publicly funded research data. In this respect, it will be important to get to a point where truly open access to data implies more than the availability of pdf documents, but direct access to data in reusable (i.e. standardised) formats.

4. Standards

Both CLARIN-ERIC and DARIAH consider standards a major concern of their activities and have institutionalised their respective work. CLARIN-ERIC has set up a Standards Committee to advise the Board of Directors on the adoption of standards to be supported by infrastructure. In the DARIAH network, various working bodies share the declared intention of working on standards and the formulation of best practises. As a particular form of language resource, standards, technical specifications and best practises are thus to be regarded as important cornerstones of digital infrastructures and should be considered infrastructure components in their own right.

When creating digital lexicographic resources, several standards and de-facto standards have to be considered. There are, for example, LMF (Lexical Markup Framework, ISO 24613:2008) and the dictionary module of the Guidelines of the Text Encoding Initiative. The bundle of documents created by ISO-TC37 (Terminology and other language and content resources) also contains a number of relevant specifications, such as ISO 639 (Codes for the representation of names of languages) or ISO 24610-1:2006 (Language resource management – Feature structures – Part 1: Feature structure representation), that should be considered.

As to the format used by the software components of the proposed infrastructure services, the goal was to come up with solutions that would be as open and flexible as possible. The core data of the initial phase of the project will be encoded in TEI P5². This is in particular due to the fact that the contributing partners provide data in this format. The involved projects are mostly rooted in humanities disciplines that have a long tradition in making use of the TEI guidelines.

The guidelines of the TEI comprise an ample set of well-tried, and in many parts thoroughly discussed, specifications for a wide range of encoding scenarios. It has grown as the de-facto encoding standard for dictionaries digitized from print sources. Interestingly, the most recent versions of the TEI Guidelines contain a passage that indicates that the authors are actually aiming at a much wider range of dictionaries:

² <http://www.tei-c.org/Guidelines/P5/>

... The elements described here may also be useful in the encoding of computational lexica and similar resources intended for use by language-processing software; they may also be used to provide a rich encoding for word lists, lexica, glossaries, etc. included within other documents. (TEI Consortium P5 2012, 247)

The idea of extending the scope of the TEI dictionary module for use with language-processing software is not as far-fetched as it may seem at first glance. The interest in the issue has been clearly documented by the large audience of the workshop “Tightening the Representation of Lexical Data: A TEI Perspective”, which was held at the 2011 Annual Conference and Members’ Meeting in Würzburg (Germany).

The dictionaries to be published in the first round share a common schema which was developed on the basis of the TEI dictionary module. This schema is made up of a comparatively small subset of elements and imposes a number of clearly defined constraints to make the resulting dictionaries interoperable with one another and some other language resources.

In using the Guidelines of the TEI for linguistic and lexicographic purposes, encoders usually combine them with other standards in a complementary manner. Thus, it has become common practice in TEI encoding to make use of the global attribute @xml:lang which has been incorporated into the Guidelines from the World Wide Web Consortium’s XML Specification. TEI prescribes this attribute to identify both linguistic varieties and writing systems. In this hybrid approach, the value of the attribute should be constructed in accordance with the Internet Engineering Task Force’s *Best Current Practice 47* (BCP 47) which in turn refers to and aggregates a number of ISO standards (639-1, 639-2, ISO 15924, ISO 3166).

An equally important tool applied in the encoding of these dictionaries is ISOcat, the ISO TC 37 (Terminology and Other Language and Content Resources; Kemps-Snijders et al. 2009) Data Category Registry³ that has been set up as a publicly available pool for definitions of widely accepted linguistic concepts. ISOcat can, for instance, be applied in TEI when annotating word forms with word class information. The ISOcat database assigns each data category a unique persistent identifier which makes them universally identifiable.

Additional infrastructure components to be contributed by the Austrian partners of CLARIN-ERIC and DARIAH belong to the third type of the above introduced data-tools-interoperability triad. It is not only important to adhere to standards. To enable others to work along similar lines, thorough documentation and examples are needed that in turn can serve as the basis of new projects and further developments in ongoing standardisation processes.

³ <http://www.isocat.org>

5. Status

At the time of preparing this report, most of the components described here are functioning and in use by researchers for their everyday work. Distributable prototypes of Dictionary-in-a-box are currently being tested and a first version will be available by early next year. The dictionary editor is already available and can be freely downloaded through the Language Resources Portal of the Institute of Corpus Linguistics and Text Technology⁴.

6. Conclusions

In this report, we introduced a suite of easy-to-adopt tools for collaborative lexicographic work and their embedment in evolving SSH research infrastructures. All of this work is driven by a vision of a growing ecosystem of freely accessible and distributable lexical resources being used by growing communities of researchers. We hope that our open concept and the readily available infrastructure will create new and sustainable dynamics in the field of lexicographic data production.

7. Acknowledgements

The work described in this paper has been made possible by funds provided by the Austrian Federal Ministry for Science and Research (Bundesministerium für Wissenschaft und Forschung), the University of Vienna and the Austrian Academy of Sciences.

8. References

- Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T. & Trippel, T. (2012). CMDI: a Component Metadata Infrastructure. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*, p. 1.
- Budin, G., Mörth, K. (2011). Hooking up to the Corpus: the Viennese Lexicographic Editor's Corpus Interface. In *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, Bled, 10–12 November 2011*, edited by Iztok Kosem and Karmen Kosem. Ljubljana: Trojina, pp. 52–59. Institute for Applied Slovene Studies.
- Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries. In *Journal of the Text Encoding Initiative (TEI and Linguistics)* 3.
- DARIAH-EU Coordination Office (2013). *Introducing DARIAH-EU*. Accessed at:

⁴ <http://oeaw.ac.at/iclitt/vle>

- www.dariah.eu/indexcc3b.pdf.
- European Commission (2010). *Legal framework for a European Research Infrastructure Consortium – ERIC. Practical Guidelines*. Accessed at: ec.europa.eu/research/infrastructures/pdf/eric_en.pdf.
- European Science Foundation (2011). Research Infrastructures in the Digital Humanities. In *Science Policy Briefing 42*.
- European Strategy Forum on Research Infrastructures (2010). *Strategy Report on Research Infrastructures. Roadmap 2010*. Luxembourg Publications Office of the European Union. (doi:10.2777/23127)
- Ide, N., Kilgarriff, A. & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In *Proceedings of the Ninth EURALEX International Congress: EURALEX 2000*: Stuttgart, Germany, August 8th–12th, 2000, 113–126. Stuttgart: Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.
- ISO-24612:2012 (2012). *Language resource management – Morpho-syntactic annotation framework*.
- ISO-24613:2008 (2008). *Language resource management – Lexical markup framework (LMF)*.
- Kemps-Snijders, M., Windhouwer, M., Wittenburg, P. & Wright, S.E. (2009). ISOcat: Remodelling Metadata for Language Resources. In *International Journal on Metadata, Semantics and Ontologies 4*: pp. 261–276.
- Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. June 2002. Accessed at: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>.
- Romary, L., Salmon-Alt, S. & Francopoulo, G. (2004). Standards Going Concrete: From LMF to Morphalou. In *Workshop on Enhancing and Using Electronic Dictionaries*. Geneva: Coling.
- Romary, L. (2010). Standardization of the Formal Representation of Lexical Information for NLP. In *Dictionaries: An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography. Accessed at: <http://arxiv.org/abs/0911.5116>.
- Romary, L. (2010). Using the TEI Framework as a Possible Serialization for LMF. Paper presented at RELISH workshop, August 4–5, 2010, Nijmegen, Netherlands. Accessed at: <http://hal.archives-ouvertes.fr/docs/00/51/17/69/PDF/NijmegenLexicaAugust2010.pdf>.
- TEI Consortium (2013). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.3.0. Last updated on 17th January 2013. Accessed at: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.
- Wandl-Vogt, E. (2010). Multiple access routes. The Dictionary of Bavarian Dialects in

Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In *eLexicography in the 21st century. New challenges, new applications, Proceedings of eLex2009*. S. Granger & M. Paquot (eds), Louvain-la-Neuve, Presses Universitaires de Louvain: pp.451-455.

Language Web for Frisian

Hindrik Sijens, Anne Dykstra

Fryske Akademy, Doelestrjitte 8, 8911 DX Ljouwert / Leeuwarden, The Netherlands
E-mail: hsijens@fryske-akademy.nl, adykstra@fryske-akademy.nl

Abstract

The Fryske Akademy has developed a web portal, the Frisian Language Web, which consists of an online spell checker, a machine translator ('Oersetter') and a dictionary portal. These three applications will make a unique language tool for native speakers and learners of Frisian to help them to write proper Frisian. An important part of the Language Web will be a standardized word list of Frisian. The standard word list will be incorporated into a database underlying the spell checker. This database also contains a list of non-standard words that are linked to standard forms. This makes it possible to use the spell checker to guide users who write non-standard Frisian towards use of the standard language. The 'Oersetter' service will be a statistical machine translator based on a bilingual Dutch–Frisian parallel corpus. The dictionary portal will consist of existing, relatively recent, lexicographic material. Paper dictionaries and terminology lists were digitized, xml-parsed and linked to each other. The dictionary portal gives access to a wealth of information not (easily) accessible in the paper counterparts of the various dictionaries. By the different nature of the individual dictionaries, the user can draw on a wealth of lexical material. He has access to the portal through two languages, Dutch and Frisian. The dictionary portal will be the starting point of a new project: an online bilingual dictionary Dutch–Frisian.

Keywords: Standard wordlist; Language Web; Dictionary Portal; Frisian Language

1. Introduction

One of the themes of the fourteenth Euralex congress, which was held in Leeuwarden in August 2010, was the lexicography of lesser used and non-state languages. Keynote speaker Anne Popkema reported on a survey that was conducted by the Fryske Akademy about the state of the art of the lexicography of these languages (Popkema, 2010). The questionnaire contained several questions about the lexicographic output in a given region. With regard to the level of diversity of lexicographic output, West Frisian¹ got eight out of the maximum of ten points. The fact that Frisian had a monolingual dictionary and several bilingual ones yielded a good classification. There was some reason to be proud to see Frisian nestled between much larger minority languages such as Catalan and Basque. But in terms of the level of use of modern technology in lexicographical practice, the situation was not so rosy. Frisian and

¹ The term West Frisian is used to distinguish it from North Frisian, the variant of Frisian which is spoken in northern Germany (Schleswig-Holstein). In the remainder of this paper, we will use the term Frisian, to refer to West Frisian.

Friesland received poor results in terms of the availability of online dictionaries, but had satisfactory results in terms of dictionary writing software and the use of an electronic corpus. Frisian clearly lagged behind compared to languages like Catalan, Basque, Friulian and Welsh. It is absolutely desirable that the Frisian language performs better in terms of use of modern technology in lexicographical practice. The first steps to reach a higher level were taken in the past two years: the development of ‘Taalweb’, a language web for Frisian. This facility consists of an online spell checker, a machine translator called ‘Oersetter’ and a dictionary portal.

The core of ‘Taalweb’ is a newly created standard wordlist of the Frisian language. In this paper we will outline ‘Taalweb’ for Frisian. Furthermore, we would like to introduce our next project: a new and sophisticated online Frisian–Dutch dictionary.

2. Friesland – Frisian Language

Friesland is a province in the Netherlands, with 650,000 inhabitants. It is a bilingual province, about 54% of its inhabitants have Frisian as their mother tongue and about 65% are able to speak the language. About 25% of its inhabitants have Dutch as their mother tongue. Several other vernaculars are spoken by about 10% of its inhabitants. Frisian is a Germanic language and historically it is the closest extant language to English. Recent figures on language use show that 85% of the inhabitants of Friesland are able to understand Frisian and about 75% of the inhabitants of Friesland are able to read it. Almost 64% claim they are able to speak Frisian well. Only 10% are able to write Frisian well, about 18% quite well, while some 70% are unable or almost unable to write in Frisian (Taalatlas, 2011).

Dutch is the official language of the Netherlands. It is used as the first language in formal domains such as administration, education, commerce, and the media. Frisian is the second official language of the Netherlands, but the language is used more intensively orally than in writing. The main reasons for Frisians to use Dutch and not Frisian as a written language are that Dutch has a higher status, and the spread of writing competence in Frisian is insufficient.

3. Standard Wordlist

Like most lesser-used and non-state languages, Frisian encounters difficulties in developing a standard. Because there exists no long and extensive tradition in written language, and because of the fact that there are different coexisting dialects, Frisian has no fixed standard. Consequently there are quite a few frequent dialectal differences in the written language and therefore also in dictionaries. One example of this is the Frisian word *giel* (yellow) which is pronounced as /gi.əl/ in the northern part of Friesland and as /ge:l/ in the southern regions. As the word is pronounced and written in two different ways, there are two entries in the dictionaries: *giel* and *geel*. Another example of variation in the spoken language, which we also find in the

written language, is the paradigm of the irregular verb *gean* (to go). The first person past tense comes in three forms: *ik gie nei hûs*, *ik gyng nei hûs*, *ik gong nei hûs* (I went home). All forms occur in the written language.

To give an idea about how many possibilities there are for some frequently used adverbs, take for example the word *eigentlichen* (actually, in fact, really). The existing dictionaries recorded twelve variants:

- ***eigentlichen***
- ***eigentlich***
- *eigentlichs*
- *eigenliken*
- *eigenlik*
- *eigenliks*
- *einliken*
- ***einlik***
- *einliks*
- *einken*
- *eink*
- ***eins***

On the basis of morphological principles and frequency counts we have chosen four of these forms to be included in the standard wordlist: ***eigentlichen***, ***eigentlich***, ***einliks***, ***eins***.

But not only is this variation or these dialectical differences, such as like *gie*, *gong* or *gyng*, part of the language, but Dutch-isms are too. Frequently used Dutch words such as *lui* (lazy) or *gebeure* (to happen) are often included in the dictionaries, together with their proper Frisian equivalents *loai* and *barre*.

Of course, dialectical variation demonstrates the richness of a language, but also creates uncertainty for hesitant users and doubting language learners. What form should they choose: *giel* or *geel*, *ik gie*, *ik gong* or *ik gyng*, all correct Frisian forms? The same can be applied to the occurrence of Dutch-isms in Frisian like *lui* and *gebeure*.

It can be difficult to choose between sometimes obsolete but correct Frisian words like *loai* en *barre* and contemporary, frequently-used Dutch-isms *lui* and *gebeure*. This doubt regarding correct usage is rooted in a lack of education and routine in writing Frisian. Frisian only became a compulsory school subject in the second half of the last century. In addition, because this obligation applied only to primary schools and because written Frisian in daily life plays a minor role, most Frisian people are not proficient in writing their own language. Language learners, as well as native speakers, are insecure in their language use; they fear to make mistakes. Even language professionals such as journalists, editors, translators and novelists experience these kinds of problems. Since the lack of a standard is felt to be an

obstacle to the use of written Frisian, language professionals uttered a desire to standardize the language. At the same time, the policy of the provincial government of Friesland is to promote written Frisian. The desire to standardize Frisian is in accordance with provincial policy. The provincial authorities therefore asked the Fryske Akademy to compile a standard wordlist of Frisian.

The existing spelling system proved to be quite complex and inconsistent. The Fryske Akademy suggested a moderate spelling reform to the provincial parliament. With a solid description of the spelling rules as a starting point, the next step was to extract a list of words from the existing language corpus and dictionaries. This basic list of 145,000 lemmas had to be edited, because it contained duplicates, homonyms, dialect forms, misspelled forms, Dutch-isms and obsolete words. With the help of a specially designed database, which contains Frisian words with their morphological structures, the individual paradigms were automatically generated, with a considerable degree of success.

Another hurdle was to develop criteria to choose the standard forms. In order to create consensus, the standard forms are usually taken from the two main dialects of Frisian. In deciding which variant should be the standard, frequency plays a role, but frequency is not always decisive. In some cases we have chosen the historical lexicalized form of a word instead of the historically correct form. In other cases, the criterion distance played a role. The form most remote from the Dutch equivalent was preferred. For instance, in the case of *giel* versus *geel* mentioned above, the chosen standard form is *giel*, because this form is different from the Dutch form *geel*.

Analogy as a criterion also played a role. The form *read* (red) is realized as *read* and with d-deletion: *rea*. However, in inflected forms like *reade flagge* (red flag) the /d/ is always written and pronounced. Therefore we have chosen *read* as the standard form.

This standard wordlist is a reliable tool for anyone who wants to write Frisian. It is a benchmark for the language and a basis for the language technology products that are part of ‘Taalweb’.

4. Spelling Checker

The standard wordlist is the core of a new spelling checker tool for Frisian. The history of Frisian spelling checkers began in the early nineties of the last century. A word list derived from the then existing dictionaries and databases was implemented in WordPerfect, at that time the most common word processor. In the late nineties, the Fryske Akademy together with the Dutch language technology company Polderland, developed a spelling checker for Microsoft Word. Ten years later the same team created an electronic language assistant for Microsoft Office, consisting of two bilingual dictionaries, a spelling checker and an option to correct and improve

texts, called ‘Taalhelp’ (Language Help). The production of these spelling checkers and tools was supported by a grant from the provincial government of Friesland.

Due to the changes in new releases of Microsoft Word, the tools were no longer compatible with Office 2010. The need for a new spelling checker has since been increasingly felt. Because it is provincial policy to support the use of written Frisian, the province financed the development of a new spelling checker. The new spelling checker is a plugin compatible with Microsoft Word, but it can also be accessed online.



Figure 1: hy **ston** sich te skearen



Figure 2: hy ston **sich** te skearen

An example illustrates the design of this new tool. In the Frisian sentence *hy ston sich te skearen* (He was shaving himself), the spelling checker highlights the verb *ston*

and the pronoun *sich*. *Ston* is a dialect form of standard Frisian *stie* (stood), which is suggested by the spelling checker.

The reciprocal pronoun *sich* is marked as Dutch-ism. In Frisian the pronouns *him* (him) or *har* (her) should be used and in this context, *him* is the most likely.

On the back end of this unique language tool we have stored a complex system of alternative, non-standard forms and Dutch-isms, all linked to the preferred standard form. Whenever an incorrect or a non-standard form is encountered by the spelling checker, the author will receive suggestions to improve and correct his text.

However, solving one problem is creating another. The spelling checker correctly marks an alternative form like *ston* (pret. 1 sing.) as ‘variant’ of *stie*. The verb *wurde* (to get, to become) however has a standard paradigm form *wurde* in the present tense which is identical to a variant form in the past tense.

	StandardLem	StandardPara	Variant1
2	stean		
4		stean	ston
5		stiest	stonst
6		stiet	ston
7		stean	steane
8		stie	ston
31	wurde		
32		wurd	
34		wurde	
35		waard	wurde
37		waarden	wurden
38	wurd		
39		wurdsje	
41		wurden	
42		wurden	

Figure 3: paradigm of verb *wurde*

Since the spelling checker is not a grammar checker, the non-standard form *wurde* cannot be marked in the same way as *ston* has been marked in the previous example. However, the user must be drawn to the fact that he may have typed a non-standard form. But as long as there is no grammar checker available, we use a practical solution for this shortcoming.

When a user types a sentence like *hy wurde ilk* (he became angry), using the variant form *wurde* instead of *waard*, the spelling checker marks *wurde* and alerts the user: this is a standard form, but it can also be a non-standard, dialect form. The form *waard* is proposed, but if the user deliberately chooses to write dialect forms, he can

ignore the suggestion. And of course, if the user has typed a sentence in the present tense, for instance *wy wurde lilk* (we become angry) he also can ignore the suggestion.

This problem does not occur only in verbs, but also with homonyms. The numeral *alve* (11, eleven) has a non-standard form *elf*. But *elf* is also a noun which refers to a figure that appears in fairy tales and fantasy films.

As already mentioned, the wordlist represents the standard language, but this does not imply that the non-standard forms are always incorrect. Non-standard word forms still can be Frisian word forms. And the author can deliberately choose to use variants because they belong to his dialect or personal language. But it is to be expected in future that the standardized words will increasingly displace non-standard forms in written Frisian.



Figure 4: *waard* and (non) standard *wurde*

It is not our intention to rebuke the Frisian writing people in a pedantic way, or to discourage them from writing in Frisian. One of our aims is to guide people from their own (local) variant to the Frisian standard language. Moreover, the standard will be prescribed in teaching and strongly recommended in official language. And it is our expectation that writers, editors and journalists will also use this new list.

5. Machine Translator

Another feature of ‘Taalweb’ is a statistical machine translation system called ‘Oersetter’. The system is able to translate from Dutch into Frisian and Frisian into Dutch and is intended to help non Frisian speaking people to understand Frisian. It is also a nice and easy way to create a basic translation, which can subsequently be edited with the other features of ‘Taalweb’.

'Oersetter' has been developed at the Radboud University Nijmegen. The translation system is built around the open-source, phrase-based SMT software Moses. The Fryske Akademy has compiled a Frisian–Dutch parallel corpus. After sentence-alignment, the corpus comprised a total of 44,503 sentence pairs, containing 701,782 words of Frisian and 673,277 words of Dutch, including punctuation marks. The monolingual corpus used to create a Frisian language model consists of 594,975 sentences and 10,043,516 words, making it considerably larger than the parallel corpus. The Frisian portion of the parallel corpus has also been included in the corpus that was used for the language model. The corpus contains texts from 1980 onwards. Though the FA tried to cover as many domains as possible, a major part of the corpus inevitably consists of literary texts. A more detailed description of the background of the machine translator can be found in Van Gompel et al. (2010). While testing the translation system, the results were satisfactory and encouraging. Frisian text generated with this translation system may be spell checked to see if it is in accordance with the standard wordlist.

6. Dictionary Portal

The third part of 'Taalweb' consists of a dictionary portal. As already indicated, the state of affairs concerning online lexicography in Friesland was not sufficient.² Back in 2010, the bilingual dictionaries Dutch–Frisian and Frisian–Dutch were partially available online. The interface provided only translations of headwords. Contexts, idioms, multi-word expressions and proverbs were absent. Unfortunately, the extensive monolingual dictionary, which was published in 2008, was not accessible online. The 25 volumes of the scholarly Dictionary of the Frisian Language were put online at the Euralex Congress in 2010.

The Fryske Akademy used custom-made dictionary writing software, which consisted of a simple text editor and a BRS/search database. It was a full-text database and information retrieval system which used a fully-inverted indexing system to store, locate, and retrieve unstructured data (Sijens and Depuydt, 2010). Furthermore, a non-tagged language database was available for dictionary compilation purposes. This corpus was established in the preceding decades and contained at that time some 24 million words. The available electronic lexicographic products were digitized versions of paper dictionaries. It is needless to say that we are not dealing here with proper electronic lexicography.

There was a lack of online dictionaries and there was also the desire to establish a new Dutch–Frisian dictionary. The most recent bilingual Dutch–Frisian dictionary was published in 1985, so it is rather outdated. Besides that, it contains too few examples to provide good, accurate and modern translations in Frisian. In order to

² Data are taken from the questionnaire response for Frisian, cf. Questionnaire, 2010.

fill the existing gap, a new project was conceived: a dictionary portal. This new online service contains several Frisian lexicographic products compiled in the years 1984 to 2008: a Frisian–Dutch dictionary (1984) with 56,000 entries, a Dutch–Frisian dictionary (1985) containing 53,000 entries, a juridical dictionary Dutch–Frisian (2000), which has 13,000 entries and finally a monolingual dictionary (2008) with 70,000 entries.

In addition to the dictionaries, the portal also contains a number of bilingual terminology lists ranging from administrative terms, through food terminology to terminology of school subjects such as geography, biology and physics. Newly compiled lists with terminology for these domains fill several lexical gaps.

The basic idea behind the dictionary portal was that linked information fields of the joint dictionaries and lists would provide much more useful information to the user than a stand-alone, digitized paper dictionary. For this purpose, the following information fields in the dictionaries were made searchable: headword, translations, idiom, synonyms and proverbs. Not every dictionary contains all fields, but that is hardly a problem. The bilingual dictionary Frisian–Dutch for instance lists more than 1,800 proverbs, a comprehensive list containing the most common Frisian proverbs. If a user is looking for a proverb, then what this dictionary provides will be sufficient to the user and the fact that the juridical dictionary does not deal with proverbs is no problem.

The recently published monolingual dictionary obviously lacked the field ‘translation’; however, since ‘translation’ is the main objective of the portal, we had to add an extra field with Dutch keywords to that dictionary.

One of the functions of the portal is to bridge the gap between the old bilingual dictionaries and a new Dutch–Frisian dictionary. The 1985 outdated Dutch–Frisian dictionary often lacks modern words that belong to the domains of computer science, modern media and sports. The dictionary portal is a cluster of lexicographical products which covers a period of almost thirty years, from 1984 until 2008. Often, when the old dictionaries fail to give a translation for a modern concept, the more recent ones offer a complement. The 1985 Dutch–Frisian dictionary has an entry for *kompjûter* (computer) but not a single compound with *kompjûter*-. The 2008 monolingual dictionary additionally offers 23 compound words with *kompjûter*.

The dictionary portal provides more translations and examples than a stand-alone online bilingual dictionary. Where the Dutch–Frisian dictionary of 1985 has its limitations, the linked dictionaries of the portal offer much additional information and many more possibilities. Take for example the Dutch adverb *vliegensvlug* (very quickly, at top speed). This one-word expression has no one-word equivalent in Frisian. The Dutch–Frisian 1985 dictionary translates the headword *vliegensvlug* with three multi-word expressions:

- *fleanende hurd*
- *mei kûgelsfeart*
- *as de reek*

While searching the entire database, including the field ‘idiom’, yields more hits:

- *as de duvel - vliegensvlug*
- *op in giseldraaf rinne - vliegensvlug draven*
- *dat giet der koers troch - dat gaat vliegensvlug, razendsnel.*
- *gean, rinne, fleane, jeie as it spoar - vliegensvlug gaan, lopen, draven, rijden.*
- *it giet, rint as it spoar - het gaat vliegensvlug*

All these matches are from the Frisian–Dutch dictionary of 1985, taken from the field of idioms. When translating from Dutch to Frisian, these additional alternatives for *vliegensvlug* can help people who want to write in Frisian, to create more varied and better texts. The standardized wordlist is envisaged in 2013. The lexicographic products that are part of the portal contain many words and variants that are not part of the standard language. Therefore, these works all have to be adapted to the standard wordlist once it will be official.

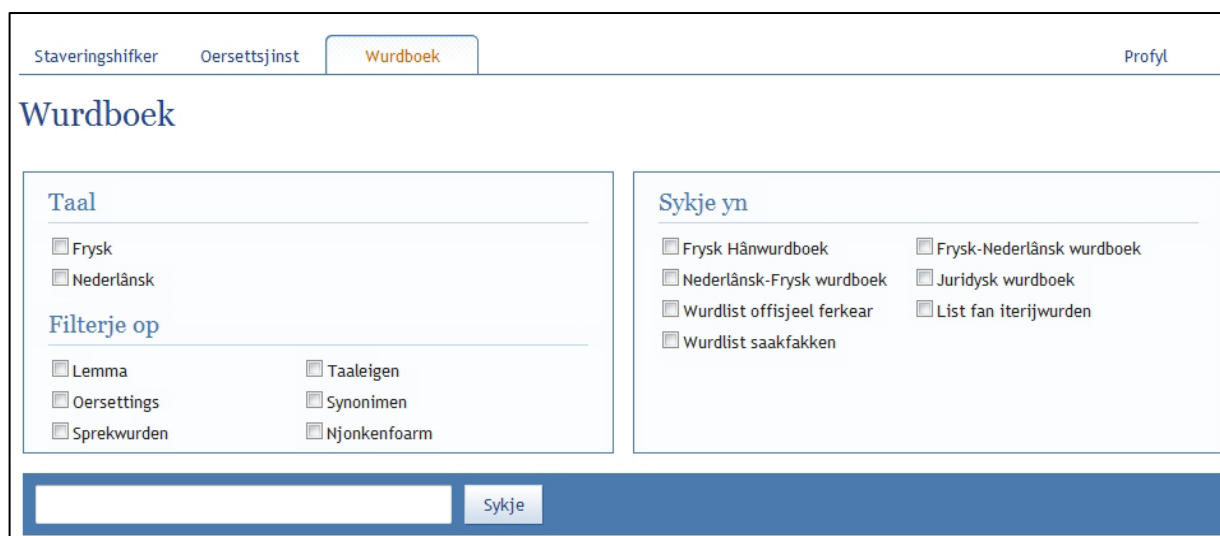


Figure 5: Opening screen ‘Dictionary Portal’

7. Future Dutch–Frisian Dictionary

The state of affairs of internet lexicography for Frisian has greatly improved with the completion and launch of the dictionary portal. However much information and new possibilities the portal offers, it will not be able to offer the same as a completely new online Dutch–Frisian dictionary can. This new dictionary will be a lexicographical database tailored to the needs of the user. To this end we will perform a study of

recent literature on both online bilingual dictionaries and dictionary use.

The target audience for this dictionary consists of learners and native speakers. The needs of two user groups have to be satisfied: Firstly, non-Frisian-speaking people will use it for translating Dutch into Frisian. Secondly, for native speakers the dictionary will be employed as an aid to using proper Frisian. Frisian speakers often have had too little mother tongue education, resulting in a lack of knowledge of their own language. In order to serve both user groups, the dictionary will offer them about 70,000 Dutch headwords with their standard Frisian equivalents. Its microstructure will contain many examples, multi-word expressions, phrases and idioms that will enable the users to produce proper and varied Frisian.

This project provides new opportunities to compile an up- to-date lexicographic database with a user-friendly interface. The new dictionary will be part of the Frisian Language database, a database system intended to open up eight centuries of Frisian. A demo version of the Frisian Language database can be accessed at <http://tdb.fryske-akademy.eu/tdb>.

8. Conclusion

For a small language community like Frisian, it is difficult to create a good lexicographical infrastructure. In his Euralex keynote lecture, Anne Popkema stated ‘Factors like magnitude of the language community and governmental recognition will be of influence on what medium a lexicographer chooses, since such factors for a considerable part determine the quintessential factor for any lexicographical endeavour: funds.’ (Popkema 2010: 87). In some way, this also applies to Friesland. The Fryske Akademy has only a small lexicographical staff at its disposal. As a scientific research center the academy is required to conduct high-quality research. This has yielded a scholarly dictionary of Modern Frisian. At the same time the academy is required to use the acquired knowledge about lexicography for the benefit of Frisian society. Therefore it is obvious that the Fryske Akademy should produce dictionaries and tools for the community within which it is part. With the financial support of the provincial government, it is possible to develop the required lexicographical infrastructure in cooperation with fellow institutes, universities and language technology supplied by IT companies. The ‘Taalweb’ with the dictionary portal is a step forward on this path. We hope to integrate the various tools in such a way that, for example, a user will be able to go from a misspelled form to the correct one via (selected) dictionary information.

Or, when offered automated translations, the user will be able to call up the relevant dictionary entries, so as to improve the automated suggestions. We like to think that even in its present state the ‘Taalweb’ will offer much to the professional user of Frisian and that it will be a quite useful tool for language learners, whether or not in the context of a language course.

9. References

9.1 General:

- Adamska-Sałaciak, Arleta (2013). Equivalence, Synonymy, and Sameness of Meaning in a Bilingual Dictionary. *International Journal of Lexicography* 26(3), pp. 329-345.
- Fuertes-Olivera, Pedro A. and Sandro Nielsen (2012). Online Dictionaries for Assisting Translators of Lsp Texts: The Accounting Dictionaries. *International Journal of Lexicography*, 25(2), pp. 191-215.
- Gompel, M. van, A van den Bosch, A. Dykstra (2013). Oersetter: Frisian - Dutch Statistical Machine Translation. In P. Boersma, H. Brand and J. Spoelstra (eds.) *Philologia Frisica anno 2012. Lêzings fan it njoggentjinde Frysk Filologekongres fan de Fryske Akademy op 13, 14 en 15 juny 2012*. Leeuwarden / Ljouwert: Afûk - Fryske Akademy (forthcoming).
- Gouws, Rufus H. (2013). Contextual and Co-Textual Guidance Regarding Synonyms in General Bilingual Dictionaries. *International Journal of Lexicography*, 26(3) pp. 346-361.
- Ilsou, R. (2013). The Explanatory Technique of Translation. *International Journal of Lexicography*, 26(3), pp. 386-393.
- Kwary, D.A. (2012). 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.
- Popkema, A.T. (2010). State of the Art of the Lexicography of European Lesser Used or Non-State Languages. In A. Dykstra and T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, (Leeuwarden, 6-10 July 2010)*. Ljouwert: Fryske Akademy - Afûk, pp. 65-98.
- Questionnaire (2010). *Questionnaire concerning lexicography of European lesser used languages. Q026 West Frisian* (unpublished).
- Sijens H. and K. Depuydt (2010). Wurdboek fan de Fryske taal / Dictionary of the Frisian Language Online: New Possibilities, New Opportunities. In A. Dykstra and T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, (Leeuwarden, 6-10 July 2010)*. Ljouwert: Fryske Akademy - Afûk, pp. 726-732.
- Taaltatlas 2011. *De Fryske taal atlas 2011. Fryske taal yn byld*. Ljouwert / Leeuwarden: provinsje Fryslân.

9.2 Dictionaries:

- Boersma, P. / K.F. van der Veen (1984-2011). *Wurdboek fan de Fryske Taal / Woordenboek der Friese Taal*. Ljouwert / Leeuwarden: Fryske Akademy. Accessed at: <http://gtb.inl.nl>.

- Duijff, P. (2000). *Juridisch Woordenboek Nederlands - Fries, met een index Fries - Nederlands*. Groningen / Leeuwarden: Martinus Nijhoff / Fryske Akademy.
- Duijff, P. en F.J. van der Kuip (2008). *Frysk Hânwurd- boek*. Leeuwarden: Fryske Akademy / Afûk.
- Visser, W. (1985). *Frysk Wurdboek 2, Nederlânsk - Frysk*. Leeuwarden: A.J. Osinga Uitgeverij.
- Zantema, J.W. (1984). *Frysk Wurdboek 1, Frysk - Nederlânsk*. Leeuwarden: A.J. Osinga Uitgeverij.

Can we determine the semantics of collocations without using semantics?

Pol Moreno¹, Gabriela Ferraro², Leo Wanner³

¹School of Informatics, University of Edinburgh,

Appleton Tower, 11 Crichton Street, Edinburgh EH8 9LE

²NICTA (National ICT Australia), London Circuit 7, Canberra City ACT 2601, Australia.

³Institució Catalana de Recerca i Estudis Avançats (ICREA) and

Department of Information and Communication Technologies, Pompeu Fabra University,
Roc Boronat, 138, 08018 Barcelona

E-mail: polmorenoc@gmail.com, gabriela.ferraro@nicta.com.au, leo.wanner@upf.edu

Abstract

The extraction of collocations from corpora has been actively worked on since the late eighties. However, so far, an important task of collocation processing, namely the semantic interpretation of the collocate, did not receive much attention, although the semantics of a given word when used as collocate very often varies from the semantics of the same word when used in a free co-occurrence. In this paper, we tackle this problem. Our aim is the automatic semantic disambiguation of collocates, or, more precisely, the classification of collocations with respect to the typology of lexical functions (LFs) introduced in the Explanatory Combinatorial Lexicology. The two main questions underlying our research that seeks a scalable solution independent of any external semantic resources are: (i) how well can we semantically classify collocates without the use of explicit semantic features; and (ii) to what extent can we dispense with explicit lexical information when classifying collocates. To answer these two questions, we carried out machine learning experiments in which we used different training feature sets and LF typologies of different abstraction. So far, we worked on Spanish verb-noun and noun-adjective collocations from the lexicographic field of emotion nouns. However, our approach is, strictly speaking, language-independent.

Keywords: collocations; semantics; lexical functions; classification

1. Introduction

The recognition and extraction of collocations from corpora has been actively worked on since the late eighties (e.g. Choueka, 1988; Church and Hanks, 1989; Smadja, 1993; Evert and Kermes, 2003; Kilgarriff, 2006; Evert, 2007; Pecina, 2008; Bouma, 2010; Wible and Tsao, 2010).¹ However, so far, an important task related to collocation recognition, namely the semantic disambiguation (or classification) of the collocate,²

¹ Not all of these works use the term “collocation”, but all of them nonetheless extract co-occurrent word combinations.

² Here and henceforth, we use the terminology as introduced by Hausmann (1989): the *base* is the semantic head of the collocation and the *collocate* is its dependent. Thus, in the collocation *strong tea*, *tea* is the base and *strong* is the collocate; in *take a rest*, *rest* is the base and *take* is the collocate, etc.

has received only very limited attention by the main stream research in the field. It is important to disambiguate the collocate because the semantics of a given word when used as a collocate very often differs from the semantics of the same word when used in a free co-occurrence. For instance, the meaning of *conduct* in *conduct an investigation* is different from its meaning in *conduct an orchestra* or in *conduct electric current*, and all three differ from its meaning as an isolated lexical item (as in *John conducted himself abominably*). Therefore, it is only when we know the meaning of the collocate in combination with the base that we can understand the meaning of the collocation as a whole and use it appropriately. This is also why in collocation dictionaries the collocates of a lemma are usually grouped according to their meaning and why automatic techniques for semantic classification of collocation collocates should be involved when, e.g., compiling collocation dictionaries from corpora.

In what follows, we tackle the problem of the semantic interpretation (or semantic disambiguation) of collocates. As in Wanner (2004), Wanner et al. (2005; 2006a; 2006b) and Gelbukh and Kolesnikova (2012), we use as reference classification the fine-grained semantic typology of collocations that underlies *lexical functions* (LFs) (e.g. Mel'cuk, 1995). Our goal is also the same: to be able to assign to the collocate of any given collocation in context a semantic class tag from the LF typology. However, unlike these previous works, which use external lexico-semantic resources (namely EuroWordNet; see Vossen, 1998), we aim to explore techniques that do not use any external resources and that are thus more scalable and universal. The two main questions underlying our research are: (i) how well can we semantically classify collocates without the use of explicit semantic features; and (ii) to what extent can we dispense with explicit lexical information when classifying collocates.

So far, we worked on Spanish collocations from the lexicographic field of emotion nouns. The corresponding corpus annotated with LFs has been provided to us by the DICE team of the Universidad de La Coruña (<http://www.dicesp.com>), Spain. We have chosen Spanish since, to the best of our knowledge, only for Spanish an LF-annotated corpus is available. However, as will become clear from the presentation below, our approach is to a large extent language-independent.

In the next section, we briefly introduce the LF typology. Section 3 outlines the experiments we carried out to assess to what extent the classification of LF instances in the corpus is feasible by exclusively using features encountered in the textual context of these instances. Section 4 comprises a discussion of the outcome of these experiments. Section 5, finally, summarizes the insights we obtain and outlines the directions of our future work on this topic.

2. On the Semantic Collocate Typology

Earlier approaches to collocation extraction from corpora tended to consider any pair

of tokens that shows a significant co-occurrence tendency (a *strong association norm* in terms of Church and Hanks, 1989) to be a collocation, with the consequence that the result lists contained such pairs as *doctor – nurse*, *professor – university*, or *smoker – cigarette*; see, e.g., (Choueka, 1988; Church and Hanks, 1989). While being useful, for instance, for the construction of relational lexica, these pairs do not find their way into collocation dictionaries since they are not, strictly speaking, collocations. Nor can they be used in such tasks as lexicalization in Natural Language Text Generation, where lexical co-occurrence resources have shown to be of great value (e.g. Wanner, 1997).

Most of the more recent collocation extraction strategies have corrected this generous interpretation of co-occurrence and handle only word occurrences that form valid syntactic structures (Smadja, 1993; Evert and Kermes, 2003; Kilgarriff, 2006).³ But this is not the end of the story: between the base and the collocates of a collocation not only a syntactic but also a semantic relation holds. This relation is often of abstract nature, such that it applies to a large number of collocations. For instance, the same relation can be said to hold between *speech* and *deliver*, *suicide* and *commit*, *step* and *take*, etc. It is the same in the sense that *deliver*, *commit*, and *take* contribute to their respective base the same semantic features. A possible label for these features is ‘perform’. Obviously, the same label can be used to tag the meaning of *deliver*, *commit*, and *take* in these co-occurrences. The typology of lexical functions (LFs) as proposed in the framework of the Explanatory Combinatorial Lexicology (ECL) (Mel’cuk, 1995) captures this kind of semantic relations between the elements of collocations. The typology consists of about 30 classes of the type ‘perform’, ‘react’, ‘begin to perform’, ‘continue to perform’, ‘take place’, ‘originate from’, ‘become involved’, ‘intense’, ‘positive’, etc.

Table 1 displays, for illustration, examples for ten of these classes. In the first column, we add in parentheses the names of the LFs (Latin abbreviations) as used in the ECL literature and as we will use for the sake of brevity in the paper.

The LF typology is not the only semantic classification of collocates used in lexicography. As already mentioned above, all major collocation dictionaries tend to group collocates of a given lemma in accordance with semantic criteria. Consider, e.g., a fragment of the entry for INITIATIVE in the *Oxford Collocations* dictionary:

undertake | plan | develop | announce | introduce, launch, set up, start | become involved | lead | approve | reject | sponsor | endorse, support ...

where ‘|’ separates the semantic groupings of collocates.

³ However, we obviously acknowledge that some researchers prefer to continue to work in the Firthian tradition of the term “collocation” and interpret any pair of tokens which co-occur with statistical significance as collocation. We think that both interpretations can cohabit as long as the authors clearly state the notion that they adopt.

Parallels of this grouping to (an abstracted) LF typology cannot be overlooked. Therefore, we have decided to use the following as reference typologies: (a) the genuine LF typology, because of its clear formal definition and potential of systematic abstraction; and (b) a generalized LF typology which is in its nature very similar to the implicit typologies used in broad distribution collocation dictionaries.

‘act’/‘perform’ (Oper1)	<i>take</i> – walk, <i>give</i> – talk, <i>hold</i> – reception
‘undergo’/‘meet’ (Oper2)	<i>receive</i> – blow, <i>encounter</i> – obstacle, <i>run into</i> – resistance
‘act accordingly’ (Real1)	<i>succumb to</i> – illness, <i>win</i> – match, <i>keep</i> – promise
‘originate from’ (Func1)	blow – <i>come from</i> , proposal – <i>stem from</i> , analysis – <i>be due to</i>
‘be fulfilled by’ (Fact1)	illness – <i>carry off</i> , benefit – <i>proceeds</i> , generosity – <i>pay off</i>
‘begin to act/ perform’ (IncepOper1)	<i>open</i> – dispute, <i>fall in</i> – love, <i>enter</i> – war
‘begin to originate from’ (IncepFunc1)	hatred – <i>come over</i> , panic – <i>seize</i> , routine – <i>catch up with</i>
‘become more intense’ (IncepPredPlus)	love – <i>grow</i> , voice – <i>become louder</i> , debate – <i>heat up</i>
‘reduce intensity’ (CausPredMinus)	<i>ease</i> – shortage, <i>contain</i> – inflation, <i>alleviate</i> – pain
‘intensify’ (CausPredPlus)	<i>increase</i> – pressure, <i>augment</i> – presence, <i>steer up</i> – hatred

Table 1: Samples of semantic classes of the LF typology (the collocates are in italics)

3. Experiments

In order to assess to what extent it is possible to identify the semantic labels of collocates in context, we carried out a series of experiments in which we interpreted the task of the semantic label identification as a machine learning-based classification task. As already mentioned above, others (e.g. Wanner, 2004; Wanner et al., 2006a,b) address the same problem using semantic features of the collocation elements from EuroWordNet (Vossen, 1998) to assess the similarity of a candidate co-occurrence with the samples of each given LF class. However, we do not use any external resources. Rather, we intend to explore to what extent semantic knowledge-poor techniques similar to those used for the extraction of collocations can be used for this purpose. In the case of a positive outcome, we furthermore want to explore: (i) whether these techniques also serve for the classification of collocations with respect to a generalized LF typology (of the kind found in broad coverage collocation dictionaries such as the *Oxford Collocations Dictionary* or

McMillan Collocation Dictionary); and (ii) whether lexical features (i.e., concrete words) are crucial for the classifier accuracy, or in other words, how semantic field-specific the classifier needs to be. ⁴

3.1 Setup of the experiments

For our experiments on the classification with respect to the genuine LF typology, we focused on the ten LFs listed in Table 1. Table 2 displays the number of samples of each LF in the DICE corpus.

Collocate class	#
Oper1	1470
Oper2	149
Real1	147
Func1	179
Fact1	160
IncepOper1	152
IncepFunc1	244
IncepPredPlus	201
Caus Pred Minus	409
Caus Pred Plus	301

Table 2: Number of samples of each collocate class in the DICE corpus

For the experiments on a generalized fragment of the LF typology, we used five generic collocation categories proposed by colleagues from La Coruña; the generalization, including the subcategories of the general semantic categories, is displayed in Table 3. For readers interested in the actual LFs that compose the categories, they are listed in the Appendix.

For the classification experiments with respect to both typologies, we used the Weka machine learning environment, together with the LibSVM implementation. A linear kernel was chosen to generate the Support Vector Machine (SVM) models since it proved to be adequate for text classification tasks, which usually need to cope with a high amount of features. The following features were used:

- *Lexical features*: all tokens in the sentence + base + collocate + base-collocate pair.⁵
- *POS-features*: POS of the base + POS of the collocate + POS of the tokens in the windows of size 2 to the left and to the right of the base and the collocate +

⁴ Recall that the DICE corpus contains only collocations from the field of emotions.

⁵ In one of the experiments (see below), we suppressed the base and the base-collocate pair from feature set.

POS-trigrams of the POS of the base and the POS of its immediate left and right context + POS-trigrams of the POS of the collocate and the POS of its immediate left and right context.

- *Morphological features*: gender, number, person of the base + number, person, tense, and mode of the collocate + POS pairs of the syntactic dependents of the base and the POS of the base + POS pairs of the POS of the syntactic head of the collocate and the POS of the collocate + POS pairs of the POS of the collocate and the POS of all its remaining dependents.
- *Syntactic dependency features*: syntactic relation between the collocate and the base + syntactic relation between the collocate and its head + syntactic relations between the collocate and its remaining dependents + syntactic relations between the base and its dependents.

Semantic category	Subcategory	# of instances
Intensity	‘high intensity’	50
	‘intensity increase’	491
	‘intensity decrease’	468
Phase	‘preparation’	14
	‘initiation’	406
	‘continuation’	309
	‘termination’	523
Manifest	‘manifestation’	1062
	‘lack of manifestation’	407
Cause	‘causation’	1001
Experimenter	‘experimentation’	1478

Table 3: Fragment of the generalized LF typology

The POS and the morphological and syntactic dependency features were obtained by parsing the corpus with Bohnet’s (2009) syntactic dependency parser.⁶ We trained 10 binary classifiers on separate positive and negative corpora for each of the ten LFs. In the positive corpus, each sentence contained at least one collocation whose collocate was an instance of the given LF. The negative corpus consisted of the sentences with occurrences of the other LFs. Due to the high amount of negative class instances compared to the positive instances, we balanced each set by under-sampling the majority class.

⁶ This parser performed best on Spanish in the CoNNL 2009 shared task.

O1	tener ‘have’ – admiración ‘admiration’, tributar ‘tribute’ – respeto ‘respect’, experimentar ‘experience’ – disgusto ‘annoyance’, tener ‘have’ – pudor ‘modesty’, sentir ‘feel’ – bochorno ‘embarrassment’, pasar ‘pass’ – apuro ‘rush’, abrigar ‘nourish’ – ilusión ‘illusion’
O2	gozar ‘enjoy’ – admiración ‘admiration’, recibir ‘receive’ – consideración ‘consideration’, gozar ‘enjoy’ – respeto ‘respect’, sufrir ‘suffer’ – desprecio ‘contempt’, tener ‘have’ – sorpresa ‘surprise’
R1	disfrutar ‘enjoy’ – felicidad ‘happiness’, degustar ‘taste’ – felicidad ‘happiness’, morir ‘die’ – [de ‘of’] pena ‘pity’, aplicar ‘apply’ – pena ‘sentence’, sucumbir ‘succumb’ – [al ‘to’] miedo ‘fear’
Fu1	desprecio ‘contempt’ – anidar ‘nest’, alborozo ‘joy’ – reinar ‘reign’, satisfacción ‘satisfaction’ – reinar ‘reign’, felicidad ‘happiness’ – sonreír ‘smile’, desazón ‘discomfort’ – asaltar ‘assault’
Fa1	tristeza ‘sadness’ – sacudir ‘shake’, pena ‘pity’ – comer ‘eat’, desazón ‘discomfort’ – quemar ‘burn’, temor ‘fear’ – paralizar ‘paralyze’, aprensión ‘apprehension’ – atezar ‘grip’, aflicción ‘grief’ – azotar ‘hit’
IO1	aversión ‘aversion’ – tomar ‘take’, caer ‘fall’ – [en ‘in’] abatimiento ‘disheartenment’, coger ‘catch’ – miedo ‘fear’, cobrar ‘gain’ – miedo ‘fear’, tomar ‘take’ – aprensión ‘apprehension’
IF1	sentimiento ‘feeling’ – invadir ‘invade’, tristeza ‘sadness’ – entrar ‘enter’, desazón ‘discomfort’ – asaltar ‘assault’, miedo ‘fear’ – aparecer ‘appear’, pasmo ‘amazement’ – dar ‘give’, odio ‘hatred’ – surgir ‘surface’
IPP	admiración ‘admiration’ – aumentar ‘augment’, respeto ‘respect’ – crecer ‘grow’, esperanza ‘hope’ – aumentar ‘augment’, angustia ‘distress’ – crecer ‘grow’, amistad ‘friendship’ – intensificar ‘intensify’
CP M	enfriar ‘freeze’ – entusiasmo ‘enthusiasm’, aliviar ‘alleviate’ – desprecio ‘contempt’, paliar ‘palliate’ – sentimiento ‘feeling’, mermar ‘diminish’ extrañeza ‘estrangement’, frenar ‘brake’ – euforia ‘euphoria’
CPP	aumentar ‘augment’ – respeto ‘respect’, reafirmar ‘reaffirm’ – entusiasmo ‘enthusiasm’, intensificar ‘intensify’ – desprecio ‘contempt’, avivar ‘enliven’ – aversión ‘aversion’, promover ‘promote’ – bienestar ‘well-being’

Table 4: Correctly classified individual LF instance samples (‘O1’ = Oper1, ‘O2’ = Oper2, ‘R1’ = Real1, ‘Fu1’ = Func1, ‘Fa1’ = Fact1, ‘IO1’ = IncepOper1, ‘IF1’ = IncepFunc1, ‘IPP’ = IncepPredPlus, ‘CPM’ = CausPredMinus, ‘CPP’ = CausPredPlus)

For the experiments that targeted the exploration of the semantic field specificity of the classification, we had removed the lexical features from the feature lists.

3.2 Results of the experiments

Due to the context-driven nature of our classification procedure, classification examples should, in fact, always be shown together with their context rather than in isolation. However, in order to keep our presentation as clear and as simple as possible, we nonetheless cite in Tables 4 and 5 a few examples of the output of our LF classification in isolation. Table 4 illustrates some correctly classified samples of individual LFs. Table 5 below displays some of the correctly classified samples of the generalized LF typology.

I	sentir ‘feel’ – admiración ‘admiration’, rebajar ‘reduce’ – exasperación ‘exasperation’, aumentar ‘augment’ – bienestar ‘well-being’, aplacar ‘appease’ – ira ‘anger’, mitigar ‘mitigate’ – nostalgia ‘nostalgia’
P	sospecha ‘suspicion’ – persistir ‘persist’, conservar ‘conserve’ – desapego ‘indifference’, desesperación ‘desperation’ – invadir ‘invade’, cariño ‘affection’ – desaparecer ‘disappear’, vergüenza ‘shame’ – entrar ‘enter’
M	testimoniar ‘testify’ – afectar ‘affect’, satisfacer ‘satisfy’ – orgullo ‘pride’, ocultar ‘hide’ – pudor ‘chastity’, expresar ‘express’ – admiración ‘admiration’, contener ‘control’ – desencanto ‘disappointment’
C	ahogar ‘drown’ – pena ‘pity’, despertar ‘wake up’ – encono ‘lingering anger’, conseguir ‘achieve’ – excitación ‘excitation’, suscitar ‘stimulate’ – resentimiento ‘resentment’, causar ‘cause’ – aprensión ‘apprehension’
E	constituir ‘form’ – felicidad ‘happiness’, sentir ‘feel’ – alegría ‘joy’, tener ‘have’ – despreocupación ‘disregard’, abrigar ‘harbor’ – ilusión ‘illusion’, poseer ‘possess’ – temor ‘fear’

Table 5: Correctly classified generalized LF instance samples (‘I’ = Intensity, ‘P’ = Phase, ‘M’ = Manifest, ‘C’ = Cause, ‘E’ = Experimenter)

If a sample occurs in the corpus several times (which is usually the case), each occurrence is analyzed separately, such that the same sample may be classified differently in different contexts. Sometimes, this is incorrect. Consider, e.g.:

- 1) ... *por ser oral fundamentalmente, ser transmitida de generación en generación que aumenta el apego del pueblo a su propia lengua...* ‘for being basically oral, being transmitted from generation to generation, which strengthens the attachment of the people to their own language’
- 2) ... *a medida que aumenta el apego al cuerpo, el sufrimiento también aumenta* ‘as the attachment to the body increases, the suffering also increases’

In both (1) and (2), *aumentar* – *apego* ‘increase – attachment’ is an instance of IncepPredPlus. However, in (1) it has been erroneously classified as CausPredPlus. On the other hand, the distribution-based classification procedure is sensitive to

fine-grained features that are decisive for the distinction between semantically very similar LFs. Thus, in (3), *augmentar – admiración* ‘increase – admiration’ is an instance of IncepPredPlus, while in (4), the same co-occurrence is an instance of CausPredPlus, such that multiple classification seems necessary.

- 3) *Su admiración aumenta al recordar la naturalidad con que se dirige a su marino* ‘His admiration increases when he remembers the naturalness with which he talks to his seaman’.
- 4) *... tiene uno buen caldo de cultivo para aumentar su admiración por la hasta entonces controvertida figura del cretense* ‘... has a fertile breeding ground to augment his admiration for the until then controversial figure of the Cretan’.

The classification procedure correctly classifies the two co-occurrences.

3.3 Evaluation

To test the accuracy of our classifier models, we used a 10-fold cross-validation scheme. Tables 6 and 7 display the results of the classification obtained with respect to the genuine LF typology and the generalized LF typology, respectively.

The second and third columns in Table 6 show the results obtained with classifiers trained on the complete set of features; the fourth and fifth columns show the results obtained with classifiers trained on a set of features that did not contain the lexical tokens of the base. In the second and fourth columns, the accuracy of the classification of a given collocation as the LF in question is indicated; in the third and fifth, the accuracy of the recognition that a given collocation is not an instance of the LF in question is provided.

LF class	F-score (all features)		F-score (no lex. base feature)	
	+	-	+	-
CausPredMinus	0.90	0.99	0.68	0.89
CausPredPlus	0.84	0.98	0.57	0.79
Fact1	0.76	0.99	0.63	0.83
Func1	0.72	0.98	0.61	0.81
IncepFunc1	0.88	0.99	0.55	0.75
IncepOper1	0.85	0.99	0.65	0.86
IncepPredPlus	0.85	0.99	0.68	0.87
Oper1	0.91	0.95	0.64	0.80
Oper2	0.58	0.98	0.52	0.80
Real1	0.69	0.99	0.48	0.76

Table 6: Classification results per LF

For the two LFs with larger numbers of samples, Oper1 and CausPredMinus, we also performed an evaluation with a data split. For this purpose, we split the corresponding corpora into training and testing sets, with an 80% to 20% ratio (using the full set of features). For Oper1 classification, we then obtained a weighted average F-score of 0.93 and for the CausPredPlus an average F-score of 0.97. This is comparable with the performance obtained with the 10-fold cross-validation. For smaller samples, a data split proved to have negative consequences since the training sets of 80% were too small.

Table 7 displays the precision and recall figures of the classification with respect to the generalized LF typology with and without lexical features.

LF class	all features		no lex. base feature	
	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>
Intensity	0.947	0.917	0.338	0.388
Phase	0.887	0.909	0.387	0.30
Manifest	0.925	0.904	0.367	0.446
Cause	0.82	0.828	0.442	0.346
Experimenter	0.906	0.92	0.538	0.567

Table 7: Classification results per generalized LF category ('p' = precision; 'r' = recall).

4. Discussion of the Evaluation

4.1 Classification using the LF typology

Table 6 shows that when using the full set of features, i.e., including the lexeme of the base, the classification with respect to the full-fledged LF typology achieves rather high accuracy scores (ranging from 0.58 for the recognition of Oper2-instances to 0.91 for the recognition of Oper1-instances); the variation of the accuracy is first of all due to the varying size of the training sets. The classification of negative instances is even better (between 0.95 and 0.99). This high accuracy is likely to be motivated by the distribution of the collocates of the collocations in a given semantic field (recall that we are dealing with a corpus on emotions here): in accordance with the Zipf law, a small number of collocate lexemes is very frequent, while the large rest occurs with a very limited frequency. Consider, for illustration, Table 8, where the share of the three most frequent collocates for four LFs in the DICE-corpus is given. It remains to be verified whether similar distributions can be observed in other semantic fields; our working hypothesis is that this is the case.

In the light of this distribution, an interesting research question is to what extent semantic field features influence the accuracy of the classification. Since the base lexemes are the most prominent features of a field (in our case, emotion nouns), the outcome of the second experiment in which we removed them from the feature lists is

of relevance; cf. columns 4 and 5 in Table 6. The accuracy is lower for all LFs, but not to an extent that would suggest that for each semantic field, separate collocate classifiers must be used. Since in both experiments positive instance classification turned out to be less accurate than negative instance classification, we focused in our error analysis on false positives.

Oper1	Freq.	Real1	Freq.
tener ‘have’	26.80%	descargar ‘unload’	9.52%
sentir ‘feel’	20.74%	dar ‘give’	8.84%
ser ‘be’	8.57%	disfrutar ‘enjoy’	7.48%
Total	56.11%		25.85%
CausPredMinus		CausPredPlus	
aplacar ‘soothe’	12.46%	aumentar ‘augment’	34.21%
mitigar ‘moderate’	10.02%	acrecentar ‘increase’	8.97%
aliviar ‘alleviate’	9.53%	avivar ‘brighten up’	7.64%
Total	32.01%		50.85%

Table 8: Collocate lexeme distribution in the DICE corpus

Table 9 shows the performance statistics for the classification with respect to four of the LFs using the complete set of features.

LF	\# Corr.	\# Inc.	\# FP
Oper1	4039	262	153
Real1	4220	81	23
CausPredPlus	4205	96	76
CausPredMinus	4228	73	39

Table 9: Error statistics in the individual LF classification

The second column contains the number of correctly classified instances (Corr.), the third the number of incorrectly classified instances (Inc.), and the fourth indicates how many of the incorrectly classified instances are false positives (FP).

A more detailed analysis reveals the following major confusion figures shown in Table 10.

Oper1:	Func 1 (36), Incep Pred Plus (31)
Real1:	Oper1 (6), Real2 (5)
CausPredPlus:	IncepPredPlus (35), CausPredMinus (6)
CausPredMinus:	IncepPredMinus (18), CausPredPlus (17)

Table 10: Classification confusion figures

As expected, the classifiers more commonly confuse LF-instances with very similar syntax. Consider, for instance, Real1 vs. Oper1 vs. Real2: here, we need to capture the semantic difference between, e.g., *keep a promise* vs. *give a promise* vs. *hold / fulfill to a promise* – which is hard, although not impossible, using the distributional semantic features we exploited so far. The confusion in the case of CausPredPlus and CausPredMinus is analogous, but still more subtle and thus more difficult to capture: the difference between CausPredPlus respectively CausPredMinus and the LFs with which they are confused consists of a few deep semantic features (‘begin to increase’ vs. ‘increase’, ‘decrease’ vs. ‘increase’, etc.). Thus, for example, many of the instances of CausPredPlus that have been classified as IncepPredPlus contain the collocate *augmentar* ‘augment’; see above, and these examples:

augmentar – placer ‘pleasure’, *augmentar* – confusión ‘confusion’, *augmentar* – sensación ‘sensation’, *augmentar* – admiración ‘admiration’, *augmentar* – abatimiento ‘disheartenment’

4.2 Classification using the generalized typology

A comparison of the figures in Tables 6 and 7 reveals that the balanced F-score achieved during the classification with respect to the generalized LF-typology is persistently higher than the average F-score across the individual LFs that constitute the generalized categories. For instance, the average F-score for recognition of the instances of the three LFs CausPredPlus, IncepPredPlus, and CausPredMinus using lexical features is 0.863, while the recognition of instances of ‘Intensity’ (which includes, among others, the above three LFs) achieves an F-score of 0.932. This can be interpreted as a sign of quality of the generalized LF-typology: similar LFs that were still confused in the individual LF classification exercise have been gathered into (more) homogeneous semantic categories, with clearer (first of all lexical) discrimination boundaries. However, with the generalized typology confusions obviously also occur. The corresponding confusion matrix in Table 11 reveals that ‘Intensity’ is confused more with ‘Phase’ than with other categories, ‘Phase’ and ‘Manifest’ with ‘Cause’, ‘Cause’ with ‘Experimenter’ and vice versa. The confusions can be explained by a more detailed analysis of the composition of the generalized categories, or, in other words, by the proximity of the individual LFs that compose the categories. Since this would imply a detailed introduction to the LFs, we refrain from such an analysis here. For the convenience of readers who are familiar with LFs, we provide the list of LFs of which each category is composed in the Appendix.

	I	P	M	C	E
Intensity (I)	944	41	19	16	9
Phase (P)	17	1212	30	48	27
Manifest(M)	19	45	1415	59	28
Cause (C)	11	39	42	985	94
Experimenter (E)	6	29	24	73	1521

Table 11: Confusion matrix in the generalized classification with lexical features

In contrast to the generalized classification which uses lexical features, the classification in which no lexical features have been used cannot compete with individual LF classification; cf. the p and r figures in columns 4 and 5 of Table 7: both precision and recall are considerably lower. The lack of lexical features penalizes the classification with respect to the generalized LF typology more than it does with respect to the individual LF typology. The confusion matrix in Table 12 shows that the confusion patterns also change. Thus, while ‘Intensity’ is still mostly confused with ‘Manifest’, ‘Phase’ is now confused most often also with ‘Manifest’ and not with ‘Cause’, ‘Manifest’ with ‘Experimenter’, etc. This is because the syntactic and contextual features of the LFs between these categories are more similar than are the lexical features. A more detailed study is needed to improve on the overall accuracy of generalized classification without the use of lexical features.

	I	P	M	C	E
Intensity (I)	395	150	337	58	98
Phase (P)	254	400	343	126	211
Manifest(M)	250	217	693	127	279
Cause (C)	131	119	238	374	219
Experimenter (E)	138	147	277	161	930

Table 12: Confusion matrix in the generalized classification without lexical features

5. Conclusions and Future Work

We have presented an excerpt of ongoing work on the semantic classification of collocates, which has until now been a largely neglected aspect of collocation processing but which we believe to be very important. To the best of our knowledge, the only existing works on the problem are those presented in Gelbukh (2012), Wanner et al. (2006a,b) and Wanner (2004). In contrast to these previous works, we do not use any external semantic resources and thus avoid two major disadvantages: (i) that the results could be negatively affected by the incompleteness and bias of the Spanish EuroWordNet towards English; and (ii) that an external semantic resource

for a specific language could limit the scalability and porting of the developed tool to other languages. Thus, our approach is much more flexible. The results obtained so far using the corpus of emotions and the genuine LF typology as reference typology are very encouraging, particularly if we take into account that the LF typology is very fine-grained. The preliminary experiments on the generalized LF typology need to be further extended since they have the potential to provide rich (and already appropriately grouped) input material for general public collocation dictionaries. In the immediate future, we plan to extend our experiments to generic corpora and to combine collocate classification with collocation identification, such that automatic semantic labeling of collocates in corpora becomes a realistic task.

6. Acknowledgements

The research reported in this paper has been partially funded by the Spanish Ministry of Economy and Competition (contr. 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; many thanks to Margarita Alonso Ramos and Orsolya Vincze for their support. We are also grateful to two anonymous reviewers for their helpful comments. At the time of the reported research, the first and second authors were members of the NLP group, Department of Information and Communication Technologies, UPF.

7. References

- Bohnet, B. (2009). Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*.
- Bouma, G. (2010). Collocation extraction beyond the independence assumption. In *Proceedings of the ACL 2010, Short paper track*. Uppsala.
- Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pp. 34–38.
- Church, K.W. & P. Hanks. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pp. 76–83.
- Evert, S. & H. Kermes. (2003). Experiments on candidate data for collocation extraction. *Companion Volume to the Proceedings of the 10th Conference of the EACL*. 83–86.
- Evert, S. (2007). Corpora and collocations. In *Corpus Linguistics. An International Handbook* ed. by A. Lüdeling & M. Kytö. Berlin: Mouton de Gruyter.
- Gelbukh, A. & O. Kolesnikova. (2012). *Semantic Analysis of Verbal Collocations with Lexical Functions*. Springer.

- Hausmann, F.-J. (1989). Le dictionnaire de collocations. In Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds). In *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin/ New York: De Gruyter. 1010-1019.
- Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress*. Torino.
- Mel'čuk, I.A., (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: Benjamins, 37-102.
- Mel'čuk, I.A. (1995). Phrasemes in Language and Phraseology in Linguistics. In *Idioms: Structural and Psychological Perspectives* ed. by M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder. 167–232. Hillsdale: Lawrence Erlbaum Associates.
- Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 54–57. Marrakech.
- Smadja, F. (1993). Retrieving Collocations from Text: X-Tract. *Computational Linguistics*.19.1:143–177.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.
- Wanner, L. (2004). Towards Automatic Fine- Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*. 10.2:95–143.
- Wanner, L. (1997). *Exploring Lexical Resources for Text Generation in a Systemic Functional Language Model*. PhD Dissertation. Universität des Saarlandes.
- Wanner, L., B. Bohnet & M. Giereth. L. (2006a) 'What Is Beyond Collocations? Insights from Machine Learning Experiments'. In *Proceedings of the EURALEX Conference*, Turin.
- Wanner, L., B. Bohnet & M. Giereth. (2006b). Making Sense of Collocations. *Computer Speech and Language*. 20.4:609–624.
- Wible, D. & N.L. Tsao. (2010). Stringnet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL-HLT Workshop on Extracting and Using Constructions in Computational Linguistics*. Los Angeles.

8. Appendix

The following table shows the composition of the generic LF categories by individual LFs. For definitions of the LFs, see, e.g. Mel'čuk (1996).

Semantic category	LF (# of instances)
Intensity	Magn+Oper1 (48), Magn+ Caus1Manif (2), CausPredPlus (292), IncepPredPlus (199) CausPredMinus (412), IncepPredMinus (63)
Phase	PreparReal1 (7), IncepOper1 (129), IncepFunc1 (234), Magn + IncepFunc1 (43), ContOper1 (94), CausContFunc0 (82), CausContFunc1 (1), ContFunc0(80), ContFunc1 (52), FinOper1 (113), LiquOper1 (36), Liqu1Func0(256), FinFunc0 (109), FinFunc1 (9)
Manifest	CausManif (610), AntiVer+Caus1Manif (6), Magn+Caus1Manif (2), Caus1Manif (2), Conv21Manif (86), IncepManif (22), PredA1Manif (6), Perm1Manif (3), Real1 (141), Caus1Func2 (5), Mang+Caus1Func2 (1), Fact1 (148), Magn+Fact1 (32), nonPermFact0 (96), nonPerm1Manif (261), nonFact1 (2), AntiReal1 (48)
Cause	V (155), CausFunc0 (186), MagnCausFunc0 (1), Caus2Func1(200), Caus2Func2(116), CausOper1 (102), Magn+CausOper1 (39), Func3 (18), Oper2 (143), Plus+Oper2 (1), Real2 (49)
Experimenter	Oper1 (1311), nonOper1 (3), Func1 (164)

Online Platform for Extracting, Managing, and Utilising Multilingual Terminology

**Mārcis Pinnis, Tatiana Gornostay,
Raivis Skadiņš, Andrejs Vasiļjevs**

Tilde, Vienības gatve 75a, Rīga, Latvia, LV-1004
{marcis.pinnis, raivis.skadins, tatiana.gornostay, andrejs}@tilde.lv

Abstract

In this demonstration paper we present an innovative platform “Terminology as a Service” (TaaS) for acquiring raw terminological data, and cleaning up, sharing, and reusing them, based on cloud computing. The platform serves, among other things, the needs of specialised lexicography. The proposed solution aims to fill the gap of collaborative terminology management and effective sharing of existing terminological data thus speeding up the development of specialised dictionaries. It also aims to build a bridge for the reuse of existing terminology between different groups of users, e.g. human users, such as lexicographers, translators, terminologists, and others, and machine users, such as computer-assisted translation tools, machine translation systems, third party terminology management solutions, and others.

Keywords: specialised lexicography, terminography, specialised dictionary, terminology service

1. Introduction

Lexicography, as the theory and practice of dictionary development, is one of the most labour-intensive human activities in the field of linguistics. The creation of a new dictionary from scratch and its delivery to an end user requires considerable resources in terms of time, man power, and finance. The main drawback of a conventional “paper” dictionary is its static and out-of-date content. For example, a particular paper terminological dictionary was already out-of-date by about 5–6 years when it was published and distributed (Shaikevich, 1983). In specialised lexicography, or terminography, it is even more critical since terminology is developing rapidly along with its subject field, or domain, and science in general.

Accurate handling of terminology is dramatically important in any professional language work—domain expertise, terminological analysis, documentation authoring and translation, professional (corporate and industry) communication, brand and product management, and other processes.

A paper terminological dictionary is somewhat a static fragment of a certain subject field in a certain language at a certain period of time.

To overcome the shortcomings of conventional lexicography, an electronic

punch-card machine was first used to create a prototype of a modern electronic dictionary by Roberto Busa in the XXth century. His first work was based on the automatic linguistic analysis (lemmatisation) of the works of Saint Thomas Aquinas. Roberto Busa compared the invention of an “electronic book” (instead of a printing book) to the introduction of a printing book by Gutenberg (instead of a manuscript) (Busa, 1961). Since that time automated lexicography has been developing rapidly.

Nowadays, with the evolution of information technologies, the Internet, and data (e.g., open data on the Web, free parallel and comparable corpora, and many other resources), the task of automated, or computational, specialised lexicography becomes a priority. Routine processes have been delegated to a computer. An electronic, or computer-based dictionary is easy to update and manage, and its main advantage is its flexible, dynamic, and extensible (e.g., in terms of new languages) character. Moreover, the new era of information technologies offers new ways of dictionary representation, e.g., on tablet, mobile, and other devices, and the usage patterns of a dictionary are changing with the course of time.

Lexicographers can have access to data and process them – analyse, tag, extract information etc. The integration of natural language processing tools have made it possible to grammatically and semantically analyse and tag a text and then to extract required pieces of information from the text. In the specialised lexicography, or terminography, developers can analyse and extract term candidates for further processing, e.g., automatic clean-up, sharing, and reusing in further processing and/or other applications (see section 3 and 4 below). Thus it has become possible to consider hundreds of thousands of terms specific to a certain domain in comparison with that time when only several thousands, usually no more than 2000, could be included in a conventional dictionary.

In this demonstration paper we present an innovative cloud-based platform “Terminology as a Service” (TaaS) for acquiring raw terminological data, cleaning it up, sharing and reusing terminological data cleaned up by users. The platform serves, among other things, the needs of specialised lexicography.

The proposed solution aims to fill the gap of collaborative terminology management and effective sharing of existing terminological data and thus speeding up the development of specialised dictionaries. It also aims to build a bridge for the reuse of existing terminology between different groups of users, e.g., human users, such as lexicographers, translators, and terminologists (human-oriented specialised dictionaries), as well as machine users, such as computer-assisted translation (CAT) tools, machine translation (MT) systems, third party terminology management solutions etc. (machine-oriented specialised dictionaries). The paper is structured as follows: section 2 provides a brief overview of the TaaS platform, section 3 describes the workflow for the creation of a bilingual terminological collection from user-provided documents, terminology sharing and reusing possibilities offered by

the platform are outlined in section 4, and available interfaces for machine users are briefly drafted in section 5. Finally, the paper is concluded and future work is outlined in section 6.

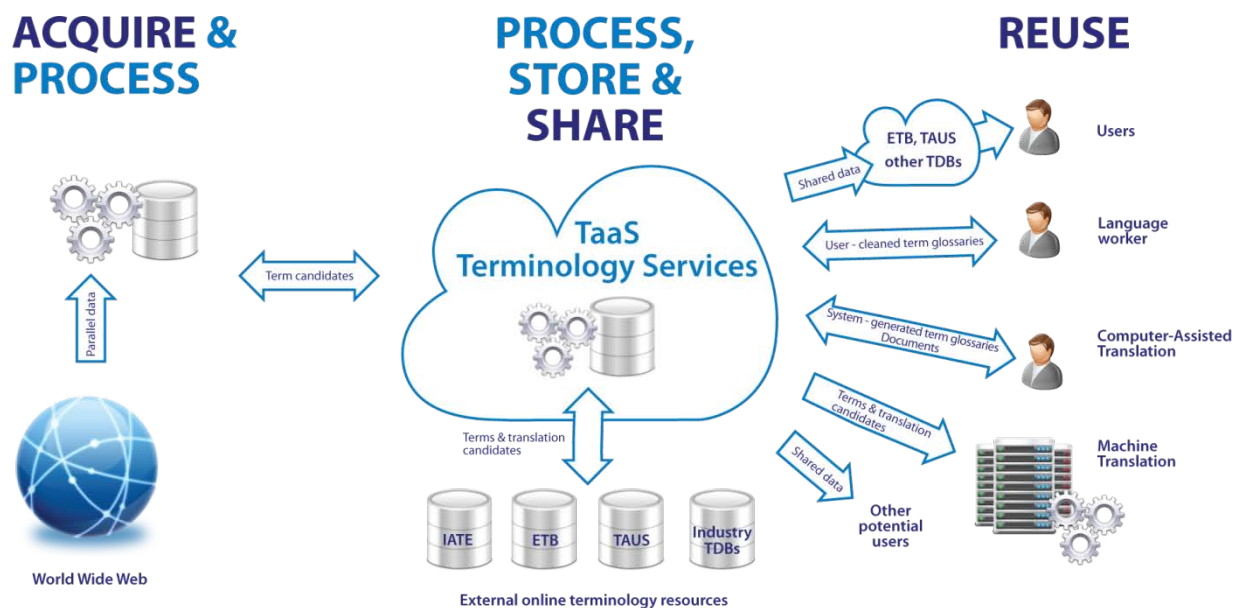


Figure 1: The concept of the innovative cloud-based platform for terminology services

2. TaaS in an Outline

The TaaS platform is being developed in an industry-research collaboration project within the EU Seventh Framework Programme for Research and Technological Development.

The motivation of the TaaS platform is to facilitate terminology work in practical translation scenarios by providing a number of online terminology services.

User surveys have shown that translators, editors, technical writers, and other language workers spend up to 30% of their working time on terminology research, looking for terms in multiple local and online sources, acquiring terminology, and organising proprietary terminology glossaries (Blancafort et al., 2010). In some cases, terminology research can consume more than 30% of overall working time, e.g. in the translation of technical specifications (Massion, 2007). A language worker usually needs immediate answers to terminology requests but due to time and cost constraints proper terminology search is often skipped. Resulting errors in term usage affect not only translation/localisation productivity and overall costs but also influence further stages of documentation life cycle, e.g., failures in product technical support, client request processing etc.

TaaS addresses these needs by establishing a cloud-based platform to provide online terminology services for key terminology tasks – term identification in user-uploaded documents, translation equivalent recognition in existing terminology resources for identified term candidates, terminological collection creation, acquisition of translation equivalent candidates from the Web (parallel and comparable Web resources), crowd-sourced clean-up of terminological data, sharing these data and reusing them in crucial usage scenarios, and thus becoming a part of the multifaceted global cloud-based service infrastructure.

The platform offers automated workflows and facilities for the following activities:

- Automatic identification of monolingual term candidates in user-uploaded documents using state-of-the-art linguistically and statistically motivated term extraction techniques.
- Automatic recognition of term translation equivalent candidates for terminological units identified in user-uploaded documents using the largest publicly available terminology databases, such as IATE¹ and EuroTermBank², as well as statistical terminological lists acquired from domain-specific comparable corpora and publicly available parallel corpora found on the Web.
- Collaborative terminology clean-up (creating, editing, deleting term entries) and monolingual and multilingual terminological collection creation.
- Sharing and reusing user-created and publicly available terminology (including monolingual and bilingual terminological collections) with the help of import/export application programming interfaces (APIs) for automated processes and easy-to-use graphical user interfaces for human users.

The TaaS platform provides also Web service-based interfaces for CAT tools and MT systems that allow terminology look-up in external terminology databases and user-created private and public terminological collections, specialised terminological collection retrieval and term translation candidate mark-up within translatable documents for CAT, MT, and other automated tasks. The conceptual design of the TaaS platform is depicted in Figure 1.

3. Workflow for the Creation of a Bilingual Terminological Collection

Translators, technical writers, terminologists, and other language workers, when working on domain-specific writing tasks (i.e., translation, documentation etc.), require in-domain terminological dictionaries (monolingual and multilingual) that

¹ <http://iate.europa.eu/>

² <http://www.eurotermbank.com>

can aid them in their effort to produce content that simultaneously has correctly and consistently applied terminology. The TaaS platform provides a workflow for *Bilingual Terminology Collection Creation* that allows human users to create terminological collections (i.e., raw terminological dictionaries) semi-automatically from user uploaded documents and comparable and parallel corpora found on the Web.

3.1 Monolingual Term Extraction

The TaaS platform allows semi-automatic terminological collection creation from multiple key formats that have been identified in a prior target user survey described in Gornostay et al. (2013) as the most used by the community including the Open Document³ formats, PDF and several parallel data exchange formats, e.g., TMX and XLIFF.

In the first step, plaintext is extracted from user-uploaded documents and terms are tagged in the documents with statistically and linguistically motivated term extraction methods following Pinnis et al. (2012) in three steps. At first, term candidates are acquired using part-of-speech pattern filtering. Then, terms are weighed using different statistical association measures; the weights are normalised with the help of the TF*IDF (Spärck Jones, 1972) measure using reference corpora statistics (i.e., an inverse document frequency list calculated on a broad domain corpus). The platform supports term tagging for all 23 official languages of the European Union and also for Russian and Croatian.

After term tagging, a monolingual terminological collection in the TBX⁴ format is created. At first, all unique terms are extracted from the tagged documents and normalised (i.e., transformed from the term morpho-syntactic surface forms to the morpho-syntactic base forms). As term normalisation is a language dependent task, it is currently available for selected languages (including English, Hungarian, Latvian, Lithuanian, and other project languages). If normalisation is applied, monolingual terms are consolidated (i.e., different surface forms of the same term are grouped together as one term entry) using term normalised forms and the respective morpho-syntactic descriptions of the normalised forms. If normalisation is not applied, monolingual terms are consolidated using term lemma sequences and part-of-speech sequences.

When automated processes are completed, the user can perform terminology clean-up or execute term translation lookup in order to proceed to multilingual terminological collection creation.

³ Open Document Format for Office Applications

⁴ TBX is a terminology exchange format originally created by the Localization Industry Standards Association (LISA) and later standardised by ISO as international standard ISO 30042:2008.

3.2 Retrieval of Term Translation Equivalents

After extracting terms from user-provided documents, the TaaS platform creates a bilingual terminological collection by finding potential translation equivalents for each of the extracted terms. For this, TaaS queries several sources of terminological data looking for entries that match the search term, are in the same subject field, and have target language equivalents.

Four types of terminological data are queried:

- private (confidential) TaaS terminological collections of the particular user,
- terminological collections shared by other TaaS users,
- external terminology databases,
- and the TaaS Statistical Database, which contains translation equivalent candidates acquired from comparable and parallel corpora found on the Web.

Let us briefly describe the sources mentioned above. The TaaS platform provides facilities to store all terminological collections created by the user. The user can create a collection either by applying TaaS workflows on the user-provided documents or by importing his/her locally created dictionary into the TaaS platform. Common formats, such as TBX and CSV, are supported for importing user terminology.

By default, user terminology is private, i.e., accessible only to the user. The owner of the terminology can provide individual access rights to his/her terminology to other users within the working group.

We encourage users to share their terminology with other users by changing their status to *Shared* (public). By sharing their terms, users participate in a collaborative effort to increase the size and scope of publicly available terminology resources. Shared terminological collections are accessed and used by both TaaS users and by TaaS workflows.

TaaS also searches several well-established online terminology databases:

- EuroTermBank: an online multilingual terminology portal providing consolidated access to 2.6 million terms from 137 terminology resources in more than 30 languages (Vasiljevs et al., 2008),
- IATE: an EU inter-institutional terminology database containing 1.4 million multilingual entries used in different EU legislative acts and other documents,
- TAUS Data Repository: a large collection of shared translation memories (TM) provided by members of TAUS (Translation Automation User Society).

It should be noted that TAUS translation memories consisting of sentences and text

segments with their translations cannot be considered as a terminological resource. But in some fields, such as information technology, TM include many terms and their translation originates from software interface and product documentation, and TM strings with exact match are retrieved by TaaS.

Querying terminology resources and TM is a relatively straightforward process. But as new terms in different areas are appearing very frequently and they have to be translated in many languages, even the best terminology databases include only a fraction of terms that are appearing in the daily workload of translators.

To assist in translating terms that do not have translation equivalents in terminology databases, TaaS provides means of finding possible translations in Web data. For this purpose TaaS collects parallel and comparable data from the Web, aligns it at sentence and word levels and extracts potential term candidates with their translation candidates. Comparable corpora consist of original source-target language document pairs on the same topic, thus not translations of each other.

For data collection and extraction, TaaS uses an updated version of the ACCURAT Toolkit (Pinnis et al., 2012). The ACCURAT Toolkit provides tools and workflows for acquisition and processing of comparable corpora in order to acquire multilingual parallel data (including terminological data). Parallel and comparable Web data are collected from multilingual news feeds, focused Web domains, and Wikipedia. This workflow runs periodically in the background and stores resulting terms in the TaaS Statistical Database.

3.3 Collaborative Terminology Clean-up

The progress in information technologies and their role in the modern specialised lexicography cannot be overestimated. However, a specialist is the one who decides whether a linguistic unit is a term or not. This is about the unithood and termhood of a term and is out of scope of this paper, although it is one of the important steps in a term life cycle. Professionals seek joint collaboration and exchange of terminological data, and the TaaS platform offers these functionalities. Several of the data clean-up functions that are provided by the TaaS platform are: deletion of term candidates from the terminological collections, editing of various data categories of term entries within the terminological collections, changing status of the term candidates etc.

4. Sharing and Reusing Terminology

The concept of sharing, unfortunately, is not present to a considerable extent in the current models of major terminology resources – instead of providing the opportunity for consumers to contribute, reuse, and share their data, major terminology resources (term banks and databases) typically keep to the traditional one-way communication of their high quality pre-selected content.

According to a recent survey, there is a need for collaborative solutions and sharing models – 60.5% of respondents (out of 1782 participants) are willing to share their terminology with colleagues (Gornostay et al., 2013).

The core objective of the TaaS platform is, however, to align the speed of terminology resource management with the speed at which multilingual documentation is created. In order to achieve this goal, the TaaS platform allows its users to take full control of their monolingual and multilingual terminological collections and lets them decide with whom to share their terminology, to whom to grant the rights of collaborative improvement of terminological content, and to whom to grant access rights to the user terminology.

The TaaS platform provides human users with simple terminological collection importing and exporting methods in TBX, CSV (comma-separated values), and TSV (tab-separated values) formats. When terminological collections are imported within the TaaS platform, they are immediately accessible to other third party systems that support the TaaS platform's API (provided that the user has access to the third party systems).

The user can also make his/her terminological collections completely public, thus sharing them with every user of the TaaS platform.

5. Interfaces for CAT Tools and MT Systems

Multilingual consolidated and harmonised terminology in the form of monolingual and multilingual terminological collections is already utilised as an important resource in the process of human translation. However, a dictionary user is not necessarily a human specialist but could be an automated system: a machine user. Therefore, the TaaS platform also offers access to multilingual terminological collections through a dedicated Web service API. Typical machine users that may benefit from the service are, for instance, CAT tools, MT systems, authoring and (multilingual) documentation and content management systems, terminology management systems, indexing systems, search engines, Web crawlers, information retrieval systems (including cross-lingual information retrieval), and others. Many of the abovementioned systems already have integrated workflows for terminology management; therefore, the linking to the TaaS platform will offer a wider access to existing and shared terminology. The latter systems (starting from search engines) may also exploit term lists as seeds for acquiring Web data or to focus their search for data in domain-specific (or search query specific) directions. The Web-based API offers three main functions: lookup of terms in existing terminological collections, import of multilingual terminological collections, and export of collections from the TaaS platform. Terminological collections can be imported and exported using the TBX format. Additionally, for machine users that provide human users with the functionalities to search, create, delete and clean up terminology (like a terminology

management system), the TaaS platform offers advanced interfaces that operate similarly to the services offered for human users accessing the TaaS platform directly. In the TaaS project we study and evaluate the benefits of having access to multilingual terminology collections for two specific machine users. At the time of writing this demonstration paper, the TaaS platform's API interface was supported by the memoQ⁵ CAT tool and the LetsMT⁶ SMT platform (Vasiljevs et al., 2012).

6. Conclusions

In this demonstration paper we have presented an innovative platform "Terminology as a Service" (TaaS) for acquiring raw terminological data, cleaning up, sharing, and reusing terminological data, based on cloud computing. The platform serves, among other things, the needs of specialised lexicography.

During the conference we will demonstrate the fully functional prototype of the platform. The live demonstration workflow will include extraction of terms from user-provided documents, as well as finding corresponding translation equivalents in terminology databases and in statistically aligned corpus data.

7. Acknowledgments

The work within the TaaS project has received funding from the European Union under grant agreement n° 296312.

We would like to thank the development team, our colleagues Andis Lagzdiņš and Pēteris Ņikiforovs, for their work on the TaaS platform.

8. References

- Busa, R. (1961). Les travaux du Centre per l'automazione dell'analisi letteraria. *In Cahiers de Lexicologie*. Vol. 26. No 1.
- Gornostay, T., Vopodiyanova, O., Vasiljevs, A., & Schmitz, K.-D. (2013). Cloud-Based Terminology Services for Acquiring, Sharing and Reusing Multilingual Terminology for Human and Machine Users. *Proceedings of the conference TRALOGY II: Futures in Technologies for Translation. The quest for meaning: where are our weak points and what do we need?* Paris.
- Blancafart, H., Daille, B., Gornostay, T., Heid, U., Méchoulam, C., & Sharoff, S. (2010). TTC: Terminology extraction, translation tools and comparable corpora. *In Proceedings, 14th EURALEX International Congress*, pp. 263-268.
- Massion F. Управление терминологией: роскошь или необходимость?

⁵ memoQ is available at: <http://kilgray.com/products/memoq>.

⁶ LetsMT is accessible at: <https://www.letsmt.eu>.

Профессиональный перевод. Выпуск 12, 2007.

- Pinnis, M., Ljubešić, N., Ștefănescu, D., Skadiņa, I., Tadić, M., & Gornostay, T. (2012). Term Extraction, Tagging, and Mapping Tools for Under-Resourced Languages. *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012)*, pp. 193-208. Madrid.
- Pinnis, M. (Tilde), Ion, R., Ștefănescu, D., Su, F., Skadiņa, I. (Tilde), Vasiljevs, A. (Tilde), & Babych, B. (2012). ACCURAT Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 91-96. Jeju: Association for Computational Linguistics.
- Shaikevich, A. (1983). *Проблемы терминологической лексикографии = Problems of the Terminological Lexicography*. Moscow.
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Volume 28, pp. 11-21.
- Vasiljevs, A., Rirdance, S., & Liedskalnins, A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pp. 213-220. Hong Kong
- Vasiljevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: a cloud-based platform for do-it-yourself machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 43-48. Jeju: Association for Computational Linguistics.

Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons

Núria Gala⁽¹⁾, Thomas François⁽²⁾, Cédric Fairon⁽²⁾

(1) LIF-CNRS, Aix Marseille Université, 163 av. de Luminy case 901,
13288 Marseille Cedex 9, France

(2) CENTAL, Université Catholique de Louvain, Place Blaise Pascal 1,
1348 Louvain-la-Neuve, Belgique

E-mail: nuria.gala@lif.univ-mrs.fr, {thomas.francois}{cedrick.fairon}@uclouvain.be

Abstract

The readability of a text depends on a number of linguistic factors, among which its lexical complexity. In this paper, we specifically explore this issue: our aim is to characterize the criteria that make a word easy to understand independently of the context in which it appears. Yet such a concern must be addressed in the context of particular groups of individuals. In our case, we have focused on language production from patients with language disorders. The results obtained from corpus analysis enable us to define a number of variables which are compared to information from existing resources. Such measures are used in a classification model to predict the degree of difficulty of words and to build a lexical resource, called *ReSyf*, in which the words and their synonyms are classified according to three levels of complexity.

Keywords: lexical resource, readability, simplification, natural language processing, language model.

1. Introduction

There has been a significant number of works on the readability and simplification of texts over the last 80 years. Most of them take into account the lexicon in an assessment of text difficulty. For instance, Flesch (1948) used the number of syllables per word as a measure of word complexity. Smith (1961) instead suggested using the mean number of letters, since this is easier for a computer to calculate. Stenner and Burdick (1997) predicted text difficulty from the logarithm of word frequencies.

However, although all these studies were concerned with the impact of the lexicon on text difficulty, they did not directly assess the complexity of the lexicon. Efforts at this level were more concerned with designing lists of ‘easy’ words. Such lists have been produced for teaching purposes in different languages, relative to a first language (L1) or a second language (L2). Among them, some of the most well-known are, for English, the *Teachers’ Book of Words* (Thorndike, 1921) and the *Basic English* (Ogden, 1930) and, for French, *Le Français Fondamental* (Gougenheim, 1958) and the *Listes Orthographiques de Base du Français* (Catach, 1985).

Although these lists were subsequently used for text readability purposes (Dale and Chall, 1948), their use presents several limitations in terms of assessing the difficulty of a whole lexicon. First, the lists are based on a single criterion, such as the

frequency of words (Thorndike, 1921), or the percentage of words known by 80% of schoolchildren from the fourth grade (Dale, 1931). More importantly, their coverage is generally limited to a set of a few hundred ‘easy’ words, making them too restricted to be used, for instance, in text simplification systems. The problem of coverage is accentuated as the vocabulary of a language is in constant evolution.

Therefore, it appears that a more integrated approach, using Natural Language Processing (NLP) techniques, could be suitable for automatically predicting the difficulty of words.

To our knowledge, the only readability study that proposes a formula directly at the lexicon level is that of Bormuth (1966). He first used the cloze test procedure¹ to yield a corpus of 20 educational texts annotated in terms of difficulty at the word level. Then, he modelled word difficulty with four variables: the number of syllables, the number of letters, a frequency index, and the word depth as defined by Yngve (1962). When combined, these four variables produced a multiple correlation coefficient (R) of 0.505, a far lower score than that obtained by the text level model ($R = 0.934$). From this study, it appears that predicting the difficulty of words is surprisingly harder than predicting text difficulty.

In this paper, we first explore a larger set of variables to predict the degree of difficulty of a word. Then, using these scores, we build a synonym lexicon where each word has a difficulty index. Such a resource is to be used (1) by humans for language comprehension or production and (2) by a language model for automatic simplification. To our knowledge, no existing lexical resource, except for graded scholar word lists, offers its users the possibility to select words according to their degree of difficulty.

The article is organized as follows. In the next section, we discuss which characteristics of a word make it simple or difficult, according to psycholinguistic studies and linguistic variables that we have defined. In section 3, we describe the resources we use to compare our features on two sets of words: (a) words used in a given task by patients affected by Parkinson’s disease, (b) words from a large general lexical list. In Section 4, we report experiments and methodology to design a first gold-standard graded list and a model of lexicon difficulty. Finally, we conclude with some remarks on the limitations of our present approach and proposals for future work.

2. How simple can a word be?

Identifying how simple a word can be has been of interest to psycholinguists for many

¹ This test, designed by Taylor (1953) to measure reading comprehension, requires readers to read a text with regular blanks (one every five words) and fill in as many blanks as possible.

years. Experiences of the complexity of words with regards to various recognition tasks (lexical decision, semantic categorization, etc.) have been intensely reported in the literature (Ferrand, 2007). One of the main findings is the word frequency effect: a high-frequency word is recognized more easily than one of low frequency. The close correlation between frequency and difficulty has been highlighted in many studies (Howes and Salomon, 1951; Brysbaert et al., 2000).

Other word-level effects have been stressed in psycholinguistic literature, such as the familiarity effect (Gernsbacher, 1984), the age of acquisition effect (Morrison and Ellis, 1995), the orthographic neighbour effect (Andrews, 1997), the length of words (O'Regan and Jacobs, 1992), etc. Most of these effects are indeed correlated with the difficulty of texts (François and Fairon, 2012) and are likely to be also a valuable source of information for a model of word complexity.

A second source of information about word simplicity comes from linguistic studies on levels lower than the word unit: morphemes or phonemes. Intra-lexical factors, such as familiarity of phonemes, regularity in pronunciation, fixed stress, consistency of the sound-script relationship, inflexional and derivational regularity, morphological transparency, generality, register neutrality, or number of meanings per form, affect vocabulary learning (Laufer, 1997). For Schreuder and Baayen (1997), the number of morphemes correlated with the size of the derivational family has an impact on visual word recognition.

To various extents, all these factors combine to explain word difficulty. It is acknowledged that the combination is dependent on a given group (or 'class') of individuals (François, 2012). What may be simple for one group may not be for another, especially since there is a wide variety of readers who do not have the same needs. However, we believe that, in order to describe how simple words can be, there are some general characteristics that can be related to fine-grained linguistic criteria. NLP methods are useful in formalizing such features and checking them on large amounts of data.

For the purposes of this study, we have identified a set of variables from the two following families:

- *Intra-lexical variables*: (1) number of letters, (2) number of phonemes, (3) number of syllables, (4) syllable structure, (5) consistency of sound-script relationship, (6) spelling patterns, (7) number of morphemes, (8) composition, and (9) affix frequency (for derived word).
- *Psycholinguistic variables*: (10) phonological neighbourhood, (11) orthographic neighbourhood, (12) abstract-concrete or imageability, (13) lexical frequency, (14) size of the derivational family, (15) absence/presence from Gougenheim list (Gougenheim et al., 1964), etc.

To check how these variables relate to difficulty, we performed two experiments. First, we computed their values on a *simplified* language corpus (see Section 4.2 for implementation details of the variables) and compared these results with values obtained from a general language lexical database. In seeking a corpus attesting some simplified language, we considered that linguistic productions from people with speech-related disorders might be a good start for observing ‘simple’ vocabulary. We therefore collected a corpus containing language productions of sufferers of Parkinson’s disease (other types of speech-related disorders might be considered in the future). Second, we analysed how those variables vary within a lexicon of graded words for French, intended for schoolchildren (see Section 4.2.3).

3. Resources

In this section, we present the four resources used in our experiments. First, we describe a corpus with simple language productions. Second, we introduce a lexical database for French, Lexique3 (New et al., 2005), that is a representation of the general vocabulary. These two resources enable us to test some variables that potentially account for simple words. We also describe Manulex (Lété et al., 2004), a list of word frequencies at various school grade levels. Lastly, we present JeuxDeMots (Lafourcade, 2007), a lexical network that helped us to build *ReSyf*, our list of graded synonyms.

3.1 Parkinson corpora

The general public mainly recognizes Parkinson's disease through its motor symptoms (rest tremor, akinesia, and rigidity). However, the pathology may also entail language and speech impairments², namely dysarthria (Pinto et al., 2010), which includes hypophonia (reduced voice volume), monotone speech, and difficulties with articulation of certain sounds and syllables, as well as increased frequency and duration of hesitations and pauses (McNamara, 2010). Sentence structures are simplified (shorter), with an increase in the ratio of open-class items (nouns, verbs, adjectives, and adverbs) to close-class items (determiners, prepositions, conjunctions, etc.).

For our study, we used a corpus of twenty recordings from twenty Parkinson’s patients describing the same picture (a short scene of an everyday situation)³. Patients were recorded whilst in ‘off state’, that is, with no medication that could have alleviated the effects of the disease.

After transcribing the twenty recordings, we obtained a corpus of 2,271 tokens that

² <http://www.sciencedaily.com/releases/2011/02/11020262.htm>

³ The authors are grateful to S. Pinto from the Laboratoire Parole et Langage (LPL-CNRS, Aix Marseille Université) for providing the corpora and valuable insights on the disease.

we tagged using TreeTagger (Schmid, 1994). All marks of disfluencies, except repeated words, were removed (hesitations, truncated words, etc.). The average number of words per file was 113, the shortest file contained 42 words and the longest 233.

3.2 A lexical database for general French words

Lexique3⁴ (New et al., 2005) is a free lexical database containing 142,728 words (47,342 correspond to a lemma; the other entries are inflected forms). Each word is described with phonological and morphological information (phonetic transcription, part of speech, morphological features [gender, number, tense, etc.], number of phonemes, number of syllables, syllable structure, number of morphemes, etc.). The database also provides estimates on the frequencies of occurrence of the words in books and film subtitles.

Figure 1 displays an example of the information available for the entry *armures* ('armours'):

ortho	phon	lemma	pos	gender
armures	aRmyR	armure	NOM	fem
number	V morpho	freq bks	freq films	nb phon
plu	-	5.46	8.11	5
struct lett	struct pho	syllables	nb lett	nb syll
VCCVCVC	VCCVC	aR-myR	7	2
sy struct pho	sy struct lett	nb homoph	nb homogr	nb morph
VC-CVC	ar-mu-re	1	0	1

Figure 1: The entry *armures* ('armours') from Lexique3.

Only some of the most significant fields are presented here, in the following order: spelling form, phonemic form, lemma, part-of-speech, gender, number, verbal morphology (tense, etc.), frequency estimated from a book corpora, frequency computed from film subtitles, number of phonemes, letter structure, phonemic structure, syllables, number of letters, number of syllables, syllable structure (phonemes) and syllable structure (letters), number of homophones, number of homographs, and number of morphemes.

3.3 A lexicon with scholar levels

To obtain a list of graded words, we used Manulex⁵ (Lété et al., 2004), a list of French words whose frequencies have been extracted from primary school textbooks. For a

⁴ <http://www.lexique.org>

⁵ <http://www.manulex.org>

given word, the authors computed several measures (raw frequency, frequency of use over one million words, dispersion index and standard frequency index) for the three following levels of education:

- First year of primary school (children of 6 years old).
- Second year of primary school (7 years old).
- The three following years of primary school (8 to 10 years old).

Figure 2 provides an example of four entries, *pomme* ('apple'), *vieillard* ('old man'), *patriarche* ('patriarch') and *cambricoleur* ('burglar'). Only the raw frequency of each word per level of education is shown:

lemma	pos	Fq Lvl	Fq Lvl	Fq Lvl
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambricoleur	N	2	-	33

Figure 2: Sample of four entries in Manulex.

For the purpose of building a list of graded words, we transformed the frequency distributions over the three levels into a class system where a word can be assigned to only one class. As a first approach, we defined three classes corresponding to the three levels of education listed above and it was assumed that a given word would belong to the textbook level where it was first observed (e.g. level 1 for *pomme* and *cambricoleur*, but level 3 for *patriarche*). This straightforward classification has obvious shortcomings. For instance, it assigns the same level (level 1) to the words *pomme* and *cambricoleur* from Figure 2, whereas they present very different frequency distributions.

From this example, it seems that using a more complex function to transform the frequency distributions might produce a better classification. The idea is to give a different value to words, such as *pomme* – those that are more frequent at level 1 than at the other levels – and words such as *cambricoleur* that rather belong to levels 2 and 3. We thus experimented with the following formula:

$$N_c = N + e^{-r}, \quad \text{where } r = \frac{\sum_{k=1}^i U_k}{\sum_{i+1}^N U_k}$$

N_c is a continuous score that is used at the word difficulty level instead of N , the level predicted by our first simple method describe above. N_c is obtained by summing N and a quantity e^{-r} that is inferior to 1 and is exponentially related to the ratio of the frequencies U_k at level k .

However, using this new scale did not lead to significant improvement for the

experiments described in Section 4.2, so we decided to use the simple approach throughout the paper. After applying the simple function and deleting grammatical words, we thus obtained a list containing 19,037 lemmas from Manulex, distributed as follows: 5863 words (31%) corresponding to level 1, 4023 words (21%) for level 2 and 9151 (48%) words for level 3.

At this stage, we compared the lemma list from the Parkinson corpora to the graded list obtained from Manulex, and the results were the following: 94.30% of the words in our corpora are tagged as belonging to the level 1 of Manulex, 1.45% are tagged as level 2, while only 1.63% belong in level 3 (the remaining 2.62% correspond to tagging errors, i.e. words tagged differently in the corpus and in Manulex). This confirms that the Parkinson list contains simple language productions.

3.4 A semantic network

JeuxDeMots⁶ (JdM) is a freely available lexical network that is under development in the framework of a game for leveraging crowd-sourcing (Lafourcade, 2007). Given a trigger word, the game consists of proposing related words corresponding to a specific semantic or thematic relation. The resulting resource contains 163,543 words (in May 2013) with at least one lexical relationship (associated term, synonym, antonym, agent, patient, etc.).

Figures 3 and 4 display the information collected for the word *cambricoleur* ('burglar'). There are 114 thematic associations (*cheater, break in, thief, robbery, steal*, etc.) in which this word has been the trigger (Figure 3).

There are 71 relations (Figure 4) in which this word has been the target when asking for, line 4, 'agent of the verb *steal*', line 5 'who could hurt with a *weapon*', line 6 'synonym of *thief*', etc.

```
114 relations ==>
• cambrioleur ---r_associated#0:420--> escroc
• cambrioleur ---r_associated#0:410--> cambrioler
• cambrioleur ---r_associated#0:390--> malfaiteur
• cambrioleur ---r_associated#0:380--> cambriolage
• cambrioleur ---r_associated#0:380--> dérober
• cambrioleur ---r_associated#0:370--> voleur
• cambrioleur ---r_associated#0:280--> monte-en-l'air
• cambrioleur ---r_associated#0:260--> voler
```

Figure 3: Outgoing relations in Jeux de Mots

⁶ <http://www.jeuxdemots.org/>

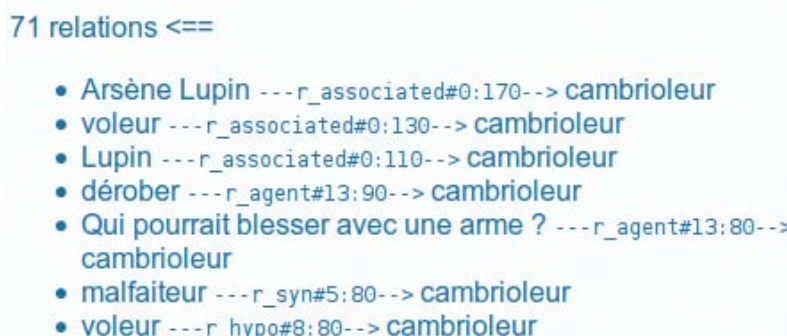


Figure 4: Ingoing relations in JeuxDeMots.

4. Building a graded synonym lexicon

Automatic acquisition of linguistic knowledge from corpora (raw texts or lexical resources) is a widespread trend in NLP. Over the last decades, many unsupervised and semi-supervised approaches have become a real alternative to manual development – too costly and time consuming (Gala and Lafourcade, 2011). More recently, collaborative approaches have emerged, based on the principle of sharing contributions (Calzolari, 2013), especially through games with a purpose (*gwap*), the lexical network JdM being an outstanding example of this trend.

For the purpose of creating a graded synonym lexicon, we first acquired information from Manulex in order to obtain a gold-standard list of graded words. Second, we implemented some of the identified intra-lexical and statistical features in order to automatically grade words outside our gold-standard list.

4.1 Acquiring information from existing resources to establish a gold-standard list of graded synonyms

We have indications that Manulex offers accurate difficulty classification: 94.3% of the words from the Parkinson’s patients corpora correspond to level 1, which is consistent with what we know about language productions of Parkinson’s patients. Therefore, we consider that Manulex grading can be used as a gold standard to create a first list of words with graded synonyms.

To this aim, we checked whether the 19,037 words of Manulex could be linked to synonyms in the lexical network JdM. From the initial 19,037 words in Manulex, 17,870 (93.87%) were present in JdM (the remaining words can be present in JdM, but with no known synonym relation). The distribution by level is as follows:

Level	Proportion	Counts
1	30.1%	5,375
2	21.0%	3,755
3	48.9%	8,740
1–3		17,870

Figure 5: Distribution of Manulex words in JdM.

From this list of 17,870 words, we gathered their synonyms in JdM: 10,975 have at least one synonym with a level in Manulex (we temporarily removed words absent from Manulex; they will be graded with our difficulty model). We obtained 12,687 graded synonyms, distributed as follows:

Level	Proportion	Counts
1	35.3%	4,477
2	21.7%	2,749
3	43.0%	5,461
1-3		12,687

Figure 6: Distribution of synonyms by level.

Figure 7 shows a sample of such a graded list containing synonyms from JdM along with their levels from Manulex:

Armure (1): protection(1), cuirasse(2), harnais(3)
Piétiner (2): marcher(1), fouler(3), piaffer(3), trépigner(3)
Patriarche (3): chef(1), père(1), vieillard(2)
Cambrioleur (1): malfaiteur(3), voleur(1), aigrefin(3)

Figure 7: Sample of ReSyf entries (*armour, protection, breastplate, harness; trample, walk, stamp one's feet, paw the ground; patriarch, chief, father, old man; burglar, criminal, thief, crook*).

We consider this list of graded words our gold-standard. Words absent from this list will be graded using our system for the automatic assessment of lexicon difficulty.

4.2 Towards a difficulty model for lexicon

This section presents the difficulty model we used to assess the difficulty of synonyms absent from Manulex. First, we detail which predictors of the lexicon complexity were implemented and how (at the time of writing this paper, only some had been tested). Then, we report two experiments performed on our three resource corpora that aimed to better understand which features are the most useful in predicting lexicon difficulty. Finally, we describe the model designed to assign a word to one of the Manulex levels.

4.2.1 Predictors of lexicon difficulty

As mentioned in Section 2, a large number of lexical predictors have been described in the literature. We implemented several of them, as follows:

a) Intra-lexical variables

(1) *number of letters*: we counted the number of alphabetical characters.

(3) *number of syllables*: we adopted a hybrid syllabification method. For words included in Lexique3, we used the gold syllabification included in the dictionary. For all other words, we generated API phonetic representations with *espeak*⁷, and then applied the syllabification tool provided with Lexique3 (Pallier, 1999). The accuracy of this combined process exceeded 98% on a small test list.

(2) *number of phonemes* and (4) *syllable structure*: obtained from the syllabification system. For the syllable structure, we defined three categories of increasing difficulty, using their frequencies in the Parkinson corpus as a criterion: the most frequent structures (CYV, V, CVC, CV)⁸, a group of less frequent structures (CCVC, VCC, VC, YV, CVY, CYVC, CVCC, CCV) and a final group containing only rare combinations.

(5) *consistency of sound-script relationship*: computed by comparing the number of letters and phonemes. We parameterized the output as three possible outcomes: 0 for complete transparency; 1 for a difference not higher than 2 characters, and 2 for words particularly obscure (difference higher than 2 characters).

(6) *spelling patterns*: defined as five categories of difficult patterns:

- double vowels ('oo', 'éé'),
- double consonants ('bb', 'cc', 'ff', 'gg', 'll', 'mm', 'nn', 'pp', 'rr', 'ss', 'tt'),
- other digraphs in French ('ck' and 'qu' [k], 'ch' and 'sh' [ʃ], 'ph' [f], 'gn' [ɲ]),
- nasal vowels written with digraphs ('an' [ɑ̃], 'in' [ɛ̃], 'on' [ɔ̃], 'un' [œ̃])
- oral vowels written with digraphs ('ai' [e], 'au' [o], 'eu' [œ], 'ou' [u]).

There is work in progress concerning the remaining variables:

(7) *number of morphemes* and (8) *composition*: the hypothesis being that constructed words are more difficult to grasp.

(9) *affix frequency* on derived words: the difficulty of a derived word may depend on the frequency of the affix. In French, some affixes are very productive (-age with verbal basis as in *lavage* ['wash'], *balayage* ['weep'], *tournage* ['filming'], etc.). Other affixes are quite rare (-is as in *treillis* ['canvas'] or *tournis* ['dizziness']). The effect of affix frequency might have an impact on the level of difficulty of a word.

b) Psycholinguistic variables

(11) *orthographic neighbours*: computed from a list of neighbours distributed under

⁷ <http://espeak.sourceforge.net>

⁸ C stands for consonant, V stands for vowel and Y stands for semi-vowels [j], [ɥ] and [w].

the Lexique3 project, which includes 128,919 inflected forms. Based on findings in the cognitive psychology literature, we modelled this effect from different angles: the number of neighbours (11a), the cumulative frequency of all the neighbours (11b), and the number of more frequent neighbours (11c).

(13) *lexical frequency*: we used the lemma frequencies from Lexique3, which contains about 50,000 lemmas. Their frequencies were obtained from movie subtitles and were smoothed with the simple Good-Turing algorithm (Gale and Sampson, 1995) to assign a default frequency to out-of-vocabulary words. Preliminary experiments showed that it was better to use the logarithm of the frequencies, as commonly reported in the literature.

(15) *presence* in a list of simple words: a convenient proxy of the ‘simplicity’ of words. We then used a binary feature telling us whether this word is in the Gougenheim list (Gougenheim et al., 1964) or not. Since it was not obvious which size of list would be the best, we experimented with several sizes, ranging from 1,063 to 8,875 words.

We are currently testing the remaining variables:

(10) *phonological neighbourhood*: the number of words having a maximum number of phonemes in common (minimal series such as ‘bain’ [bɛ̃], ‘main’ [mɛ̃], ‘pain’ [pɛ̃], etc.). Our hypothesis is that the higher the number of neighbours, the easier the word.

(12) *abstract-concrete* and *imageability*: concrete words, as well as vocabulary from familiar contexts, would have a lower level of difficulty than abstract words.

(14) *size of the derivational family*: as shown by Schreuder and Baayen (1997) for visual word recognition, the bigger the family, the lower the difficulty a word would have as a result of proximity.

4.2.2 Analysis of the variable efficiency

In this section, we analyze how a simple lexicon (obtained from the Parkinson corpus) deviates, according to our variables, from general trends in the language, as represented by Lexique3.

For each variable listed in the previous section⁹, we compared its distribution on both corpora using statistical tests. More precisely, a T-test (t) was applied to parametric interval variables, a Mann-Whitney test (U) to non-parametric interval variables, and a Chi-square test (X^2) to nominal variables (see Howell, 2008 for details). Figure 8 reports the means on both corpora (when meaningful) along with the p-values of the statistical tests.

⁹ Presence in the Gougenheim list (15) was not considered for this step of the analysis, since this feature is not an intrinsic characteristic of words.

	Park.	Lex3	p-value¹⁰
1. # letters	6.3	8.6	< 0.001 (t)
2. # phonemes	4.7	6.8	< 0.001 (t)
3. # syllables	1.96	2.89	< 0.001 (t)
4. syll. struct.	/	/	0.6 (X ²)
5. sound-script	1.05	1.14	0.0004
6. # ortho.	0.75	0.96	0.007 (X ²)
11. #	3.88	1.31	< 0.001 (U)
13. frequencies	756.7	19.5	< 0.001 (t)

Figure 8: Variation in means from both corpora and significance of the difference between means.

The mean number of letters, phonemes and syllables is lower in our simple lexicon than in the language as represented by Lexique3. Words used by Parkinson speakers have, on average, 6.3 letters, 4.7 phonemes and 1.96 syllables; whereas words in Lexique3 have, on average, 8.6 letters, 6.8 phonemes, and 2.89 syllables. All three differences are significant, which is not surprising since these variables have been known for long in the readability literature as good proxies for the lexical complexity of a text.

Word frequency (13) is another feature that has proven useful for text readability measures. We also notice a significant difference ($p < 0.001$) between the frequencies of simple words, which are more frequent on average than the terms from Lexique3.

More innovative approaches of the lexicon difficulty include our variables based on the sound-script correspondences (5) and the difficulty of specific spelling patterns (6). Interestingly, both variables show significant differences between both lexicons. It appears that a simple lexicon contains significantly less complex correspondences between the sound and the written form. Also, simple words comprise fewer complex spelling patterns: 0.75 on average for simple words and 0.96 for the general lexicon.

Finally, simple words have significantly more orthographic neighbours (11) ($p < 0.001$). According to psycholinguistic literature (Andrews, 1997), this characteristic yields a facilitation effect in English, but not in French. Our result appears inconsistent with these experimental findings, but this is likely due to the fact that we did not control for the frequency of words. Since simpler words are also more frequent and shorter, they also tend to have more neighbours. It is worth noting that this type of inter-correlation between our variables is a well-known issue that must be taken care of when variables are combined within a statistical model, such as in Section 4.2.3.

¹⁰ The threshold alpha used in this study is 0.05, which means that any lower p-value in this table represents a significant difference between the distributions in the Parkinson corpus and Lexique3.

To conclude this analysis, we have shown that all our variables, except the syllabic structure of words, have a different behavior on a simple lexicon and on the general vocabulary. This can be interpreted as a validation of their effectiveness in predicting the difficulty of terms. The next section further investigates these predictive abilities, using a lexicon of words annotated in terms of their complexity (i.e. Manulex).

4.2.3 The difficulty model

Having confirmed that most of our predictors can be used in order to discriminate between simple and complex words, we used Manulex as a gold standard to describe more precisely the relation existing between one of our variables and word difficulty. This relation, captured through a Spearman correlation¹¹, informs us how a given variable varies in relation to the three levels of difficulty in Manulex. This analysis precedes a more integrated approach, where all efficient variables are combined within a statistical model, which will also be used to assess the difficulty of words.

Name of the variables	Spearman corr. ¹²
1. # of letters	0.27
2. # of phonemes	0.3
3. # of syllables	0.27
11a. # neighbours	-0.25
11b. cumulative freq. of neighbours	-0.25
13. word log-frequencies	-0.51
15. presence in the 5000 first words from the Gougenheim list	-0.41
6. complex spelling patterns (nasal)	0.08
6. complex spelling patterns (sum)	0.05

Figure 9: Spearman correlation for the most meaningful variables.

The total number of variables we tested amounts to 27 (including the variants described in Section 4.2.1). Correlations for the most efficient of them are reported in Figure 9. A positive correlation infers that the difficulty of words increases as the value of the variable increases (e.g. longer words tend to be more complex), whereas a negative correlation corresponds to the opposite relationship (e.g. complex words tend to be less frequent).

¹¹Spearman correlation formula is described among others in Howell (2008). We did not use the Pearson correlation here, since some of our variables do not have a linear relationship with difficulty (e.g. those based on orthographic neighbours).

¹²Due to the large number of words in Manulex, all correlations reported in this table are significant at the level $p < 0.001$.

One should note that among the set of predictors which do not significantly correlate with word difficulty are our three classes of syllabic structures. This finding is consistent with our previous analysis on the Parkinson corpus. More surprisingly, spelling patterns and the difference between the oral and written forms do not account for much of the word difficulty. On the contrary, the two best predictors are the logarithm of word frequencies and the presence/absence from the 5,000 first words of the Gougenheim list.

As a result of this analysis, we selected a subset of nine predictors from our 27, which correspond to the best variables, as listed in Figure 9. These variables were combined using support vector machines (Boser et al., 1992) – generally abbreviated in SVM. It is a generalized linear classifier widely used in automatic classification¹³.

We trained the final classifier on all Manulex words, but first estimated its performance on new words using a five-fold cross-validation approach. This consisted of splitting the data into five folds, training a model on four folds and testing it on the last fold. The accuracies thus obtained are averaged to yield an estimate of the mean accuracy of our model. It is also worth noting that SVMs require setting some parameters: the kernel used, the cost (*C*) and *gamma*. We opted for a radial basis function (RBF) kernel and explored by grid search a limited amount of combinations of values for *C* and *gamma*. The best model (with *C* = 1 and *gamma* = 0.5) attained a 62% classification accuracy.

Such a performance certainly leaves room for improvement, but should also be considered against the difficulty of the task. As we reported previously, Bormuth's (1966) study stressed the complexity of automatically predicting word difficulty. Moreover, our current model's accuracy is nearly twice as good as a random classification.

5. Results and discussion

Applying our lexicon difficulty model to JdM words absent from Manulex, we were finally able to produce a list of 17,870 graded words with graded synonyms, which stands as the first gold-standard list of French words to be used for language comprehension or production. The resource is available at:

<http://cental.uclouvain.be/resyf>

As the synonyms were extracted from a contributive lexical network, they correspond to the target word with a precision rate of 100%. However, some drawbacks can be identified for some lexical units, as a result of using word forms instead of senses.

¹³ For an implementation of SVM available in Python, we relied on scikit-learn (Pedregosa et al., 2011).

5.1 Drawbacks requiring a more fine-grained study of the vocabulary

By and large, we have identified two kinds of issues:

a) Semantics

Polysemy and homonymy are not yet taken into account, neither is the difference between concrete and figurative senses. As a consequence, our resource assigns the same difficulty level to the various senses of a given word. For example, the word *renard* in French means ‘fox’ in a literal sense, but it also refers to an ‘intelligent or smart attitude’. The list of synonyms for this word is the following one:

renard(1) futé(1), malin(1) / goupil(2), canidé(1)

(fox / smart / canid)

The two senses should be distinguished and should probably get a different difficulty score. The same applies to the word *hospitalier* (‘related to hospitals’ in a first sense, ‘friendly and welcoming’ in a second interpretation)¹⁴.

Another problem with the synonyms obtained is the register or language level. Three levels could be defined: familiar or slang, current, formal. A tag indicating the appropriate language register should be added. To give an example, *policier* (‘police officer’) has two synonyms belonging to a familiar register (*flic* and *poulet*, corresponding to ‘cop’). Whether the lexicon is used by someone affected by a language difficulty or by a machine for a lexical simplification task, such information on senses and register should be taken into account.

b) Compounds

In Manulex, compounds mostly belong to levels 2 or 3, for example:

papier-monnaie(3) argent(1), billet(1)

(paper money / money, bill)

homme-orchestre(2) musicien(1)

(band man / musician)

However, in some cases, the semantics of the target word can be obtained by the ‘sum’ of the senses of the word-forms integrating the compound word:

¹⁴ Identifying the semantic structure of lexical units is a crucial issue in NLP. In future work we will follow existing proposals already defined in the literature, (Ploux & Victorri 1998) among others.

yéti(3) abominable(2) homme(1) des neiges(1)

(Yeti / abominable snowman)

These intuitive examples show the interest of investigating compounding and lexicalization mechanisms. In future work, we intend to evaluate how to automatically relate semantic compositionality or opacity (which are not trivial to measure) to word difficulty.

5.2 NLP for building specialized lexicons

As in many disciplines, the use of semi-automatic methods and specific software has become widespread over the last decades. Responsibility for key lexicographic tasks has been transferred from people to computers, especially for those tasks at which the computer excels, namely, counting, clustering, treating large amounts of data, extracting patterns, and identifying salient neighborhoods between words, etc.

Since the 1980s, lexicographers benefit from ever-growing volumes of data and either the collection or the analysis of such data has become largely streamlined (the ‘drudgery’ in the words of M. Rundell, 2009). Progress in computational linguistics has permitted a deeper investigation of the data, discriminating surface differences and highlighting more fine-grained representations at the morphological, syntactic or even semantic level (Grefenstette, 1998).

As mentioned in previous sections of this article, statistics computed by machines on large volumes of data have shown interesting results on determining how simple a word can be (frequency effect). However, we show in this paper that more sophisticated measures have to be considered and that NLP methods are useful for obtaining them. In a first step, basic linguistic treatments (tokenizing, lemmatizing and part of speech tagging) allow us to identify lexical units in corpora. Counting phonemes or letters, syllabification or on the consistency sound-script (difference between number of letters and phonemes) are simple tasks for a computer. More difficult tasks may imply the use of computational lexicons with structured information. To give an example, to obtain the number of morphemes, a list of affixes is required, as well as some linguistic knowledge on phonological alternations. Similarly, to identify senses on polysemic words, explicit linguistic knowledge has to be gathered on available resources and clustering heuristics have to be implemented to regroup senses. Lastly, as we have shown, the design of a language model is crucial to predict the level of difficulty of a word by combining and weighting the different predictors over large amounts of data.

Judging from these examples, computational linguistics enables the formalization of fine-grained linguistic phenomena which, in turn, provides a better comprehension of such phenomena. As a result, specialized lexicons with explicit information can be created, for human or automated usages in NLP tasks.

6. Conclusion

In this paper, we presented the first version of a French lexicon of synonyms graded with a tag indicating the level of difficulty (*ReSyf*). The data and the tags were obtained from existing resources and from a lexicon difficulty model based on a set of lexical measures. Such measures describe fine-grained intra-lexical features as well as some statistical or psycholinguistic properties of words.

Although we present preliminary work, our contribution demonstrates that natural language techniques can be used to create lexical resources with specific information (in this case, the difficulty levels) gathered and tested over different kinds of corpora.

Yet, there remain important aspects that have to be taken into consideration. We already mentioned that a more accurate sampling of the levels in Manulex is required to refine the gold-standard list. Ideally, a more precise training resource should be obtained through large scale subject testing. In addition, some variables that we introduced have yet to be implemented and integrated to our model. Finally, we also highlighted the importance of a more semantic-oriented approach to the lexicon complexity (as word forms are ambiguous).

To conclude, our future research will continue to focus on the identification of the features that make words easier for a given population class (in particular populations with language impairments) as well as on the automatic assessment word difficulty. We thus foresee a comparison of pedagogical data with pathological data to obtain deeper insights, while adapting the model to take the senses into account. Finally, we expect to use *ReSyf* in the context of automatic text simplification. The integration of a graded resource of synonyms indeed seems likely to impact the efficiency of such systems.

7. Acknowledgements

This project is partly financed by the Programme Hubert Curien (PHC) Tournesol 2013 (France-Fédération Wallonie-Bruxelles).

8. References

- Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval : Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4(4):439–461.
- Biran O., Brody S. & Elhadad, N. (2011). Putting it simply: a context aware approach to lexical simplification. Proceedings of the 49th *Annual Meeting of the Association of Computational Linguistics (ACL 2011)*, pages 496-501. Portland, Oregon.
- Boser, B. and Guyon, I. et Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

- Bormuth, J. (1966). Readability: A new approach. *Reading research quarterly*, 1(3):79–132.
- Brybaert, M., Lange, M. and Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1):65–85.
- Calzolari, N., Gurevych, I. and Kim, J. (2013). *The People's Web Meets NLP: Collaboratively Constructed Language Resources* annotated edition. Springer-Verlag Berlin and Heidelberg GmbH & Co. K.
- Collins-Thomson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. Proceedings of *Human Language Technologies (HLT-NAACL)*, pages 193-200.
- Dale, E. (1931). A comparison of two word lists. *Educational Research Bulletin*, 10(18): 484–489.
- Dale, E. and Chall, J. (1948). A formula for predicting readability. *Educational research bulletin*, 27(1):11-28.
- De Belder, J. and Deschacht, K. (2010). Lexical Simplification. Proceedings of the *1st International Conference on Interdisciplinary Research on Technology, Education and Communication (ITEC 2010)*. Kortrijk.
- Ferrand, L. (2007). *Psychologie cognitive de la lecture*. De Boeck, Bruxelles. (ISBN-13 9782804159030).
- François, T. (2012). Lexical and syntactic complexities: a difficulty model for automatic generation of language exercises in FFL. PhD thesis. Université Catholique de Louvain, Louvain-la-Neuve.
- François, T. and Fairon, C. (2012). An “AI readability” formula for French as a foreign language. In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012), Jeju, 466-477.
- Gala, N. and Lafourcade, M. (2011). NLP lexicons: innovative constructions and usages for machines and humans. Proceedings of Electronic Lexicography (E-Lex 2011). Bled (Slovenia).
- Gale, W. and Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.
- Gernsbacher, M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology : General*, 113(2):256–281.
- Grefenstette, G. (1998). The Future of Linguistics and Lexicographers: Will there be Lexicographers in the Year 3000?, in Fontenelle et al. (Eds) Proceedings of the Eighth *EURALEX Congress*. Liege: University of Liege: 25-41. Reprinted in Fontenelle, T (Ed.) *Practical Lexicography: A Reader*. OUP 2008.
- Gougenheim G. (1958). *Dictionnaire fondamental de la langue française*, Paris : Didier. (ISBN 2-208-00133-8).
- Howes, D. and Solomon, R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.
- Lafourcade, M. (2007). Making people play for Lexical Acquisition. Proceedings of

- the 7th *Symposium on Natural Language Processing (SNLP-2007)*. Pattaya, Thaïlande, 8 pages.
- Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In S CHMITT, N. et M C C ARTHY, M., editors: *Vocabulary: Description, Acquisition and Pedagogy*, pages 140–155. Cambridge University Press, Cambridge.
- Lété, B., Sprenger-Charolles, L. and Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments & Computers*, 36, 156-166.
- Levelt, W. J. M. (1989). *Speaking*. Cambridge, MA: MIT.
- Ogden, C. K. (1930). *Basic English: A General Introduction with Rules and Grammar*. Cambridge.
- McNamara, P. (2010). Parkinson's Disease-Related Speech and Language Problems. Retrieved April 3, 2013, from http://parkinsons.about.com/od/signsandsymptomsfpd/a/speech_problems.htm
- Morrisson, C. and Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1):116–133.
- New B., Pallier, C., Ferrand, L. and Matos R. (2005). Une base de données lexicales du français contemporain sur Internet: Lexique. *L'Année Psychologique*, 101, 447-462.
- O'Regan, J. and Jacobs, A. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):185–197.
- Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.
- Pinto, S. and Ghio, A. and Teston, B. and Viallet, F. (2010) La dysarthrie au cours de la Maladie de Parkinson. Histoire naturelle de ses composantes: dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, vol. 166, no. 10. 2010, p. 800-810.
- Ploux, S. and Victorri, B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, vol. 39(1) :161-182.
- Rundell, M. (2009). The road to automated lexicography: First banish the drudgery... then the drudges? In *Proceedings of eLexicography in the 21st Century Conference*, Louvain-la-Neuve, Université Catholique de Louvain.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.
- Schreuder R. and Baayen R. H. (1997). How simplex complex words can be. *Journal of Memory and Language* 37, 118-139.

Taylor, W. (1953). Cloze procedure : A new tool for measuring readability. *Journalism quarterly*, 30(4):415 433.

University of Groningen (2011). Parkinson's disease undermines language processing. *ScienceDaily*. Retrieved April 3, 2013, from <http://www.sciencedaily.com/releases/2011/02/11020262.htm>

Going Online with a German Collocations Dictionary

Tobias Roth

University of Basel, Deutsches Seminar, Nadelberg 4, Basel, Switzerland
tobias.roth@unibas.ch

Abstract

Although a lot of dictionaries are available on the Web, there are no well-established ways to present collocations dictionaries for language learners online. In the online version of the collocations dictionary for German we are working to overcome certain shortcomings of printed collocations dictionaries. A major issue when they are used in production situations (e.g. writing processes) is how to find collocations efficiently. Another difficulty for users is to transfer the information found to their own language use. The lexicographic challenge consists of conceiving a microstructure that assists users in finding a collocation without having to read complete articles. At the same time, enough information has to be given in order for learners to be able to use a collocation appropriately. Our online dictionary uses present-day electronic search facilities for improved access, as well as a presentation of dictionary articles on two levels: a minimalistic view for the search and navigation stage and a more detailed view once a collocation is found.

Keywords: online dictionary, collocations, dictionary design, learners' dictionary, German language

1. Introduction

Many dictionaries are available on the Web today. However, as yet there are no well-established ways of how to present collocations dictionaries for language learners online.

Major issues are retrievability and information transfer. How can a collocation be found efficiently, i.e. without needing to read complete dictionary articles? And how is the information best presented so that users understand the entries and can effectively use collocations found in the dictionary themselves?

The *Kollokationenwörterbuch*¹ – the collocations dictionary we are working on – is not a pure online project. The dictionary is also intended to appear in print. Not all aspects, therefore, are optimised for the online version. Some decisions reflect a compromise between these two mediums as it would not make sense to duplicate certain structures because of slightly differing needs between online and print versions. However, as it is a completely new dictionary built from scratch, it was possible to freely choose the design of the underlying database and special attention

¹ Its full (working) title being *Kollokationenwörterbuch – typische und gebräuchliche Wortverbindungen des Deutschen*, it is accessible at <http://www.kollokationenwoerterbuch.ch>.

was paid to the possibility of media-independent publishing.

2. Collocations in online dictionaries

Most learners' dictionaries specifically dedicated to collocations are not available online² or are available only in e-book versions as electronic copies of printed dictionaries, such as Quasthoff (2011) for German. However, there are some more general online dictionaries that cover collocations and multi-word units in general. Examples for German are *Duden online*, *ellexiko* (Klosa, Schnörch & Storjohann 2006), *DWDS* (Geyken 2011; Klein 2004) and *LEO*.

These dictionaries use different strategies for presenting multi-word units. They range from writing simple listings, separate article entries for each unit and more elaborate visualisation methods, such as word clouds and network graphs.

2.1 Separate entries

In *ellexiko*'s sub-dictionary "feste Wortverbindungen", every multi-word unit gets its own article entry. Most have at least slightly idiomatic meanings, so detailed explanations are clearly justified.

For collocations that are semantically compositional, this structure is less appropriate. This would result in very small and relatively uninformative article entries, whereas important aspects such as the context of a collocation (whether there are similar collocations with the same component words) or its retrievability will be neglected.

2.2 Listings

Other dictionaries give simple listings of collocations or multi-word units for a headword. In certain cases they are presented like usage samples, although a majority of these sample combinations possess collocational characteristics and would enter a collocations dictionary (cf. e.g. *Duden online*, the dictionary part of *DWDS*).

Often such listings are not further hierarchically structured. In cases where they are, criteria are often syntactic. For example, the *Wortprofil* (word profile) in *DWDS* groups collocations by their syntactic configuration (there are groups for collocates as subjects, objects, attributes etc.; see also Geyken 2011).

Listings are easy to produce and can potentially display large numbers of collocations in a limited space, but as it becomes more extensive, navigation can become difficult.

² *DiCE*, an online collocations dictionary for Spanish, can be cited as one of the few exceptions.

2.3 Word clouds and network graphs

Word clouds and network graphs are more sophisticated tools to visualise collocations of a given headword. Word clouds are used by *Wortprofil* in *DWDS* and network graphs by *Wortschatz Leipzig*; while *Duden online* uses a combination of word clouds and network graphs to display typical word combinations.

Both word clouds and network graphs are preferred for automatically extracted collocation lists. They are hardly ever found in manually crafted articles. The advantage of both of them is a rather compact mode of presentation and the possibility to visualise context and frequencies and the strength of connections.

3. Issues in collocations lexicography

The aforementioned general online dictionaries are obviously not specialised in the presentation of collocations. To obtain a clearer idea about the difficulties one has to deal with in online collocations dictionaries, it is best to start with the analysis of the main issues in collocations lexicography for language learners.

The present project, like many other collocations dictionaries, is perceived to be an aid in text or language production. The prototypical user wants to write or say something about, e.g. a *mountain*, knows that this is *Berg* in German, and expects to find collocations with *Berg* (under the headword *Berg*) that match the meaning he/she wants to convey.

The two main problems here are navigation and information transfer.³ How should collocations be arranged in the dictionary so they can be retrieved as easily and efficiently as possible? And in what form should the information be provided for users to be able to actually integrate a collocation found in the dictionary into their own speech and writing?

3.1 Retrieval

How collocations are best retrieved is by no means a trivial question. If we consider collocations as transparent and essentially compositional in meaning⁴ we can assume that users will be able to look up a collocation under one of its component words. If a headword comprises a large number of collocations the next question is how to group and sort them to ease the search process.

³ Issues no less important, but more closely related to content, e.g. selection criteria for collocations or integration of compounds (cf. Häcki Buhofer 2011; Roth 2012b), are not discussed here.

⁴ As *idioms of encoding* (Fillmore, Kay & O'Connor 1988; Makkai 1972).

3.1.1 Node and collocate vs. base and collocator

Hausmann (1985) introduced the concept of *base* and *collocator* in collocations. The formerly used terms *node* and *collocate* (Sinclair 1966) just indicate a perspective: *node* is the word that is being looked at and its *collocates* are the partner words that form collocations with it. All components of a collocation can be both *node* and *collocate*, just depending on the perspective.

In contrast, *base* and *collocator* describe an absolute hierarchy within a collocation. Rather vaguely defined, the *base* is the word a prototypical user would look up in order to find a collocation; the *collocator* its counterpart. According to Hausmann (1985), the noun is the most important word class for *bases* because nouns denote things and phenomena in the world that we talk about:

Die wichtigste Basiswortart ist das Substantiv, weil es die Substantive sind, welche die Dinge und Phänomene dieser Welt ausdrücken, über die es etwas zu sagen gibt. Adjektive und Verben kommen als Basiswörter nur insoweit in Frage, als sie durch Adverbien weiter determiniert werden können. (Hausmann 1985, p. 119).

In verb-noun collocations the *base* is the noun; in adjective-noun collocations it is also the noun; in verb-adverb collocations it is the verb, etc. Even if the concept has its problems (cf. e.g. Handl 2009; Herbst 2009; Roth 2012b; Steyer 2000) it has been widely adopted by current collocations dictionaries (Le Fur 2007; Lo Cascio 2012; OCDSE 2009; Quasthoff 2011; Rundell 2010). In printed dictionaries it allows for a reasonable navigation structure without the need of duplicated entries: collocations are printed in the base article only, not under the collocator.

3.1.2 Grouping and sorting

Several proposals have been made on how to arrange collocations within an article. Grouping and sorting criteria are mainly morphosyntactic, syntactic and semantic. As outlined above, the search process on this level is semantically motivated: users look for a collocation that fits, as closely as possible, the meaning they want to express. They might have an idea of how the construction of the whole sentence will appear, hence the morphosyntactic and syntactic criteria, but essentially it is a semantic choice.

Most collocations dictionaries have at least two grouping levels below the headword.⁵ Quasthoff (2011) groups according to word class (verb, adjective). Noun-verb collocations are subgrouped according to the grammatical case of the base noun, whereas collocations with adverbs and adjectives contain semantically motivated subgroups. The OCDSE (2009) and Rundell (2010) both consider word class groups on the top level, but include positional information (e.g. *X + verb* vs. *verb + X*). Subgroups are semantically motivated; in the case of Rundell (2010) the content of a

⁵ The exact number depends on whether the splitting of different meanings of a headword is considered as a grouping level or not.

semantic subgroup is explicitly stated. Le Fur (2007) also forms groups by word class, positional information and semantically motivated subgroups. Finally, Lo Cascio (2012) forms the same top level groups (word class and positional information), but no sub-groups; instead, the collocations are in alphabetical order.

3.2 Information transfer

Once a suitable collocation is found, a user needs to know its exact form and properties so as to be able to actually use it. In a preliminary study on article structure conducted at different schools, students preferred less abstract citation forms and articles with more example sentences (Siebenhüner 2010). They often displayed difficulties in deriving the correct usage of a collocation from collocators only or from abstract citation forms without examples.

This need to be more explicit in order to facilitate information transfer contradicts in some ways the need to be as compact as possible in order to facilitate navigation and retrieval. Most current collocation dictionaries focus on compactness rather than on explicit information presentation.

The majority provide the base form of the collocator but no citation form more explicit (Le Fur 2007; OCDSE 2009; Quasthoff 2011; Rundell 2010). An exception to this is Lo Cascio (2012) who provides extended citation forms. Others try to convey grammatical information mainly by their structure of groups and subgroups (see above). Le Fur (2007) additionally indicates certain grammatical or other features by means of abbreviations in superscript next to the collocator. Example sentences are given by most of the dictionaries quoted above. Exceptions are Quasthoff (2011) who gives no example sentences at all and Lo Cascio (2012) with explicit meaning indications for a big part of the collocations.

4. A German online collocations dictionary

The present project consists of creating a German collocations dictionary with collocations of about 2000 base-vocabulary headwords (Häcki Buhofer 2011; Roth 2012b). The primary target audience is intermediate L2 learners of German. The dictionary is not an online-only project; there will also be a printed version.

The dictionary is completely new, written from scratch, so there was no need to consider the integration of older versions or other kinds of legacy data. The dictionary writing system in use has also been newly developed for this specific purpose. This offered the possibility to structure the data in such a way that would allow media-independent publication (Roth 2012b). Online and printed presentations are not completely independent, however, as they share certain common features. On one hand they share the same needs concerning some points, whereas on the other hand it would often be highly uneconomical to duplicate features with only slight differences between online and printed versions. Sometimes

there is a solution that is suitable for both versions, even if there was a more ideal solution for a particular format, and in such cases a common approach is utilised for both.

A prototype of the online version of the *Kollokationenwörterbuch* can be found under the URL <http://www.kollokationenwoerterbuch.ch>. Its main characteristics, and some proposals for solutions to the presentation issues raised above, are described below.

4.1 Search

The main means of interaction with the dictionary is a simple search field, similar to the familiar Web search engine types. When typing a word into the field, matching lemmas show up in a menu list underneath in an ‘as-you-type’ fashion (see Figure 1).

At the top of the list there are words beginning with the search term, whereas below you can find words containing the search term. Lemmas that are part of the 2000-item base vocabulary appear in bold. For these lemmas a complete collocation search including manual semantic grouping (see 4.2.) has been performed. Lemmas not in bold appear in collected collocations, but they have not been treated as a headword for the printed version and they have not undergone a systematic collocation search. These articles are dynamically assembled. As no manual semantic grouping has taken place in these cases, collocations are presented alphabetically, grouped by word class.



Figure 1: Search and navigation

If you type more than one word in the search field, the dictionary article for the first search word is fetched and subsequent search strings are highlighted in the just-loaded

article (see Figure 2-c).

Such standard search functionality for an online application helps in overcoming the problem of whether it is reliably the base that is looked up. Access through collocators is possible, also, and all collocations belonging to a word are directly shown. Articles do not strictly follow the base-collocator principle anymore, but rather show a node-collocate approach. Yet, the overall structure of an article is not greatly changed because of this. What would otherwise appear as links to other articles are now presented as full collocation entries, but displayed in a separate grouping at the end of the article.

In general, the possibility to easily search by all collocation components is a big improvement in retrievability.

4.2 Grouping and navigation layer

Once one component word (*node*) is found along with its associated collocations, the next question or challenge is how to find a suitable collocation without having to read the complete article.

In the present project, it was decided to introduce two hierarchical grouping levels. In a first step, collocations are grouped by the word class of their collocates (see Figure 2 b). Subsequently, they are further subgrouped according to semantic criteria. Collocations belonging semantically together can be found in the same subgroup. A subgroup may receive a label (see Figure 2e) that describes its content or is at least associated with the collocates of this subgroup and stands as a kind of a prototypical example. Its goal is not the meaning description proper, but to assist in navigation.⁶

This also holds for the printed version. The main difference introduced in the online dictionary is a split into two presentation layers (see Figure 2). On the first layer, still in the navigation stage, only collocates are displayed. All supplementary information, such as extended citation forms, example sentences, meaning indications, etc., is omitted. The collocates are displayed in boxes grouped by semantic similarity. With this layout, a maximum of collocations fit on one screen in a clearly arranged fashion. This should help users to more quickly find the collocations they seek. With only one word per collocation a strict minimum of information is provided with no extra information to detract from the retrieval task.

4.3 Detailed information

The second layer presents all collected information for a collocation. As soon as a suitable collocation is found the one-word-per-collocation approach has reached its goal and is then no longer informative enough. The user's next task is to find out how

⁶ Not like in Rundell (2010) where actual meaning descriptions for every subgroup are given.

exactly to use this collocation. Studies conducted in this project (Siebenhüner 2010) have confirmed that extended citation forms and a large number of example sentences are necessary for many students so that they can correctly use collocations they have looked up.

Along with extended citation forms and example sentences some additional detailed information is presented here. Some collocations that might be difficult to understand are given meaning indications. Pragmatical usage information is also provided (markers such as *informal*, *pejorative*, etc., but also more detailed usage explanations when considered necessary). Collocations are also marked for regional usage restrictions on a country level for Austria, Germany and Switzerland (Roth 2012a).

This detailed information on the second layer is accessed by a clicking or hovering action on the single collocates causing an expansion of the details window (Figure 2d). In addition, this approach has the advantage that users interact with the dictionary application with more active involvement than just by plain reading.

4.4 Internal and external links

The detailed view of a collocation provides links to internal and external targets. For the time being this feature is not extensively used; so far there are internal links to other dictionary articles and external corpus links.

Internally, collocates in the detailed view are linked to the corresponding node articles. Clicking on a collocate will open an article with the respective word as a node. If, in the example article in Figure 2, the user is unsure of the exact meaning of *ausrücken* they can click on it and navigate to the article for *ausrücken*. There, the user will find collocations that inform that the word describes something that the fire brigade (*Feuerwehr*) and the police (*Polizei*) do. If the user required a collocation describing the arrival of the fire brigade they can now click on *Feuerwehr* to obtain several verbs that can be used for this purpose.

The first external links given have the *Swiss Text Corpus* (Bickel et al., 2009) as a target. These links will open the corpus site with a *KWIC* view (key word in context) of examples for the respective collocation.

Besides links to more corpora, links to other dictionaries could also prove useful. Collocations dictionaries do not provide information about individual words, such as meaning indications, grammatical information beyond citation forms and examples as well as pragmatical usage information, which can be seen as a shortcoming. Links from individual words to a general dictionary or even to a bilingual dictionary could help users in this case (but do not form part of the current version of the *Kollokationenwörterbuch*).

4.5 Customisability

An advantage of online dictionaries and electronic dictionaries in general is that they are more dynamic. Potentially, everyone can have their own, tailor-made version of a dictionary in terms of what data are displayed and how they are displayed.

However, users also expect an online dictionary to work ‘out-of-the-box’. Their first concern will typically not be how they can customise it. In addition, it is often not very clear what special features users might expect from an online dictionary, as Müller-Spitzer, Koplenig & Töpel (2011) put it:

Nevertheless, this does not mean that the development of innovative features of online dictionaries is pointless. As we show elsewhere in detail [...], users tend to appreciate good ideas, such as a user-adaptive interface, but they are just not used to online dictionaries incorporating those features. As a result, they have no basis on which to judge the usefulness of those features. (Müller-Spitzer, Koplenig & Töpel 2011, p. 270)

Customisability in the online version of the *Kollokationenwörterbuch* is therefore kept on a low level. Users should not be overwhelmed with settings to customise, or with too many features, but they should have certain possibilities to influence the behaviour of the user interface.

The screenshot shows a web interface for a dictionary article on 'Brand'. At the top, there is a search bar containing 'brand schwer' and a magnifying glass icon. Below the search bar are two checkboxes: 'Bsp./Infos' (checked) and 'ganze Form'. To the right of the search bar are navigation links: 'Projekt', 'Aktivitäten', 'Grundlagen', 'Ergebnisse', 'Team', and 'Kontakt'. The main content area is titled 'Brand m' and is divided into several sections: 'ADJEKTIVE/ADVERBIEN', 'VERBEN', 'NOMEN', 'PHRASEN', and 'ZUSAMMENSETZUNGEN'. Each section contains a list of related terms and phrases. Annotations (a-e) are placed throughout the page: (a) is near the search bar; (b) is near the title 'Brand m'; (c) is near the adjectives 'klein', 'lodernd', 'schwer', 'zahlreich'; (d) is near the verb 'ausrücken'; (e) is near the verb 'brennen'. The interface is clean and organized, with clear sections and a consistent layout.

Figure 2: Dictionary article

Instead of the default two-level presentation outlined above, users can switch to a view where all information (examples, meaning indications, etc.) is displayed on one level (see Figure 2a). They can also toggle the display of collocates and extend citation forms. This gives users a view that resembles more the printed version of the dictionary.

4.6 Extensions

Possibilities to further extend the functionality of the dictionary⁷ include, of course, the aforementioned linking of additional dictionary sources. Linkage from other (dictionary) sites to *Kollokationenwörterbuch* could also help to put it into a more general context, a bit detached from its status of a rather specialised dictionary and towards that of a tool commonly and readily used when writing. A Web service interface would greatly facilitate integration into other websites.

Another obvious enhancement, and probably the next step in further development, would be a version optimised for mobile devices, either as a mobile app on its own or just as a mobile-friendly version of the dictionary site.

Since a primary target audience of the *Kollokationenwörterbuch* are people producing text in writing, another promising possibility would be direct integration of the dictionary into text editors (as an add-on or plug-in). Just like they already get synonyms and spelling errors, authors could get collocates for a given word.

More ideas for extensions might come up with user feedback as soon as the dictionary site has been running for some time.

5. Conclusion

The *Kollokationenwörterbuch* is one of the first specialised collocations dictionaries for learners that has a dedicated online user interface. This user interface is the main topic of the present contribution.

Solutions have been proposed to two main problems of production-oriented collocations dictionaries. The problem of retrievability and navigation is tackled by a search facility over all the component words of the collocations, as well as with a two-level presentation that hides detail in the first step. Semantically motivated grouping is another feature likely to help in navigation.

The problem of information transfer, i.e. how to actually use a collocation that has been looked up, has been of great concern in the conception of the microstructure. Measures taken include explicit citation forms, meaning and usage indications and many example sentences.

⁷ See also Roth (2012b).

In general, the online version of the *Kollokationenwörterbuch* should take the discussion on how collocations dictionaries should be presented online a step further.

6. References

- Bickel, H., Gasser, M., Hofer, L. & Schön, C. (2009). Das Schweizer Textkorpus. In *Linguistik online* 39.3, pp. 5–31.
- DiCE. Diccionario de colocaciones del Español*. Accessed at: <http://www.dicesp.com>.
- Duden online*. Accessed at: <http://www.duden.de>.
- DWDS. Digitales Wörterbuch der Deutschen Sprache*. Accessed at: <http://www.dwds.de>.
- elexiko. Online-Wörterbuch zur deutschen Gegenwartssprache*. Accessed at: <http://www.elexiko.de>.
- Fillmore, C. J., Kay, P. & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. In *Language* 64.3, pp. 501–538.
- Geyken, A. (2011). Die dynamische Verknüpfung von Kollokationen mit Korpusbelegen und deren Repräsentation im DWDS-Wörterbuch. In Klosa, A. & Müller-Spitzer, C. (eds.) *Datenmodellierung für Internetwörterbücher*. OPAL 2/2011. Mannheim: Institut für deutsche Sprache.
- Häcki Buhofer, A. (2011). Lexikografie der Kollokationen zwischen Anforderungen der Theorie und der Praxis. In Engelberg, S., Holler, A. & Proost, K. (eds.) *Sprachliches Wissen zwischen Lexikon und Grammatik. Jahrbuch des Instituts für Deutsche Sprache 2010*. Berlin: De Gruyter, pp. 505–531.
- Handl, S. (2009). Towards Collocational Webs for Presenting Collocations in Learners' Dictionaries. In Barfield, A. & Gyllstad, H. (eds.) *Researching Collocations in Another Language. Multiple Interpretations*. Basingstoke: Palgrave Macmillan, pp. 69–85.
- Hausmann, F. J. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Bergenholtz, H. & Mugdan, J. (eds.) *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch, 28.–30.06.1984*. Tübingen: Niemeyer, pp. 118–129.
- Herbst, T. (2009). Item-Specific Syntagmatic Relations in Dictionaries. In Nielsen, S. & Tarp, S. (eds.) *Lexicography in the 21st Century: In Honour of Henning Bergenholtz*. Terminology and Lexicography Research and Practice 12. Amsterdam & Philadelphia: John Benjamins, pp. 281–308.
- Klein, W. (2004). Das digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts. In Scharnhorst, J. (ed.) *Sprachkultur und Lexikographie. Von der Forschung zur Nutzung von Wörterbüchern*. Frankfurt am Main: Lang.
- Klosa, A., Schnörch, U. & Storjohann, P. (2006). ELEXIKO – A lexical and lexicological corpus-based hypertext information system at the Institut für Deutsche Sprache, Mannheim. In Marello, C. et al. (eds.) *Proceedings of the 12th EURALEX International Congress, Turin, Italy*, pp. 425–430.
- Kollokationenwörterbuch. Typische und gebräuchliche Wortverbindungen des*

- Deutschen*. Accessed at: <http://www.kollokationenwoerterbuch.ch>.
- Le Fur, D. (ed.) (2007). *Dictionnaire des combinaisons de mots*. Paris: Le Robert.
LEO. Accessed at: <http://www.leo.org>.
- Lo Cascio, V. (2012). *Dizionario combinatorio compatto Italiano*. Amsterdam: John Benjamins Publishing Company.
- Makkai, A. (1972). *Idiom Structure in English*. The Hague: Mouton.
- Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. In Kosem, I. & Kosem, K. (eds.) *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 203–208. Accessed at: http://www.trojina.si/elex2011/elex2011_proceedings.pdf.
- OCDSE (2009). *Oxford Collocations Dictionary for Students of English*. Compiled by C. McIntosh. 2nd ed. Oxford: Oxford University Press.
- Quasthoff, U. (ed.) (2011). *Wörterbuch der Kollokationen im Deutschen*. Berlin: De Gruyter.
- Roth, T. (2012a). Using Web Corpora for the Recognition of Regional Variation in Standard German Collocations. In Kilgarriff, A. & Sharoff, S. (eds.) *Proceedings of the Seventh Web as Corpus Workshop (WAC7). Pre-WWW2012 Workshop, 17 April, 2012*, pp. 31–38. Accessed at: <https://sigwac.org.uk/raw-attachment/wiki/WAC7/wac7-proc.pdf>.
- Roth, T. (2012b). *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. PhD thesis. Universität Basel.
- Rundell, M. (ed.) (2010). *Macmillan Collocations Dictionary*. Oxford: Macmillan Education.
- Siebenhüner, S. (2010). *Kollokationenwörterbuch: Schulstudie*. Universität Basel, Praktikumsbericht. Unpublished.
- Sinclair, J. (1966). Beginning the Study of Lexis. In Bazell, C. E. et al. (eds.) *In Memory of J. R. Firth*. London: Longman, pp. 410–430.
- Steyer, K. (2000). Usuelle Wortverbindungen des Deutschen. In *Deutsche Sprache* 2/00, pp. 101–125. *Wortschatz-Portal Universität Leipzig*. Accessed at: <http://wortschatz.uni-leipzig.de/>.

TERMIS: A corpus-driven approach to compiling an e-dictionary of terminology

Nataša Logar¹, Iztok Kosem²

¹University of Ljubljana, Faculty of Social Sciences, Ljubljana, Slovenia

²Trojina, Institute for Applied Slovene Studies, Škofja Loka, Slovenia

E-mail: natasa.logar@fdv.uni-lj.si, iztok.kosem@trojina.si

Abstract

This paper describes the process of compiling an online dictionary of terminology within the TERMIS project. The compilation began from a morphosyntactically tagged synchronous LSP corpus and involved automatic term recognition performed for single- and multi-word terms with the LUIZ term extraction system and the automatic extraction of lexical information from the corpus via the Sketch Engine tool. The information obtained, along with the results of the GDEX system, was imported into the dictionary editing system to the Termania web portal. A free online terminological database of the public relations field comprised of 2000 entries has been publicly available since July 2013.

Keywords: terminology, corpus, database, public relations, Termania

1. Introduction

Due to the continuous growth of scientific research, all disciplines must assure the development of terminology in their own language. In the case of Slovene, terminology development is connected to the importance of native language. Several terminological dictionaries have been published in Slovenia in the last few decades; however, there still remains a need for terminology description in many different disciplines.

New challenges in terminology have arisen as a result of the Bologna Reform and system of internationalization of higher education that, among other things, promote frequent exchange between students, lecturers, and researchers (Kalin Golob & Stabej, 2007; Humar & Žagar Karer, 2010; Kalin Golob, 2012; Kalin Golob et al., 2012). Interpretation of internationalization in its narrow sense, i.e. an increase in the number of university programs taught in English, implies a resulting abandonment of Slovene as a language of instruction in higher education. As a result, there are now more and more warnings of such a practice turning into a situation in which “Slovene would eventually become a language in which some disciplines would no longer have, or would no longer develop, its own terms, and the communication would be conducted in a foreign language only” (Humar & Žagar Karer, 2010: 9).

One of the solutions to this problem is to provide Slovene terminology with contemporary reference materials, namely terminological dictionaries and databases. This paper describes the development of a terminological database within the TERMIS project, which consisted of six key phases: (a) a corpus, (b) automatic extraction of term candidates, (c) automatic extraction of collocations and grammatical relations, (d) extraction of good examples, (e) data editing and (f) final online visualization of entries.

2. TERMIS

An applied research project titled *Terminology data banks as the bodies of knowledge: The model for the systematization of terminologies* (TERMIS; <http://www.termis.fdv.uni-lj.si/>) was conducted between July 2011 and June 2013, funded by the Slovenian Research Agency. The aim of the project was the compilation of an online terminological dictionary of public relations, with two specific objectives:

a) The development of a freely accessible online dictionary-like terminological database for the discipline of public relations. The database contains 2000 terms with definitions, English translations, and typical collocations. Each entry is linked with a specialized corpus of public relations texts called KoRP (http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl; Logar, 2007) and Gigafida, a reference corpus of Slovene (<http://www.gigafida.net>; Logar Berginc et al., 2012).

b) The development of an online dictionary editing system that is easy to use so that an expert in the field, i.e. a terminologist, can start using it without any prior knowledge. Dictionary writing systems are freely available on the Termania online portal (<http://www.termania.net>; Romih & Krek, 2012; Kompara & Holozan, 2011: 145).

This paper focuses on the first objective only.

3. Corpus

The basis of the project was KoRP, a corpus of public relations texts. The corpus contains 1.8 million words and is a monolingual and synchronous specialised corpus. The corpus has been freely accessible online since it was completed in July 2007. Recently, the corpus was lemmatized and morphosyntactically tagged with the latest statistical tagger for Slovene, called Obeliks (<http://oznacevalnik.slovenscina.eu/Vsebine/Sl/SpletniServis/SpletniServis.aspx>; Grčar, Krek & Dobrovoljc, 2012). The texts in the KoRP corpus were selected according to carefully designed criteria (Logar, 2007), which make the corpus representative of a public relations field in Slovenia.

4. Term extraction

There are many approaches to extraction of term candidates from specialized corpora. Almost all of them use a combination of linguistic knowledge of terms, and mathematical statistics on word and word sequence distribution in corpora (Vintar, 2008: 100; Vintar, 2009: 346–347 and literature therein cited). Using the LUIZ term extraction tool (<http://lojze.lugos.si/cgitest/extract.cgi>; Vintar, 2010) we have extracted from the KoRP corpus:

a) single-word term candidates: nouns, verbs, adjectives, and adverbs;

b) multi-word term candidates: noun phrases and verb phrases.

Both single- and multi-word term candidates have been extracted using morphosyntactic patterns and term weight, calculated by comparing the frequency in the KoRP corpus and the frequency in a general corpus, in our case FidaPLUS, a reference corpus of Slovene (<http://www.fidaplus.net>; Arhar Holdt & Gorjanc, 2007), and phraseological stability of an extracted terminological unit. We have identified 39 morphosyntactic patterns in total: 30 with a noun as a headword, 9 with verb as a headword. The result of the extraction was lists with 47,007 multi-word units (excluding proper nouns) and 16,190 single-word units (excluding proper nouns).

The lists were carefully analyzed and evaluated in order to determine the successfulness of the extraction method. This highlighted two issues:

a) When the top part of the list containing extracted term candidates was compared with the top parts of the noun and verb frequency lists in KoRP, we noticed only minor differences, but all in favour of the lists of extracted terms; in other words, the lists with extracted terms offered better results. Our expectations were thus confirmed, so we subsequently decided to use only automatically extracted lists of term candidates for building our headword list.

b) The analysis of the top 100 units on the lists of all 30 multi-word patterns containing a noun headword showed that the terminologically most productive patterns were *Adj N*, *Adj and Adj N* and *Adj Adj N*. Over 50% of the analyzed extracted units in the lists were proper terms and thus relevant for our headword list (see Table 1).

We were able to obtain 2000 terms for the dictionary headword list by analyzing 3000 items on the lists containing single-word noun term candidates, and 4000 items on the list of multi-word term candidates (using all 30 patterns).¹ The analysis

¹ The extraction of adjectives, adverbs, and verb phrases did not yield terminologically relevant results, so they are not discussed in this paper.

was conducted by a terminologist and two experts in the field of public relations. In the next phase of the project, we automatically extracted lexical information for the words and multi-word units (e.g. compounds) on the created headword list.

Pattern	Number of terms in the top 100 units on the list	Example
Adj N	87	<i>blagovna znamka</i>
Adj and Adj N	62	<i>notranja in zunanja javnost</i>
Adj Adj N	45	<i>integrirano marketinško komuniciranje</i>
Adj N S N	20	<i>vladni odnosi z javnostmi</i>
Adj N and N	17	<i>strateško načrtovanje in upravljanje</i>
N Adj N	17	<i>upravljanje žgočih problemov</i>
R Adj N	11	<i>cenovno občutljiva informacija</i>
N S N	7	<i>odnosi z javnostmi</i>
N Adj Adj N	6	<i>model dvosmernega asimetričnega komuniciranja</i>
N N	6	<i>vir informacij</i>

Table 1: The 10 terminologically most productive patterns containing multi-word term candidates with a noun headword.

5. Automatic extraction of lexical information

When reporting on the compilation process of a new Lexical Database for Slovene (<http://www.slovenscina.eu/spletni-slovar/leksikalna-baza>; Gantar, 2009; Gantar & Krek, 2011), Kosem, Gantar & Krek (2012: 118) said:

The decision to use automatic extraction of lexical information from the corpus /.../ comes from the need to reduce time and costs connected with the production of dictionaries, by utilizing new possibilities offered by state-of-the-art tools for corpus analysis.

Due to these very reasons, combined with the fact that we collaborated on the TERMIS project, as well as the *Communication in Slovene* project (<http://www.slovenscina.eu/projekt>), where this lexical description of contemporary Slovene has been produced, we used the method of Kosem, Gantar & Krek (2012) in our TERMIS project for extracting lexical information (syntactic relations, collocations, and examples) for single and multi-word terms from the KoRP corpus. The method uses the Sketch Engine tool and its Word sketch function (<http://www.sketchengine.co.uk/>; Kilgarriff et al., 2004; Kilgarriff & Kosem, 2012), so we had to prepare and upload the KoRP corpus in our local installation of the Sketch Engine. Due to the different nature of the project, and the corpus, some changes were necessary in the extraction algorithm and its constituent parts. For example, Sketch Grammar was slightly adapted (Krek, 2012), new GDEX (Good Dictionary Examples) configurations for good example extraction were prepared, and minor tweaks to API script (Application Programming Interface) were made (Kosem,

Gantar & Krek, 2012; Kilgarriff et al., 2008; Kosem, Husak & McCarthy, 2011). In addition, a new DTD for the Termania dictionary portal was prepared to enable importing of information in the database, as well as its visualization.

After two test automatic extractions, we divided the terms into 10 different groups according to their frequency/salience values for relations for three groups of terms:

a) single-word terms:

- verbs:
 - group 0: frequency: 1–29
 - group 1: frequency: 30–199
 - group 2: frequency: >200
- nouns:
 - group 0: frequency: 1–19
 - group 1: frequency: 20–99
 - group 2: frequency: 100–699
 - group 3: frequency: >700

b) multi-word terms (adjective + noun, noun + noun):²

- group 0: frequency: 1–9
- group 1: frequency: 10–129
- group 2: frequency: >130

For terms in groups 0, all information available in word sketch was extracted. For other groups, we set four parameters for extraction (minimum collocation frequency, minimum collocation salience, minimum gramrel frequency, minimum gramrel salience) for each grammatical relation (example of settings is shown in Table 2, and an example of information they refer to is shown in Figure 1).³

² Automatic extraction of lexical information for other patterns, e.g. noun + preposition + noun, was not possible at the time.

³ Explanation of values in Figure 1: top number in the second column indicates minimum gramrel frequency (e.g. 299 for the relation *S_kakšen?*), top number in the third column indicates minimum gramrel salience (e.g. 2.3), all the numbers in the second column indicate minimum collocation frequency (e.g. 32 for *spodbujen*), and all the numbers in the third column indicate minimum collocation salience (e.g. 11.51 for *spodbujen*).

imidž (*samostalnik*) KoRP frekvenca = 659 (300.3 na milijon)

S kakšen?		299	2.3	S s-koga-česa		181	2.2
<input type="checkbox"/>	spodbujen	<u>32</u>	11.51	<input type="checkbox"/>	preučevanje	<u>12</u>	10.22
<input type="checkbox"/>	splošen	<u>77</u>	10.77	<input type="checkbox"/>	sij	<u>4</u>	9.41
<input type="checkbox"/>	pozitiven	<u>27</u>	9.84	<input type="checkbox"/>	odsev	<u>4</u>	9.18
<input type="checkbox"/>	korporativen	<u>24</u>	9.5	<input type="checkbox"/>	spodbujanje	<u>8</u>	9.18
<input type="checkbox"/>	želen	<u>11</u>	9.2	<input type="checkbox"/>	izboljšanje	<u>7</u>	8.88
<input type="checkbox"/>	nevtralen	<u>7</u>	9.14	<input type="checkbox"/>	ovrednotenje	<u>4</u>	8.71
<input type="checkbox"/>	šibek	<u>6</u>	8.86	<input type="checkbox"/>	ustvarjanje	<u>9</u>	8.56
<input type="checkbox"/>	konsistenten	<u>5</u>	8.71	<input type="checkbox"/>	problematika	<u>7</u>	8.47
<input type="checkbox"/>	negativen	<u>8</u>	8.46	<input type="checkbox"/>	oblikovanje	<u>13</u>	7.97
<input type="checkbox"/>	turističen	<u>4</u>	7.81	<input type="checkbox"/>	vpliv	<u>13</u>	7.63
<input type="checkbox"/>	nacionalen	<u>4</u>	7.23	<input type="checkbox"/>	koncept	<u>8</u>	7.4
<input type="checkbox"/>	različen	<u>9</u>	6.29	<input type="checkbox"/>	učinek	<u>4</u>	6.1
<input type="checkbox"/>	dober	<u>6</u>	6.22	<input type="checkbox"/>	področje	<u>5</u>	4.86
<input type="checkbox"/>	javen	<u>4</u>	5.36	<input type="checkbox"/>	vloga	<u>4</u>	4.79

Figure 1: Partial word sketch for *imidž* in the KoRP corpus (the Sketch Engine).

It is worth emphasizing that we initially employed the same settings as our colleagues for compiling single-word noun and verb entries in the lexical database; however, the automatic extraction of lexical information for multi-word units (through MWU links in the Sketch Engine) was first tested in the TERMIS project. With the exception of values for minimum collocation salience for nouns and values for minimum gramrel salience for verbs, which remained unchanged, we had to reduce the minimum values for all other parameters of grammatical relations. This was expected, given the fact that the KoRP corpus (1.8 million words) is much smaller than the Gigafida corpus (1.2 billion words), used in extracting the information for the lexical database.

6. GDEX

Part of the method for extracting lexical information involves the GDEX tool. GDEX ranks corpus examples according to their dictionary potential by using criteria such as sentence length, whole-sentence form, sentence complexity, presence/absence of rare words, presence of URLs etc., and is therefore a very useful function for lexicographers (Kilgarriff et al., 2008; Kosem et al., 2011; Kosem, Gantar & Krek, 2012). It has been envisaged from the very beginning that the dictionary of public relations will include collocations as well as examples, so we yet again utilized the knowledge gained during the compilation of the Slovene Lexical Database (Kosem et al., 2011; Kosem, Gantar & Krek, 2012).

	min. coll. freq.	min. coll. sal.	min. gramrel freq.	min. gramrel sal.	gramrel type
O_količina	2	0.5	6	10.0	O
O_nedoločnik_cs	2	0.5	8	0.2	O
O_povratni_se	2	0.5	8	0.2	O
O_povratni_si	2	0.5	8	0.2	O
O_s_števili	2	0.5	6	1.0	O
O_tretja_oseba	2	0.5	8	0.2	O
O_z_lastnim_imenom	2	0.5	6	1.0	O
O_zanikanje	2	0.5	6	10.0	O
S._*_p2	2	0.5	6	10.0	S
S._*_p3	2	0.5	6	10.0	S
S._*_p4	2	0.5	6	10.0	S
S._*_p5	2	0.5	6	10.0	S
S._*_p6	2	0.5	6	10.0	S
S._*_r	2	0.5	8	0.2	S
S._*_r2	2	0.5	8	0.2	S
S._*_r3	2	0.5	8	0.2	S
S._*_r4	2	0.5	8	0.2	S
S._*_r5	2	0.5	8	0.2	S
S._*_r6	2	0.5	8	0.2	S
S._*_s2	2	-20.0	6	2.0	S
S._*_s3	2	-20.0	6	2.0	S
S._*_s4	2	-20.0	6	0.2	S
S._*_s5	2	-20.0	8	0.5	S
S._*_s6	2	-20.0	6	1.0	S
S._*_x_g2	2	-20.0	6	0.5	S

Table 2: Part of settings for grammatical relations for nouns, group 2.

We prepared five different GDEX configurations for nouns and two for verbs; the configurations differed in values of certain parameters (e.g. optimum example length: 15–40 words/15–35 words/15–30 words). After several evaluations we selected two final configurations: one for nouns (single-word and multi-word) and one for verbs. The difference in ranking of examples by different configurations was especially noticeable for more frequent nouns, i.e. nouns with frequency over 600, while the comparison of rankings for single-word nouns, multi-word nouns, and verbs with frequency under 250 displayed little or no difference; however, this is to be expected due to a smaller number of examples for each collocate of low frequency words. Nonetheless, even in the cases of collocates with fewer examples, GDEX saved valuable time by ranking, and thus selecting for automatic export, the two best examples.

If compared with the GDEX configuration used for the Slovene Lexical Database, only three changes have been required for the terminology extraction. The changes to frequency settings were expected. Table 3 shows the differences in settings for nouns.

Classifier	Slovene Lexical Database	TERMIS
penalty for examples containing tokens with frequency of less than 3	yes	no
lemma frequency	yes, frequency = 1000	no
additional classifier for second-level collocations	yes, weight 10	yes, weight 10 (min. frequency of a collocate: 2)

Table 3: Part of GDEX configuration settings for Lexical Database for Slovene and TERMIS (single-word and multi-word nouns).

In addition to the three changes in settings used for noun terms, another change was made in the extraction of information for verb terms; namely, we added a classifier for optimum position of the keyword (i.e. term), so that the examples containing the keyword in the last two thirds of the sentence were ranked higher.

After all the configurations had been prepared, we ran the API script and extracted the information in XML format, and after minor conversions (e.g. gramrel names) imported the data into the editing tool of the Termania terminology portal.

Using word sketches, sketch grammar and GDEX, we extracted collocates and good examples (two examples per collocate). Each collocate was automatically listed under the relevant grammatical relation. Using this approach, we avoided manual corpus analysis for nearly 2000 terms, including the consultation of word sketches. Manual corpus analysis was used for only 150 multi-word terms that did not contain the combination *adjective + noun* or *noun + noun*, i.e. multi-word terms where automatic extraction of lexical information was not possible.

7. Editing the data

“No matter how many features are used to summarize the data, the lexicographer still needs to critically review the summary” (Kilgarriff & Kosem, 2012: 48). One of the differences between the compilation of a terminological dictionary and the compilation of a general dictionary is that there is much less polysemy in a terminological dictionary. Consequently, the work with the dictionary editor on the Termania portal (Figure 2) mainly comprised the redistribution and grouping of semantically related collocates, identification of compounds, and moving and reordering of corpus examples. In rare cases, we were required to re-examine the word sketch of the term, and in 10% of collocates the automatically extracted examples were too similar, so we analyzed the concordances and manually selected another example. This was the case for rare words and phrases where GDEX is of little use; in fact, in many cases the manual analysis revealed that there are no

alternative different examples in the corpus, as the authors of corpus texts cited the same source in the same or a very similar manner. Even in cases when all the word sketch information was extracted (e.g. for 479 nouns with corpus frequency of less than 20), the automatic extraction reduced the time required for editing; deleting irrelevant information was quicker than the alternative, i.e. searching for relevant information in concordances and manually exporting each example.

8. Visualization of data for online dictionary

The visualization of the terminological dictionary of public relations is currently in its final stage. Online availability of the dictionary database was included in the original project proposal.

There are some important characteristics of online dictionaries or databases, including a customizable interface, filters, hyperlinks, video content, etc. (e.g. see Corr ard, 2002; Schryver, 2003: 152–160; Heid & Gouws, 2006: 981; Caruso 2011). Simultaneously, we are aware of the rather unexpected findings of M ller-Spitzer, Kopenig & T pel (2011), that multimedia content and other functionalities of online dictionaries are regarded as rather unimportant by users, especially if compared with the importance of reliability, clarity, and the up-to-date nature of information (see also Kopenig, 2011). Thus, we focussed on one particular aspect of visualization: how to present data to the user in a clear and understandable manner, considering that for a terminological dictionary there is likely to be a large amount of data for a single entry.



Figure 2: Dictionary editor of the Termania portal.

As shown in Figures 3–5, each entry (at the moment) consists of **three levels of display**:

I. The home page of Termania contains a search window that enables searches in all dictionaries included in the portal. At this level, the search results from the dictionary of public relations include the following information: headword, beginning of a (short) definition, and translation of headword into English (Figure 3).

II. By clicking on the headword, we open the second level where additional information is displayed: frequency in the KoRP corpus in the form of diamonds,⁴ grammar information, entire short definitions, two corpus examples, collocates grouped by grammatical relation (Figure 4), related entries in the same dictionary (cross references), and, in the last part of the entry, links to concordances in Gigafida and KoRP.

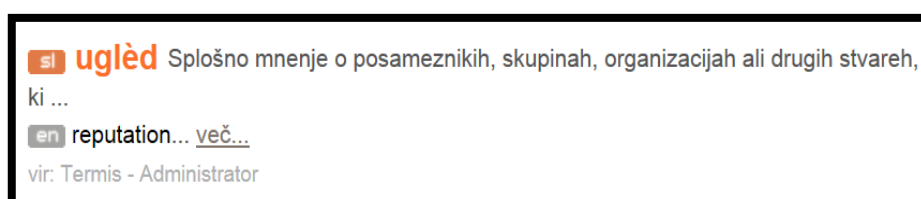


Figure 3: Termania portal: first level of entry display.

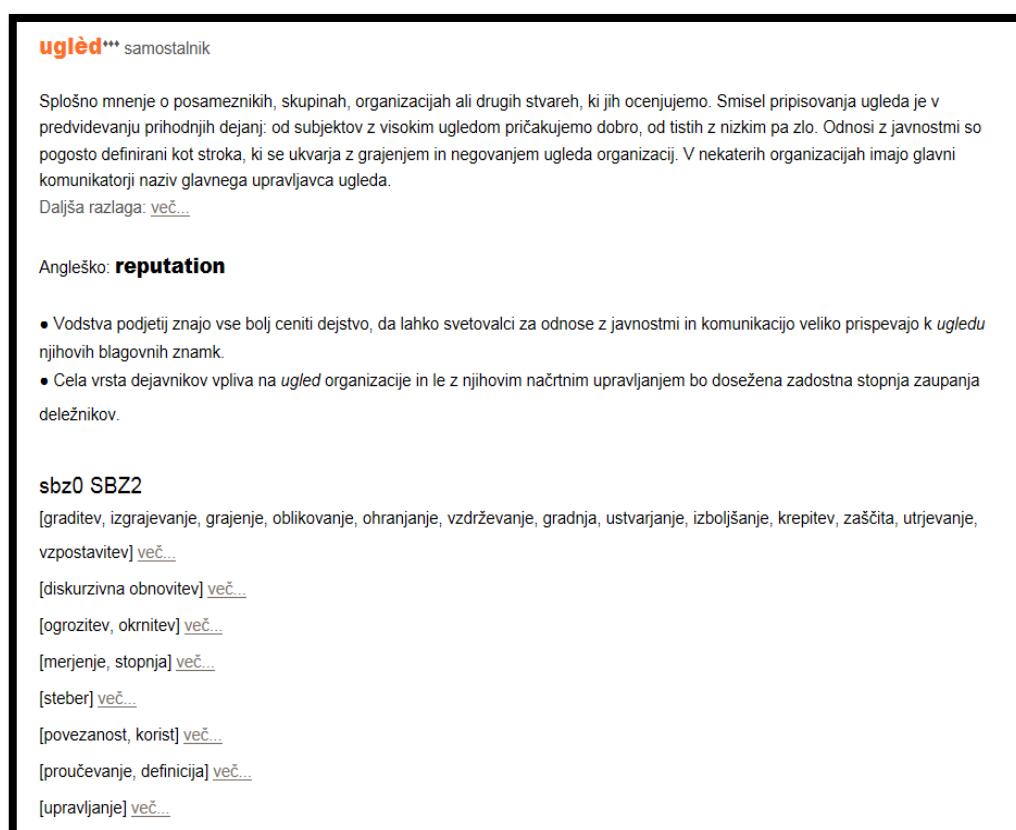


Figure 4: Termania portal: second level of entry display (partial screenshot).

⁴ Number of diamonds and related values: one diamond = frequency between 1–99, two diamonds = frequency of 100–699, three diamonds = frequency 700 and above.

III. The user accesses the third level by clicking on *več...* (*more...*), available in two places:

- a) at a short definition, where a click on *več...* reveals a longer, encyclopaedic definition (Figure 5, above),
- b) and at each group of collocates, where a click on *več...* reveals corpus examples, two per collocate (Figure 5, below).

At the time of writing the paper, we are conducting a survey among public relations experts and translators. The survey will provide information on understandability, clarity, accuracy, readability, amount, usefulness and relevance of the database contents and its structure. The survey findings will be implemented in the design of the dictionary.

uglèd*** samostalnik

Splošno mnenje o posameznikih, skupinah, organizacijah ali drugih stvareh, ki jih ocenjujemo. Smisel pripisovanja ugleda je v predvidevanju prihodnjih dejanj: od subjektov z visokim ugledom pričakujemo dobro, od tistih z nizkim pa zlo. Odnosi z javnostmi so pogosto definirani kot stroka, ki se ukvarja z grajenjem in negovanjem ugleda organizacij. V nekaterih organizacijah imajo glavni komunikatorji naziv glavnega upravljavca ugleda.

Daljša razlaga: [mani...](#)

V poslovnem svetu ima ugled ekonomsko vrednost, saj ugledni subjekti lažje in ceneje prihajajo do virov, ki jih potrebujejo pri svojem delu: ljudje se raje odločajo delati za bolj ugledne organizacije, banke jim ceneje posojajo denar, lažje prihajajo do oblastnih dovoljenj itd. Gospodarsko je ugled eden od virov konkurenčne prednosti, računovodsko se lahko vrednoti kot neopredmeteno sredstvo, v pravo pa prihaja do izenačevanja varstva pravic fizičnih in pravnih oseb v zvezi z medijskimi napadi na njihov ugled. Zato vse več podjetij in drugih organizacij strateško upravlja s svojim ugledom, in sicer z raziskovanjem in upoštevanjem širših družbenih pričakovanj, upravljanju čim boljših odnosov s svojimi deležniki in s komuniciranjem. Zato so pomembne sestavine upravljanja z ugledom tudi upravljanje spornih tem, upravljanje s tveganji in krizni menedžment s kriznim komuniciranjem. Pogosta so tudi izračunavanja uglednostnega kapitala, ki je razlika med tržno in knjigovodsko vrednostjo podjetja. Najbolj enostavno si je vrednost uglednostnega kapitala mogoče predstavljati kot razliko med likvidacijsko vrednostjo podjetja (med tistim, kar bi dobili, če bi podjetje prodali po kosih kot zemljišča, stavbe, opremo in drugo oprijemljivo in nesporno lastnino) in vrednostjo, ki jo lahko dosežemo, če podjetje prodamo kot celoto. V tem smislu ima lahko ugled tako pozitivne kot negativne vrednosti. Finančno poslovanje in ugled sta sicer vzajemno povezana: visoki zaslužki praviloma prinašajo višji ugled, ta pa povratno prispeva k višjim zaslužkom. V sociologiji je ugled sestavina kulturnega kapitala. Ugled je produkt števila ljudi, ki nas poznajo, in njihovih predstav o nas. Čeprav je ugled javna podoba o nas, lahko v različnih javnostih uživamo različno vrednotenje in imamo torej več različnih ugledov. Pregled se tudi preliva: na vrednosti ugleda podjetja vpliva kakovost izdelčnih in storitvenih znamk, panoga, v kateri posluje, država porekla ali njegovo vodstvo – vse to pa velja tudi povratno.

Angleško: **reputation**

- Vodstva podjetij znajo vse bolj ceniti dejstvo, da lahko svetovalci za odnose z javnostmi in komunikacijo veliko prispevajo k *ugledu* njihovih blagovnih znamk.
- Cela vrsta dejavnikov vpliva na *ugled* organizacije in le z njihovim načrtnim upravljanjem bo dosežena zadostna stopnja zaupanja deležnikov.

sbz0 SBZ2

[graditev, izgrajevanje, grajenje, oblikovanje, ohranjanje, vzdrževanje, gradnja, ustvarjanje, izboljšanje, krepitev, zaščita, utrjevanje, vzpostavitev] [mani...](#)

- o Dobro poznana in ugledna podjetja imajo zato oprijemljive finančne koristi in tako dolgoročno nadomestijo višje stroške, ki so jih imela z vlaganjem v graditev dobrega *ugleda*.
- o Poleg verjetno najpomembnejšega razloga, to je poslovne uspešnosti podjetja, je h graditvi *ugleda* prav gotovo pripomogla komunikacijska dejavnost podjetja.

Figure 5: Termania portal: third level of entry display (partial screenshot).

9. Conclusion

The aim of the TERMIS project was the development of a model for compiling Slovene terminological dictionaries or systematically structured databases in a relatively short amount of time. The method of automatic extraction of term candidates and lexical information (including collocations and examples) from the corpus, used in compiling the terminological database of public relations terms, is in fact language independent; individual parameters of different tools can be adapted to other languages, something that is important for modern lexicography that promotes automating as much of the lexicographic work as possible (Kosem, Gantar & Krek, 2012; Rundell & Kilgarriff, 2011). The use of language technologies and lexicographic tools, as described in this paper, has not only facilitated a quicker building of terminological database for the discipline of public relations, but has also made the analysis more objective.

It appears that user-friendliness and the availability of various multimedia functions, enabled by the online dictionary medium, are yet to be fully developed by dictionary-makers; similarly, dictionary users are still getting accustomed to these functions (Müller-Spitzer, Koplenig & Töpel, 2011). The online format has removed the need for space-saving techniques; but in contrast has raised two questions: how much data is still manageable for the user, and how should dictionary information be effectively organized in this new medium?

/O/ne of the really distinctive features of dictionaries and other lexicographical tools is that they provide quick and easy access to the specific types of data from which a specific type of users can retrieve the information that may cover their specific types of needs in a specific type of extra-lexicographical situation (Tarp 2010: 40).

We are currently conducting a survey among the users of our dictionary that will provide answers to certain questions related to design and structure of dictionary data, but only the feedback of the wider public can evaluate how successful we have been in achieving the aim of our dictionary project.

The TERMIS project has highlighted how language technologies can speed up the building of terminological databases. In addition, language technologies can be used to identify types of information that can be difficult to obtain by manual analysis.

We have developed a model for building a terminological database that could be adopted by other disciplines in Slovenia for the compilation of respective terminological dictionaries. We believe that in times of internationalization of disciplines and research, an effective way to facilitate the development of terminology is by using the approach demonstrated by TERMIS: by making state-of-the-art electronic terminological resources.

10. Acknowledgements

The findings presented in this paper are the result of collaborative work; the authors would like to thank Polona Gantar, Simon Krek and Špela Vintar.

The research was conducted as part of the project entitled *Terminology data banks as the bodies of knowledge: The model for the systematization of terminologies*, funded by the Slovene Research Agency (contract no. 1000-11-274193), Pristop d.o.o., and Chamber of Commerce and Industry of Slovenia. The project has been supported by the following sponsors: Elektro Ljubljana d.d., Mercator d.d., Pošta Slovenije d.o.o., and Zavarovalnica Maribor d.d.

11. References

- Arhar Holdt, Š. (2011). *Luščenje besednih zvez iz besedilnega korpusa z uporabo dvodelnih in tridelnih oblikoskladenjskih vzorcev*. Ljubljana: Trojina, zavod za uporabno slovenistiko.
- Arhar Holdt, Š., Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), pp. 95-110.
- Caruso, V. (2011). Online specialised dictionaries: a critical survey. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 66-75.
- Corréard, M. (2002). Are space-saving strategies relevant in electronic dictionaries. *Proceedings of the 10th EURALEX international congress*, Copenhagen: Center for Sprokhtnologi, pp. 463-470.
- Gantar, P. (2009). Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo*, 54(3/4), pp. 69-94.
- Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) *Natural language processing, multilinguality: 6th international conference*. Modra: Tribun EU, pp. 72-80.
- Grčar, M., Krek, S. & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 89-94.
- Heid, U., Gouws, R. (2006). A model for a multifunctional dictionary of collocations. *Proceedings of the 12th EURALEX international congress*. Torino: Edizioni dell'Orso, pp. 979-988.
- Humar, M., Žagar Karer, M., eds. (2010). *Nacionalni jeziki v visokem šolstvu*. Ljubljana: Založba ZRC, ZRC SAZU.
- Kalin Golob, M. (2012). Jezik slovenskega visokega šolstva: med zakonodajo, strategijo in vizijo. In V. Gorjanc (ed.) *Slovanski jeziki: iz preteklosti v prihodnost*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 95-109.
- Kalin Golob, M., Stabej, M. (2007). Sporazumevanje v znanosti in na univerzi: uboga slovenščina ali uboga jezikovna politika? *Jezik in slovstvo*, 52(5), pp. 87-91.

- Kalin Golob, M., Stabej, M., Stritar, M., & Červ, G. (2012). *Primerjalna študija o učnem jeziku v visokem šolstvu v Republiki Sloveniji in izbranih evropskih državah*. Accessed at:
http://www.mizks.gov.si/fileadmin/mizks.gov.si/pageuploads/Slovenski_jezik/FDV_-_ucni_jeziki_v_visokem_solstvu.pdf.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX international congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425-432.
- Kilgarriff, A., Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger, M. Paquot (eds.) *Electronic lexicography*. Oxford: Oxford University Press, pp. 31-55.
- Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. (2004). The Sketch engine. *Proceedings of the 11th EURALEX international congress*. Lorient: Université de Bretagne-Sud, pp. 105-116.
- Kompara, M., Holozan, P. (2011). What is needed for automatic production of simple and complex dictionary entries in the first Slovene online dictionary of abbreviations using Termania website. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 140-146.
- Koplenig, A. (2011). Understanding how users evaluate innovative features of online dictionaries – an experimental approach. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 147-150.
- Kosem, I., Gantar, P. & Krek, S. (2012). Avtomatsko luščenje leksikalnih podatkov iz korpusa. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 117-122.
- Kosem, I., Husak, M. & McCarthy, D. (2011). GDEX for Slovene. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 150-159.
- Krek, S. (2012). New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3*. Brno. Accessed at: <http://trac.sketchengine.co.uk/wiki/SKEW-3/Program#>.
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Logar, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: doktorska disertacija. Ljubljana: Filozofska fakulteta.
- Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What makes a good online dictionary? – Empirical insights from an interdisciplinary research project. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, pp. 203-208.

- Romih, M., Krek, S. (2012). Termania – prosto dostopni spletni slovarski portal. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 163-166.
- Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.) *A taste for corpora: a tribute to professor Sylviane Granger*. Amsterdam: John Benjamins.
- Schryver, G. de (2003). Lexicographers' dreams in the electronic-dictionary age. *International journal of lexicography*, 16(1), pp. 143-199.
- Tarp, S. (2010). Functions of specialized learners dictionaries. In P. Fuertes-Olivera (ed.) *Specialised dictionaries for learners*. Berlin, New York: De Gruyter, pp. 39-53.
- Vintar, Š. (2008). *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.
- Vintar, Š. (2009). Samodejno luščenje terminologije – izkušnje in perspektive. In N. Ledinek, M. Žagar Karer, M. Humar (eds.) *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, pp. 345-356.
- Vintar, Š. (2010). Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141-158.

The dynamics outside the paper: user contributions to online dictionaries

Andrea Abel¹, Christian M. Meyer²

¹ EURAC, Drususallee 1, Bolzano, Italy

² Ubiquitous Knowledge Processing Lab (UKP-TUDA),
Technische Universität Darmstadt, Germany

E-mail: andrea.abel@eurac.edu, meyer@ukp.informatik.tu-darmstadt.de

Abstract

Online dictionaries rely increasingly on their users and leverage methods for facilitating user contributions at basically any step of the lexicographic process. In this paper, we propose a novel classification of the different types of user contributions, which have not been systematically studied so far. With the help of many practical examples, we discuss three major types of user contributions and discuss multiple forms and implementations of them: (i) Direct user contributions, which comprise dictionary articles written entirely or partly by users in a collaborative effort; (ii) Indirect user contributions, which occur in different forms of explicit feedback (e.g., by e-mail or web forms) and implicit feedback through log file analysis or external user-generated content; (iii) Accessory user contributions, which go beyond the dictionary content by initiating an exchange either between the dictionary makers and their users or among the users themselves. We argue that the ease of communication and collaboration between dictionary makers and users has enormous potential, not only for keeping the dictionary up to date and of high quality, but also for developing improved, user-adapted views of, and access to, the contents of a dictionary. Studying the different types of user contribution is crucial for effectively planning online dictionaries and for future research on electronic lexicography.

Keywords: internet lexicography; online dictionaries; user contributions; collaborative lexicography

1. Motivation

The World Wide Web offers various possibilities for users to contribute to dictionaries. These range from giving feedback or correcting errors to creating new dictionary articles and discussing language-related issues beyond the explicitly encoded knowledge. The ease of communication and collaboration between dictionary makers and users has enormous potential, not only for keeping the dictionary up to date and of high quality, but also for developing improved, user-adapted views of, and access to, the contents of a dictionary.

The discussion on user contributions in lexicography is mainly linked to online dictionaries, but is not new as even print dictionaries may be strongly based on collaboration with the public. The *Oxford English Dictionary*, for example, conducted reading programs right from its inception in the 19th century to collect quotations illustrating how words are used (cf. Thier, 2013).

However, the dynamics outside the paper are obviously different as they facilitate a greater variety of user contributions as well as immediate publication and timely feedback. With the rise of social media technologies (e.g., blogs, wikis, social networks) and the Web 2.0, users can actively participate in the compilation of a dictionary. In fact, we face a new kind of lexicographical process in which the formerly clear distinction between dictionary editors and dictionary users becomes increasingly blurred. This is also captured by the neologism *prosumer*, a blend of *producer* and *consumer* (cf. Lew, 2013). Carr (1997) describes this change of lexicographic paradigms as *bottom-up lexicography* according to which dictionaries are “evolving upward from readers” – as opposed to *top-down lexicography* “from editors, through publishers, to readers”.

For the first time, we systematically study the different types of user contribution backed by multiple practical examples found in existing online dictionaries. Our analysis takes into account both individual dictionaries (e.g., the *Oxford English Dictionary*, *Duden online*) and dictionary portals, such as *LEO*, *dict.cc*, and *canoonet* (cf. Storrer, 2010; Engelberg & Müller-Spitzer, in print). As a result of our work, we propose a classification for describing the dynamics induced by user contributions. At the top level, we distinguish the following three types of user contribution:

- (i) Direct user contributions
- (ii) Indirect user contributions
- (iii) Accessory user contributions

Obviously, a single dictionary project may utilize different types of user contributions at the same time. Therefore, we provide a general, dictionary-independent classification instead of focusing on a specific project. In the paper, we first discuss related work in this area and then describe each of the three types of user contribution in detail. Table 1 shows an overview of our proposed classification. We conclude the paper with a final discussion and a summary of our findings.

Direct user contributions	Indirect user contributions	Accessory user contributions
<ul style="list-style-type: none"> • Contributions to open-collaborative dictionaries • Contributions to collaborative-institutional dictionaries • Contributions to semi-collaborative dictionaries 	<ul style="list-style-type: none"> • Explicit feedback <ul style="list-style-type: none"> – form-based feedback – free form feedback • Implicit feedback <ul style="list-style-type: none"> – log file analysis – external user-generated content 	<ul style="list-style-type: none"> • Exchange between dictionary makers and dictionary users <ul style="list-style-type: none"> – unidirectional communication – bidirectional communication • Exchange among dictionary users

Table 1: Overview of our functional classification of user contributions to online dictionaries

2. Related work

The earliest descriptions of user contributions to electronic dictionaries date back to the mid 1990s. In his well-known article, Carr (1997) introduces the terms *bottom-up lexicography* and *collaborative lexicography*, without further differentiating between them. Although Carr predominantly addresses the submission of dictionary articles or additions by e-mail, both expressions serve nowadays as umbrella terms for different types of user contributions. This also applies to other expressions that are, more or less, synonymously used to describe any type of user contribution, including *user involvement* (Lew, 2011), and contributions based on *user-generated content* (Lew, 2013).

Storrer (1998) distinguishes different types of *user participation* targeted at (i) correcting errors, (ii) identifying gaps, (iii) obtaining expert contributions on certain topics, and (iv) collecting contributions by laypeople in an entertaining and playful setting. In subsequent work, Storrer (2010) focuses on the distinction between dictionaries allowing for user contributions controlled by professional editors and dictionaries created by the users themselves in a collaborative effort.

Køhler Simonsen (2005) describes the evolution from lexicographic products to lexicographic services, which raises an increasing need for involving the users in every stage of the lexicographic process. To this end, he proposes two principles to facilitate user contributions in a specialized dictionary, and he associates each phase of the lexicographic process with the corresponding principles and objectives. By *active user involvement*, he refers to feedback on the design and the development of a dictionary by means of surveys or test groups. On the other hand, *lexicographic democracy* describes feedback on the dictionary articles and the quality of the lexicographic descriptions (e.g., submitting error corrections). The proposed classification is, however, limited to indirect user contributions, as Køhler Simonsen (2005) explicitly excludes the possibility of modifying the dictionary articles directly, as is the case, for example, in collaborative dictionaries. He argues in particular that each user contribution should be subject to editorial control.

Thus, Køhler Simonsen's definition of democracy is not to be confused with the use of *democratization* elsewhere. Fuertes-Olivera (2009), for instance, considers democratization as a result of *collective free multiple-language internet reference works* such as *Wikipedia* and *Wiktionary*, which are entirely compiled by users – without editorial control. He distinguishes them from *institutional internet reference works* that are offered by professional publishers.

A similar distinction is made by Lew (2011), who additionally introduces *collaborative-institutional dictionaries*, which, according to him, lie in between collective-free and institutional dictionaries. This type of dictionary is offered by professional publishers, but allows for direct user contributions.

Lew (2013) discusses multiple dictionary projects along the dimension of their degree of *user-generated content*. This ranges from lexicographic works that entirely consist of user-generated content (*collaborative dictionaries*) to a combination of user-generated content and professional content (comparable to concepts such as *semi-collaborative* [Melchior, 2012], or *user participation* [Storrer, 2010]), and works in which professional content dominates. Lew (2013), in line with Rundell (2012), sees potential in the combination of user-generated and professional content – especially for certain vocabulary types.

Melchior (2012; 2013) introduces the term *semi-collaborative* for his analysis of the *LEO* dictionary portal. He defines a semi-collaborative dictionary as being *supported by users* rather than *generated by users*. Thus, Melchior's use of the term relates to improving and extending existing content, as well as expanding and developing the dictionary project as a whole.

Though it is mostly discussed in the context of the quality of lexicographic products, *simultaneous feedback* (De Schryver & Joffe, 2004; De Schryver & Prinsloo, 2000) represents an important concept when thinking about user contributions, because it initiates a large amount of feedback implicitly and explicitly uttered by users. For printed dictionaries, this means releasing small-scale dictionaries, which are used to collect suggestions for a main dictionary that is being compiled in parallel (De Schryver & Prinsloo, 2001). For the electronic adaptation *fuzzy simultaneous feedback*, De Schryver & Joffe (2004) replace the traditional means of getting feedback (e.g., using questionnaires) with the generation of free implicit feedback, based on log file analysis. From the perspective of user contributions, (fuzzy) simultaneous feedback is similar to the proposal by Køhler Simonsen (2005) introduced above, in the sense that feedback occurs during different phases of the lexicographic process (cf. De Schryver & Prinsloo, 2000).

Recent studies of user contributions have become increasingly detailed. However, a comprehensive and systematic classification is still missing. Rather, there has been a variety of ambiguous and partly overlapping terms, which hampers the effective planning of forms of user contributions for new and established dictionaries. A particular problem is that most previous works are focused on one specific type of user contribution, for example, focusing on the degree of editorial control or discussing different types of feedback.

In his analysis of 88 online dictionaries according to various criteria, Mann (2010) lists three possible types of user contribution. First, *direct contributions* to the dictionary, including the compilation as well as the modification of articles. Second, *indirect contributions*, including the option to give feedback by means of forms, contact addresses, etc., which inherently implies a form of editorial control. Third, the *exchange with other dictionary users* by means of online forums. This classification comprises both collaborative approaches and user contributions based on feedback.

However, Mann (2010) provides little detail of the individual types of user contribution and omits, for instance, the forms of discourse between the dictionary makers and users, which we discuss in section 5. The goal of our contribution is therefore to classify the previously discussed dimensions of user contributions and close the gaps between existing classifications. We use the three types of user contribution proposed by Mann (2010) as a starting point.

3. Direct user contributions

By *direct user contributions* we refer to additions, modifications, and deletions of dictionary articles or parts of them performed by a dictionary user. We can distinguish between direct user contributions to open-collaborative, collaborative-institutional, and semi-collaborative dictionaries.

Contributions to open-collaborative dictionaries are neither constituted nor controlled by a predefined group of experts. Rather, the descriptions in the corresponding dictionaries are completely built by the users themselves. The open-collaborative approach has become particularly popular with the rise of the free online encyclopedia *Wikipedia*, in which users write and edit encyclopedic articles that are immediately published on the Web. Instead of expert knowledge, these user contributions are backed by the collective intelligence of a large number of authors, which has often been described as the “wisdom of crowds” (Surowiecki, 2005). According to Malone et al. (2010), the motivation for contributing to open-collaborative projects can be characterized by *money* (including any type of economic benefit and the training of personal skills), *love* (enjoyment, altruism, socializing with others), and *glory* (receiving recognition from peers).

Most open-collaborative dictionaries are based on fixed lexicographic instructions and a predefined article microstructure. The *Urban Dictionary* is one example of this, as the scope of the dictionary is made clear (i.e., slang, jargon, nonce words, and the like) and contributions are organized in a fixed web form asking for the word, a definition, example usages, and a number of keywords. Many dictionaries of this type focus on translations, for example, *bab.la* or *Glosbe*, as they are easy to model and usually only require fields for the term in the source and the target languages. Multilingual dictionaries particularly benefit from direct user contributions because of the broad diversity of the language pairs of contributing users (cf. Meyer & Gurevych, 2012).

More complex open-collaborative dictionaries that aim at compiling a general language dictionary, such as the *Kamusi project*, require extensive user interfaces to represent all encoded information types. While the majority of these dictionaries provide a dictionary-specific user interface, some of them are based on the wiki technology, such as *Wiktionary* or the *Rap Dictionary*. Wiki-based dictionaries are usually not based on fixed lexicographic instructions and a predefined microstructure.

They rather define a markup language with which the microstructure can be defined individually for each dictionary article (e.g., using bold face for encoding parts of speech). Matuschek et al. (2013) compare user contributions to a dictionary with a fixed microstructure (*OmegaWiki*) and with a loosely defined microstructure (*Wiktionary*). They find that a fixed microstructure limits expressiveness, because complex information types such as verb argument structures or hierarchically-organized word senses are often not modeled and are too complicated to add later on. The structural openness of *Wiktionary*, however, yields inconsistencies in the layout of the articles, and this hampers the fast and efficient use of the dictionary.

Since user contributions to open-collaborative dictionaries are not moderated by professional editors, they are subject to two types of quality-related flaws: (i) spam and vandalism, and (ii) unspecific, incorrect, outdated, oversimplified, or overcomplicated descriptions. In larger projects, there is hence a need for quality assurance measures. *Wiktionary*, for instance, recently introduced the *flagged revisions* feature for some of its language editions. A flagged revision marks a certain version of an article as having accomplished a basic quality standard. Permission to indicate an article as a flagged revision is only granted to active contributors after having edited at least 200 articles. So far, the flagged revisions indicate that an article is at least free of spam (type (i) flaws), but the feature also generally enables a distinction between a *sighted flag* (type (i) flaws) and a *quality flag* (type (ii) flaws).¹

In addition to that, *requests* are another quality assurance measure in *Wiktionary*. If a contributor notices a quality flaw, which (s)he cannot resolve immediately, a colored “request” banner may be added to the article stating a need for verification (e.g., the addition of sources), extension (e.g., the addition of an example sentence), clean up (in terms of content and format), or deletion of an article.

A second type of direct user contribution is *contributions to collaborative-institutional dictionaries* (cf. Lew, 2011). These dictionaries are provided by major dictionary publishers, for example, the *Merriam-Webster Open Dictionary*. The motivation for a company to publish a collaborative-institutional dictionary is to collect evidence and suggestions for improving editorial dictionaries and to keep dictionary users interested in the publisher’s activities and products. Contributions to collaborative-institutional dictionaries may address arbitrary vocabulary as in the *Macmillan Open Dictionary*, or focus on a narrower scope, such as Duden’s *Szenesprachenwiki* for neologisms.

Typically, contributions are in the form of full dictionary articles, which are checked for spam, personal offense or defamation before being published. They are, however, not edited on a large scale, as is the case for semi-collaborative and indirect user

¹ <http://meta.wikimedia.org/w/index.php?oldid=5434621> (27 April 2013)

contributions (see below). Unlike contributions to open-collaborative dictionaries, the users cannot directly modify or delete other user contributions, but are limited to submitting new articles.

In contrast, *contributions to semi-collaborative dictionaries* are carefully examined by professional editors before they are incorporated into the dictionary. One example for this is the *TechDictionary*, which asks for submissions of technology- and computer-related dictionary articles. Naber (2005) found for the semi-collaborative synonym dictionary, *OpenThesaurus*, that only a fraction of the registered users actively contribute to the project. Although user contributions are not limited to additions, he found that most of them merely add new synonyms.

Direct user contributions are also the backbone of the *LEO* project, a collection of eight semi-collaborative bilingual dictionaries. Direct user contributions have been encouraged since the launch of the project in the mid 1990s. Melchior (2013) describes different user contributions to *LEO*, which comprise multiple types of contributions according to our classification system. What we define as contributions to semi-collaborative dictionaries are the submission of new entries, which can be discussed with other users in a forum, as well as the donation of entire word lists and glossaries. After these submissions have been checked by the *LEO* editors for correctness, they are usually directly added to the actual dictionary.

4. Indirect user contributions

Indirect user contributions are suggestions, corrections, supplementary material, comments, external content, and usage data provided by users as feedback to the dictionary makers. The users do not have the possibility to directly modify dictionary articles. We distinguish between *explicit* and *implicit feedback*.

Explicit feedback refers to suggestions, wishes, and error corrections explicitly submitted by the users. Thus, users contribute to the dictionary through their feedback on existing content, by providing supplementary material for single articles (e.g., illustrative usage examples and citations), submitting corrections (e.g., spotting erroneous entries, indicating unclear definitions), or commenting on the dictionary as a whole, for example in terms of the presentation of the dictionary articles. Feedback may also include suggesting new content, e.g. in order to fill lemma gaps.

In this context, we can make a further distinction: dictionaries and dictionary portals allowing for *form-based feedback* by providing templates with predefined fields, and those allowing for *free form feedback*, where any text can be submitted, for instance using e-mail or open text fields. There can also be combinations of both types of explicit feedback.

The *LEO* dictionaries provide, for example, separate web forms for reporting errors, such as typos or imprecise translations.² The web forms in *LEO* are characterized by providing only a few fields, which are, however, obligatory. Users can also comment on the dictionary as a whole. Melchior (2012) discusses the conflicting opinions of different types of users regarding the content of the dictionary. Some users argue, for example, in favor of adding newly-coined terms even though they might be used only for a very short period of time. This conflicts with other users who complain about confusing and overloaded search results. In addition to that, the users may test beta versions of the dictionary and give feedback by e-mail or forum posts on the overall layout, the presentation of specific data, and new features, such as the presentation of inflection tables (cf. Melchior, 2013).

The *Oxford English Dictionary* provides a very detailed web form with mandatory and optional fields, allowing the users to suggest any improvements at any time.³ Aside from this kind of feedback, the editors also react to informal messages in the form of letters or e-mails. The *Oxford English Dictionary* particularly fosters initiatives to get in contact with its users, such as the search for *Science fiction citations* recording the first use of an expression. Although participants can submit their citations in an open format e-mail, they are requested to follow strict rules on what kind of information is required.⁴ In the project *Wordhunt*, the *Oxford English Dictionary* cooperated with the BBC to collect verifiable evidence of the first use of a word.⁵ Thier (2013) gives a detailed overview of these efforts.

These two examples show that there is a smooth transition between direct contributions to semi-collaborative dictionaries and indirect contributions in the form of explicit feedback. While the submission of a new translation to *LEO* (a contribution to a semi-collaborative dictionary) is directly published as part of the dictionary (if the editors agree on it), the citations sent to the *Oxford English Dictionary* (i.e., explicit feedback) represent supplementary material that requires critical verification and selection. The contributions often do not represent a separate dictionary article, but rather a specific piece of evidence that is incorporated into the actual dictionary article. The latter also holds for error corrections that are reported to the dictionary editors.

Rautmann (2013) describes that users of *Duden online* have the possibility to suggest missing lemmas and submit extensions or error corrections by clicking on a button *Wortvorschlag* (i.e., *lemma suggestion*) available at the top of each entry and leading

² http://dict.leo.org/pages/collaboration/ende/reportError_en.html (4 June 2013)

³ <http://www.oup.com/uk/oedsubform/> (4 June 2013)

⁴ http://www.jessesword.com/sf/how_to_cite (4 June 2013)

⁵ <http://public.oed.com/resources/for-students-and-teachers/balderdash-and-piffle> (4 June 2013)

to a web form.⁶ Like the *Oxford English Dictionary*, *Duden online* reacts to e-mails containing propositions and suggestions. The user feedback is considered a valuable resource for the editors to help them improve the dictionary content (Rautmann, 2013).

A different type of explicit feedback is the request for quality assessment. Under the heading “Contribute!”, *dict.cc* asks its users to improve the dictionary by rating a translation as “YES (100% correct)” or “NO / MAYBE”.⁷ The task is described in a set of guidelines and designed similarly to the increasingly popular human intelligence tasks on common crowdsourcing platforms, such as Amazon Mechanical Turk,⁸ which are frequently used for user studies in marketing, social sciences, or artificial intelligence.

The second type of feedback we define is *implicit feedback*, which is provided by users through their usage of the dictionary. This kind of feedback does not require the users to make any effort, and often they do not even realize that they are contributing.

The way a website is used and accessed by a user is often logged in *webserver log files*. Through the use of visualization tools such as *Google Analytics*, dictionary publishers are able to analyze their users and the way their dictionaries are used. *Duden online* identifies, for example, the most frequently accessed articles and lists them in a sidebar. A publisher can also analyze the search terms supplied by the users and spot lemma gaps in the dictionary. Furthermore, this kind of analysis facilitates the analysis of lookup strategies. It turned out that *Duden online* users often entered multiword expressions, such as “im Folgenden” (“hereafter”) or “des Weiteren” (“in addition”), in the search window. Thus, the editors decided to add frequently-searched multiword expressions as separate lemmas rather than treating them as subentries of one of their constituents (cf. Rautmann, 2013).

Apart from specific tools, the analysis of log files is often suggested as a means of revealing a user’s needs and improving the dictionary (cf. De Schryver & Joffe, 2004). In *Elektronisches Lernerwörterbuch Deutsch–Italienisch* (cf. Abel et al., 2003) the analysis of log files has been characterized by a user model recording the actual use of the dictionary individually for each user (e.g., the number of words looked up per visit, the type of lemma and data categories, etc.). Because of this, users have to register by creating a user account and log in before accessing the dictionary. A similar analysis has been done for the *Base lexicale du français* in order to record not only the words and word combinations used as search terms, but the whole lookup behavior of the users (cf. Verlinde & Binon, 2010).

⁶ remark of the authors (4 June 2013): function temporarily disabled

⁷ <http://contribute.dict.cc/?action=wizard> (4 June 2013)

⁸ <https://www.mturk.com> (4 June 2013)

However, the use of log files has also been criticized as yielding limited, superficial conclusions (cf. Möhrs & Müller-Spitzer, 2008; Verlinde & Binon, 2010). A particular problem is the noise introduced by robots and scripts that automatically browse through the dictionary and thus yield imprecise results. Relevant literature in this field lacks methods for properly cleaning the log files.

Many dictionaries or dictionary portals, such as *Merriam-Webster Online* or *Dictionary.com*, allow their users to sign up for a personal account. Once logged in, a user can, for instance, select their favorite articles or organize the dictionary articles in multiple word lists. Although these features are primarily intended for organizing a user's work, the publisher can utilize this information to learn about frequently-used articles or articles that are organized in the same word list and thus might benefit from being cross-referenced. *Wordnik* publishes those word lists and hence makes them part of the dictionary (McKean, 2011).

Finally, the use of *external user-generated content* is another type of implicit feedback. *Wordnik*, for instance, also includes a great deal of user-generated content from external sources, including images uploaded by users from Flickr and short text messages from Twitter. The users of these external services implicitly contribute with their content to the dictionary. An important consideration when using external user-generated content is the method of dealing with inappropriate content. Lew (2013) discusses, for instance, the use of embarrassing images in the *Google Dictionary*. The vast amount of user-generated content usually impedes checking the contents manually. The dictionaries rather rely on disclaimers, collaborative filtering (cf. Terveen & Hill, 2001), or natural language processing systems.

5. Accessory user contributions

Accessory user contributions go beyond the dictionary content by initiating an exchange either between the dictionary makers and their users or among the users themselves.

Many dictionary publishers provide blogs reporting interesting or funny facts about language use and the dictionary. The *Macmillan Dictionary Blog*⁹ features, for example, the regular series “Language tip of the week”, targeted at improving the language proficiency of learners, as well as the “Stories behind Words” series, in which they invite scholars to write about their personal meaning of a certain word. The blog posts usually contain hyperlinks to dictionary articles and thus serve the purpose of promoting the publisher's products and encouraging customers to return.

We consider blogs as a form of *unidirectional communication* for initiating an *exchange between dictionary makers and dictionary users*. Similar measures

⁹ <http://www.macmillandictionaryblog.com> (4 June 2013)

include using newsletters, social networks, or microblogging services to distribute news to the dictionary users. Thier (2013), for example, gives an overview of unidirectional communication in the context of the *Oxford English Dictionary*.

A notable type of offer is online language games. Schoonheim et al. (2012) describe, for instance, the “*Het Verloren Woord*” (The Lost Word) game of the *Algemeen Nederlands Woordenboek*. As part of this game, users receive cryptic descriptions of a ‘lost’ word and are asked to exchange ideas and submit their solution. The game attracted a large number of players and the authors mention that it serves an educational and a dictionary-didactic purpose, in addition to mere publicity.

If the users, in turn, contribute to this form of communication by commenting on or rating the posts, they can contribute to defining interesting topics and hence shape the publisher’s offer. We consider this as bidirectional communication, since it results in a mutual exchange between the dictionary makers and users.

The language blog “*Fragen Sie Dr. Bopp!*” (“Ask Dr. Bopp!”) by *canoonet* evokes another type of bidirectional communication: In keeping with the motto ‘there are no stupid questions; each question will be answered’, a user can submit a language-related question and receives an answer by a language expert. Such offers provide useful insight into the information needs of users and help in improving the dictionary. In addition, the answer to a question usually refers to dictionary articles and hence is another way of promoting the dictionary.

Accessory user contributions are not limited to communication between experts and laypeople. The technologies of the Web 2.0 also yield increasing possibilities for initiating an *exchange among the dictionary users* themselves.

A well-known example of this type of accessory contribution is the forum of the *LEO* online dictionaries. Consider the German compound *Nutzerbindung* (customer retention). At the time of writing, there is no English translation encoded in the *LEO* dictionary. However, there is an entry in the forum, in which a user seeks a translation for this term.¹⁰ The user briefly defines the term in German and proposes the literal translation *user binding* (which is obviously wrong). Answers to the forum post propose the phrases “to build a loyal customer base” and “to get repeat business (or customers)”. This example shows that accessory user contributions are an important addition to the dictionary itself, because the users can react to the specific context of a language-related question.

Other means for initiating this kind of discourse include user comments and discussion pages. *Wordnik*, for example, provides a function for commenting on the dictionary articles; this may be used to ask questions or simply to share one’s own

¹⁰ <http://dict.leo.org/forum/viewUnsolvedquery.php?idThread=88976> (7 August 2013)

opinion on a word. Discussion pages are present in *Wiktionary* allowing users to discuss each dictionary article on a separate page. Unlike the commenting function and the forum posts, user contributions to discussion pages are not bound to a linear order. Instead, utterances can be contributed at any position of the discussion page, which makes it possible to discuss multiple issues at the same time.

Accessory user contributions raise a similar issue regarding the inclusion of user-generated content: inappropriate comments are to be removed. In small projects, this can be achieved by checking each contribution manually. Larger projects make use of automatic systems such as spam filters or rely on manual checking in a collaborative effort. *Wordnik*, for instance, displays a link for reporting comments that contain spam.

6. Conclusion

Drawing on the relevant literature on user contributions to dictionaries and previous approaches to classifying them, we argue that the existing classifications are insufficient to capture the broad variety of user contributions in a comprehensive way. This is why we propose a new classification distinguishing three main types of user contributions and multiple subdivisions:

- (i) Direct user contributions comprise collaborative efforts in open-collaborative, collaborative-institutional, and semi-collaborative dictionaries. This type of user contribution is targeted towards insertions, modifications, and deletions that directly affect the dictionary articles.
- (ii) Indirect user contributions are subdivided into explicit feedback based on e-mail or web forms and implicit feedback through log file analysis or external user-generated content. Thereby, the users have only indirect means of changing a dictionary article.
- (iii) Accessory user contributions go beyond the dictionary content as they include communication either between the dictionary makers and their users in a unidirectional or bidirectional way or among the users themselves.

We described each type of user contribution with the aid of multiple practical examples relating both to individual dictionaries and to dictionary portals. We have particularly pointed out that a dictionary is not limited to a single type of user contribution. This becomes evident, for example, in the *LEO* dictionaries, which facilitate user contributions of all three main types that we distinguish.

Our proposed classification of user contributions is crucial for properly planning any online dictionary and for future research on user contributions. In this context, quality is a core aspect, which has not yet been exhaustively addressed, in particular

with regard to defining and evaluating quality (cf. Penta, 2011; Nesi, 2012). This is especially a problem if the dictionary function and target audience is not entirely clear, as is often the case with online dictionaries. This is certainly a desideratum for further research.

7. Acknowledgements

Christian M. Meyer has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

8. References

- Abel, A., Gamper, J., Knapp, J. & Weber, V. (2003). Formative Evaluation of the Web-based Learner's Dictionary ELDIT. In D. Lassner & C. McNaught (eds.) *Proceedings of Ed-Media 2003 World Conference on Educational Multimedia, Hypermedia & Telecommunications, June 23-28, 2003, Honolulu, Hawaii, USA*. Norfolk (USA), pp. 1210–1217.
- Algemeen Nederlands Woordenboek*. Accessed at: <http://anw.inl.nl>
- Bab.la*. Accessed at: <http://bab.la>
- Base lexicale du français (BLF)*. Accessed at: <http://ilt.kuleuven.be/blf>
- canoonet*. Accessed at: <http://www.canoo.net>
- Carr, M. (1997). Internet Dictionaries and Lexicography. *International Journal of Lexicography*, 10(3), pp. 209–230.
- De Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 187–196.
- De Schryver, G.-M. & Prinsloo, D.J. (2001). Fuzzy SF: Towards the ultimate customised dictionary. *Studies in Lexicography*, 11(1), pp. 97–111.
- De Schryver, G.-M. & Prinsloo, D.J. (2000). Dictionary-Making Process with ‘Simultaneous Feedback’ from Target Users to the Compilers. In U. Heid, St. Evert, E., Lehmann & Ch. Rohrer (eds.) *Proceedings of the Ninth Euralex International Congress. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, pp. 807–818.
- dict.cc*. Accessed at: <http://www.dict.cc>
- Dictionary.com*. Accessed at: <http://www.dictionary.com>
- Duden online (DO)*. Accessed at: <http://www.duden.de>
- Elektronisches Lernerwörterbuch Deutsch–Italienisch (ELDIT)*. Accessed at: <http://www.eurac.edu/eldit>

- Engelberg, St. & Müller-Spitzer, C. (in print). Dictionary Portals. In R.H. Gouws, U. Heid, W. Schweickhard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography*. Berlin/New York: de Gruyter.
- Fuertes-Olivera, P.A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective Free Multiple-Language Internet Lexicography. In H. Bergenholtz, S. Nielsen & S. Tarp (eds.) *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp. 99–134.
- Glosbe*. Accessed at: <http://glosbe.com>
- Google Dictionary*. Accessed at: <http://www.google.com/dictionary> [offline since 2011]
- The Kamusi Project*. Accessed at: <http://kamusi.org>
- Köhler Simonsen, H. (2005). User Involvement in Corporate LSP Intranet Lexicography. In H. Gottlieb, J.E. Mogensen & A. Zettersten (eds.) *Symposium on Lexicography XI. Proceedings of the Eleventh International Symposium on Lexicography May 2-4, 2002, at the University of Copenhagen*. Tübingen: Niemeyer, pp. 489–510.
- LEO*. Accessed at: <http://dict.leo.org>
- 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. (2013). User-generated content (UGC) in English online dictionaries. In A. Abel & A. Klosa (eds.) *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess* (OPAL – Online publizierte Arbeiten zur Linguistik). Mannheim: Institut für Deutsche Sprache, pp. 9–30.
- Macmillan Open Dictionary*. Accessed at: <http://www.macmillandictionary.com/open-dictionary>
- Malone, T.W., Laubacher, R. & Dellarocas, C. (2010). *Harnessing Crowds: Mapping the Genome of Collective Intelligence*, MIT Sloan School Working Paper 4732–09. Accessed at: <http://ssrn.com/abstract=1381502>.
- Mann, M. (2010). Internet-Wörterbücher am Ende der „Nullerjahre“: Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien. In R.H. Gouws, U. Heid, St.J. Schierholz, W. Schweickard & H.E. Wiegand (eds.) *Lexicographica 26*. Berlin/New York: de Gruyter, pp. 19–46.
- Matuschek, M., Meyer, C.M. & Gurevych, I. (2013). Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications, *Translation: Computation, Corpora, Cognition – Special Issue on ‘Language Technology*

for a Multilingual Europe, 3(1), pp. 87–118.

McKean, E. (2011). Wordnik: Notes from an online dictionary project. In I. Kosem & K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users (eLex2011)*. Bled, Slovenia.

Melchior, L. (2012). Halbkollaborativität und Inline-Lexikographie. Ansätze und Überlegungen zu Wörterbuchredaktion und Wörterbuchforschung am Beispiel LEO Deutsch-Italienisch. In R.H. Gouws, U. Heid, St.J. Schierholz, W. Schweickard & H.E. Wiegand (eds.) *Lexicographica 28*. Berlin/New York: de Gruyter, pp. 337–372.

Melchior, L. (2013). Ansätze zu einer halbkollaborativen Lexikographie. In A. Abel & A. Klosa (eds.) *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess (OPAL – Online publizierte Arbeiten zur Linguistik)*. Mannheim: Institut für Deutsche Sprache, pp. 31–52.

Merriam-Webster Online. Accessed at: <http://www.merriam-webster.com>

Merriam-Webster Open Dictionary. Accessed at:
<http://nws.merriam-webster.com/opedictionary>

Meyer, Ch.M. & Gurevych, I. (2012). Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In: S. Granger & M. Paquot (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, pp. 259–291.

Möhrs, Ch. & Müller-Spitzer, C. (2008). First ideas of user-adapted views of lexicographic data exemplified on OWID and elexiko. In M. Zock & C-R. Huang (eds.) *Coling 2008: Proceedings of the workshop on Cognitive Aspects on the Lexicon (COGALEX 2008)*. Manchester, August 2008, pp. 39-46.

Naber, D. (2005). OpenThesaurus: ein offenes deutsches Wortnetz. In: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung*. Frankfurt: Peter Lang, pp. 422–433.

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

OmegaWiki. Accessed at: <http://www.omegawiki.org>

OpenThesaurus. Accessed at: <http://www.openthesaurus.de>

Oxford English Dictionary (OED). Accessed at: <http://www.oed.com>

Penta, D. J. (2011). *The Wiki-fication of the dictionary: Definitng lexicography in the digital age*. Paper presented at 'Unstable platforms: the promise and peril of transition', 7th Media in Transition Conference, Massachusetts Institute of Technology, Cambridge, MA, 13-15 May, 2011.

The Rap Dictionary. Accessed at: <http://www.rapdict.org>

Rautmann, K. (2013). Duden online und seine Nutzer. In A. Abel & A. Klosa (eds.) *Ihr*

- Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess* (OPAL – Online publizierte Arbeiten zur Linguistik). Mannheim: Institut für Deutsche Sprache, pp. 53–66.
- Rundell, M. (2012). ‘It works in practice but will it work in theory?’ The uneasy relationship between lexicography and matters theoretical. In R.V. Fjeld & J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress, Oslo*: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 47–92.
- Schoonheim, T., Tiberius, C., Niestadt, J. & Tempelaars, R. (2012). Dictionary Use and Language Games: Getting to Know the Dictionary as Part of the Game. In R.V. Fjeld & J.M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress, Oslo*: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 974–979.
- Storrer, A. (1998). Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In H. E. Wiegand (ed.) *Wörterbücher in der Diskussion III*. Tübingen: Niemeyer, pp. 107–135.
- Storrer, A. (2010). Deutsche Internet-Wörterbücher: Ein Überblick. In R. H. Gouws, U. Heid, St. J. Schierholz, W. Schweickard & H. E. Wiegand (eds.) *Lexicographica 27*, Berlin/New York: de Gruyter, pp. 155–164.
- Surowiecki, J. (2005). *The Wisdom of Crowds*, New York: Anchor Books.
- Szenesprachenwiki*. Accessed at: <http://szenesprachenwiki.de>
- TechDictionary*. Accessed at: <http://www.techdictionary.com>
- Terveen, L. & Hill, W. (2001). Beyond Recommender Systems: Helping People Help Each Other. In Carroll, J.M. (ed.) *Human-Computer Interaction in the New Millennium*, Boston: Addison-Wesley.
- Thier, K. (2013). Das Oxford English Dictionary und seine Nutzer. In A. Abel & A. Klosa (eds.) *Ihr Beitrag bitte! – Der Nutzerbeitrag im Wörterbuchprozess* (OPAL – Online publizierte Arbeiten zur Linguistik). Mannheim: Institut für Deutsche Sprache, pp. 67–74.
- Urban Dictionary*. Accessed at: <http://www.urbandictionary.com>
- Verlinde, S. & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress*. Ljouwert: Fryske Akademy, pp. 1144–1151.
- Wikipedia*. Accessed at: <http://www.wikipedia.org>
- Wiktionary*. Accessed at: <http://www.wiktionary.org>
- Wordnik*. Accessed at: <http://www.wordnik.com>

A Jellyfish Dictionary for Arabic

Mohammed Attia, Josef van Genabith

School of Computing, Dublin City University, Ireland

{mattia,josef}@computing.dcu.ie

Abstract

In a festschrift to Martin Gellerstam (Gottlieb and Mogensen, 2007), an article was published by John Sinclair in which he introduced the concept of a *jellyfish dictionary*. It presented the idea of a self-updating dictionary that is able to automatically monitor language change. “It would, so to speak, float on top of a corpus, rather like a jelly-fish, its tendrils constantly sensing the state of the language.” We think that an electronic *jellyfish dictionary* should be able to perform three major tasks. It should be able to tell which words have newly appeared in a language, which words are not in use anymore, and which word usages or senses have changed based on contemporary data. In this paper we explain our methodology for realizing a jellyfish dictionary for Arabic by automatically performing the three tasks: detecting new words, flagging obsolete words, and discovering word senses.

Keywords: Arabic; automatic lexical acquisition, detection of new words, obsolete word detection, word senses

1. Introduction

A corpus is the foundation for any lexicographic work, as both a source of lexical knowledge and evidence underpinning theoretical assumptions related to dictionary entries. However, most of the lexicographic work to date has concentrated on the evidence part of the corpus, rather than the knowledge part. Today’s dictionaries are inspired and supported by corpora, rather than shaped by them. This is where the need for a *jellyfish dictionary* emerges. The idea of a *jellyfish dictionary* was first introduced in an article published by John Sinclair (Gottlieb and Mogensen, 2007) in which he put forward the concept of a self-updating dictionary that is able to automatically monitor language change. “It would, so to speak, float on top of a corpus, rather like a jelly-fish, its tendrils constantly sensing the state of the language.”

With today's corpus sizes exceeding 10^9 words, it becomes impossible to manually check corpora for new words to be included in a lexicon. The idea of a jellyfish dictionary is to develop intelligent tools to allow the corpus to manage the dictionary from top to bottom. The tendrils of the jellyfish sense changes in the sea of words in the corpus and inform us about new developments.

We uphold that an electronic *jellyfish dictionary* needs to perform three major tasks: detecting new words appearing in a language, flagging obsolete words, and observing word senses by identifying the contexts in which words usually prefer to appear. In

this paper, we present our methodology for performing these three tasks. First, we automatically detect new words in Arabic, lemmatize new words in order to relate multiple surface forms to their base underlying representations, decide on words' part of speech (POS), collect statistics on the frequency of use, and model human decisions on whether to include the new words in a lexicon or not. Second, we signal obsolete words in a traditional dictionary based on statistics from a large corpus and a number of web search sites. Third, we investigate word senses based on their preferred contexts, concentrating on the extraction of subcategorization frames and word trigrams.

In our work we use a large-scale corpus of 1,089,111,204 words, consisting of the Arabic Gigaword Fourth Edition (Parker et al., 2009) with 925,461,707 words, in addition to 163,649,497 words from news articles crawled from the Al-Jazeera web site¹. In this corpus, new words appear at a rate of between 2% of word tokens (when we ignore possible spelling variants) and 9% of word tokens (when possible spelling variants are included). For the purposes of this study, new words are words not recognized by the SAMA morphological analyzer (Maamouri et al., 2010), and spelling variants refer to alternative (sub-standard) spellings recognized by SAMA which are mostly related to the possible overlap between orthographically similar letters, such as the various shapes of *hamzahs* (أ إ إى), *taa' marboutah* and *haa'* (ة ة), and *yaa'* and *alif maqsoura* (ي ي).

Our techniques and methods in dealing with the extraction and lemmatization of new words are evaluated on a held-out manually-annotated gold standard of 2,103 form types (unique words), improving on previous work by Attia et al. (2012).

This paper is structured as follows. Section 2 presents the methodology we follow in extracting and analysing new words. Section 3 explains how obsolete words are automatically detected. Section 4 provides details on how word senses can be ranked according to their frequency in the corpus in certain contexts (subcategorization frames and trigrams), and Section 5 concludes the paper.

2. Detecting New Words

New words are constantly finding their way into any living human language. These new words are either coined or borrowed and reflect changes in our societies and lives. Words such as *تويتر* *twiytar* 'twitter', *محاصصة* *muHASaSap* 'allotting shares', *عسكرة* *Easokara* 'to militarize', and *سَيِّسَ* *say~asa* 'to politicize' are not included in current Arabic dictionaries. The inclusion of new words in a lexicon needs to address three important problems. First, the detection, or the method by which we know that a new word has appeared. Second, lemmatization, or relating multiple surface forms to their canonical representation. Third, reaching a decision on the new word; that is,

¹ <http://aljazeera.net/portal>

how we judge whether the new word should be added to the lexicon or not. We address this issue by developing an automatic technique to recognize unknown words in a large corpus of 10^9 words, and reduce them to their lemmas, predict their POS, and rank them in their order of lexicographic importance.

In previous proof-of-concept research, Attia et al. (2012), thereafter referred to as Attia2012, detect a total of 2,116,180 new types. They filter this list using a frequency threshold and a spell checker, creating a subset of 40,277 new types. After lemmatization, the list is reduced to 18,000 possible unique new lemmas. The drawback with filtering in the pre-processing stage through spell checking is that it could be throwing the baby out with the bath water. There is no guarantee that all word forms not accepted by the spell checker used are actually spelling mistakes (or even that all the ones accepted are correct).

In the research presented here we show that filtering in the pre-processing stage actually leads to discarding potentially useful information too early. In our new gold standard of 2,103 types, 1,074 were incorrectly tagged as misspelt by the automatic spell checker, resulting in only 48.93% accuracy for unknown words. Furthermore, of the terms incorrectly tagged as misspellings, 20.58% were nominated to be included in a dictionary (9.59% when excluding proper nouns).

Similar problems arise with the idea of excluding types based on their frequency. Word forms with low frequency may interact with other word forms to support a certain lemma, and throwing them out too early risks losing potentially important information. For example, in our data the word form *واديدينامياتنا* wadiynamiy~AtinA ‘and-our-dynamics’ has a frequency of one, but it interacts with 31 other sister forms (such as *والديديناميات* ‘and-dynamics’, *ديدينامياتهم* ‘their-dynamics’) with an accumulated frequency of 3,464, to support the lemma *ديدينامية* diynamiy~ap ‘dynamic’. In our new gold standard test set of 2,103 types, a subset of 701 types is selected from the frequency range of 10 repetitions or less. When analyzed, we found that 306 types of them were valid (43.65%). Of the valid types, 94 (30.72%) participated with other forms to support a certain lemma and all of them were nominated for inclusion in a dictionary.

In the current research we apply our technique to the full list of 2,116,180 unknown types from Attia2012. We test our method against a manually created gold standard of 2,103 types and show a significant improvement over the baseline and Attia2012. Furthermore, we investigate different criteria for weighting and prioritizing new words for inclusion in a lexicon depending on four factors: number of form variations of the lemmas, cumulative frequency of the forms, type of POS tag, and spelling correctness (according to a spell checker).

2.1 Lemmatization

In order to deal with new words we need to address the issue of lemmatization.

Lemmatization reduces surface forms to their canonical base representations (or dictionary look-up form), i.e., words before undergoing any inflection, which, in Arabic, means verbs in their perfective, indicative, 3rd person, masculine, singular forms, such as شَكَرَ \$akara ‘to thank’; and nominals (the term used for both nouns and adjectives) in their nominative, singular, masculine forms, such as طالب TALib ‘student’; and nominative plural for *pluralia tantum* nouns (or nouns that appear only in the plural form and are not derived from a singular form), such as ناس nAs ‘people’.

The problem with lemmatizing unknown words is that they cannot be matched against a morphological lexicon. Furthermore, the specific problem with lemmatizing Arabic words is the richness and complexity of Arabic morphological derivational and inflectional processes.

Lemmatization of unknown words has been addressed for Slovene in Erjavec and Džerosk (2004), for Hebrew in Adler et al. (2008), for Spanish in Grefenstette et al. (2002), and for English, Finnish, Swedish and Swahili in Lindén (2008). Lemmatization of Arabic has been addressed in Roth et al. (2008) and Dichy (2001). Mohamed and Kübler (2010) handle Arabic unknown words and provide results for known and unknown words in both word segmentation (stemming) and part of speech tagging. They reach a stemming accuracy of 81.39% on unknown words and over 99% on known words.

Mohammed and Kübler’s work, however, focuses on stemming rather than lemmatization, which is quite distinct albeit frequently confused. The difference between stemming and lemmatization is that stemming strips off prefixes and suffixes and leaves the bare stem, while lemmatization returns words to their canonical base forms. To illustrate this with an example, consider the Arabic verb form يقولون yaquwluwn ‘they say’. Stemming will remove the present prefix ‘ya’ and the plural suffix ‘uwn’ and leave ‘quwl’ which is a non-word in Arabic. By contrast, full lemmatization will reveal that the word has gone through an alteration process and return the canonical قال qAl ‘to say’ as the base form.

We develop a rule-based finite-state (Beesley and Karttunen, 2003; Hulden, 2009) morphological guesser that can deal with morphological concatenations and alterations and integrate it with a machine learning based disambiguator, MADA (Roth et al., 2008), in a pipeline-based approach to lemmatization.

3. Methodology

To deal with unknown (or out-of-vocabulary) words, we use a pipeline approach which predicts POS tags and morpho-syntactic features before lemmatization. In the first stage of the pipeline, we use MADA (Roth et al., 2008), an SVM-based tool that relies on the word context to assign POS tags and morpho-syntactic features. MADA internally uses the SAMA morphological analyzer (Maamouri et al., 2010), an

updated version of the Buckwalter morphology (Buckwalter, 2004). Second, we use a finite-state morphological guesser that provides all possible interpretations of a given word. The morphological guesser first takes an Arabic surface form as a whole and then strips off all possible affixes and clitics one by one until all possible analyses are exhausted, and it also reverses the effect of morphological alteration rules. The morphological guesser is highly non-deterministic as it outputs a large number of solutions. To counteract this problem, all the solutions are matched against the POS and morpho-syntactic features produced by MADA, and the analysis with the closest resemblance (i.e. the analysis with the largest number of matching morphological features between the FS guesser and MADA) is selected.

For illustration, we present the analysis of the verb form `ويتناهشونها` wa-yatanAha\$uwna-hA ‘and-they-s snatch-it’ by MADA and the different analyses by the finite state guesser sorted according to the number of features that are successfully matched with the MADA analysis of the original surface form.

MADA output for wa-yatanAha\$uwna-hA:

```
form:wytAh$wnhA num:p gen:m per:3 case:na asp:i mod:i vox:a
pos:verbprco:0 prc1:0 prc2:wa_conj prc3:0 enco:3fs_dobj stt:na
```

Finite-state guesser output for wa-yatanAha\$uwna-hA:

```
9      و+conj@+verb+pres+active+3pers+تناهش+Guess
      +masc+pl+nom@ها+objpron+3pers+sg+fem@
7      و+conj@+verb+pres+active+3pers+تناهشو+Guess
      +fem+pl@ها+objpron+3pers+sg+fem@
-2     و+conj@+adj+يتناهشونها+Guess+sg@
-2     و+conj@+noun+يتناهشونها+Guess+sg@
-2     +adj+يتناهشونها+Guess+sg@
-2     +noun+يتناهشونها+Guess+sg@
-3     +adj+يتناهشونه+Guess+dual+nom+compound@
-3     و+conj@+adj+يتناهشونه+Guess+dual+nom
      +compound@
-3     +noun+يتناهشونه+Guess+dual+nom+compound@
```

The matching uses positive scores for matches and negative scores for features found in the finite state output but not present in the MADA output. The top (highest scoring) analysis is selected as the correct lemma of the word.

Figure 1 shows the steps taken to identify, extract and lemmatize unknown Arabic words, which are summarized as follows:

- A corpus of 1,089,111,204 tokens (7,348,173 types) is analyzed with MADA to produce POS tags and morpho-syntactic features.
- The number of types for which MADA could not find an analysis in the Buckwalter morphological analyzer is 2,116,180 (about 29 % of the types). After removing common spelling variants (as detected by MADA), 1,698,852 types remained.

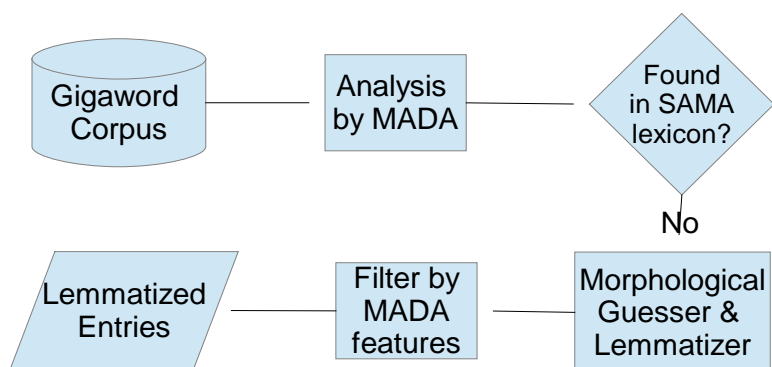


Figure 1: New word extraction and lemmatization process

- Unknown words are analyzed with our finite-state morphological guesser to produce all possible morphological interpretations and relevant possible lemmatizations.
- POS tags and morpho-syntactic features in MADA output are compared with the output of the morphological guesser and the FST guesser analysis with the highest matching score is chosen.

As lemmatization is expected to merge forms having the same lemma together, after lemmatization the list of 1,698,852 types is reduced to 982,886 lemmas, which is too large. We conduct initial filtration by removing word forms that have no supporting morphological variation and which occur only once in the corpus. This basic filtration further reduces the number to 476,349 lemmas.

4. Gold standard Creation

In order to evaluate our methodology we need to create a gold standard from a randomly selected subset of the data. As mentioned earlier, our unknown word list consists of 1,698,852 types. We find that words have varying frequency ranges with a minimum frequency of one, a maximum of 75,885 and a mean of 9.79, as shown in Table 1.

Statistic	Value
Unknown words (after discarding spelling variants)	1,698,852
Minimum frequency	1
Maximum frequency	75,885
Mean	9.79

Table 1: Frequency statistics of the unknown words

When we select a random sample of the data we find that the sample is biased towards low frequency words. Out of 3,000 randomly-selected types, there are 2745

(91.50%) with frequency of 10 or less. This is also true of the entire population where 91.03% of the unknown types have a frequency of 10 or less.

When we investigate the frequency distribution of the unknown words, we see that, as expected, they follow the Zipfian law with a few words having very high frequency and a large number of words having very low frequency (Figure 2).

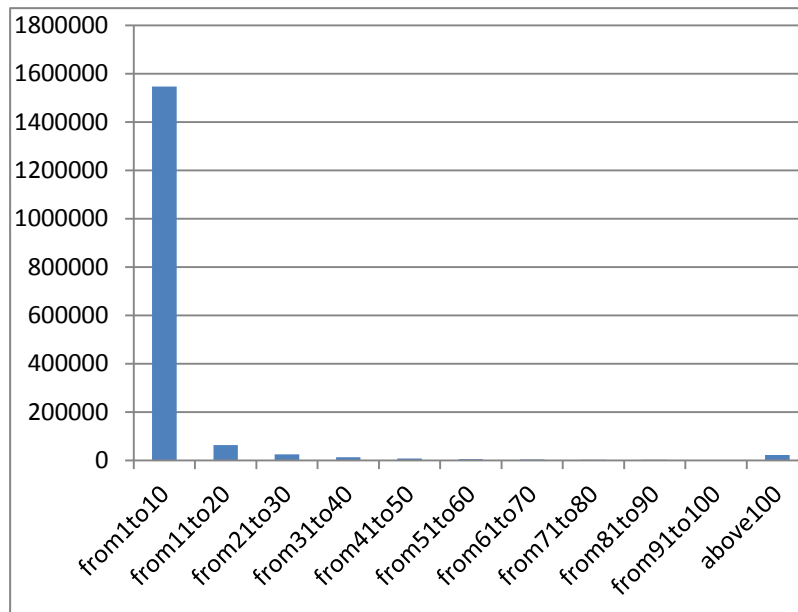


Figure 2: Frequency distribution of the unknown words

In order to avoid the bias towards low frequency words produced by pure randomization, we use a method known in corpus linguistics as ‘stratified sampling’ or what we may call here ‘stratified randomization’. We randomly select 701 words with frequency ≤ 10 , 701 words with frequency >10 and ≤ 50 , and 701 words with frequency >50 , so that our test suite becomes representative of three major frequency ranges, as shown in Figure 3.

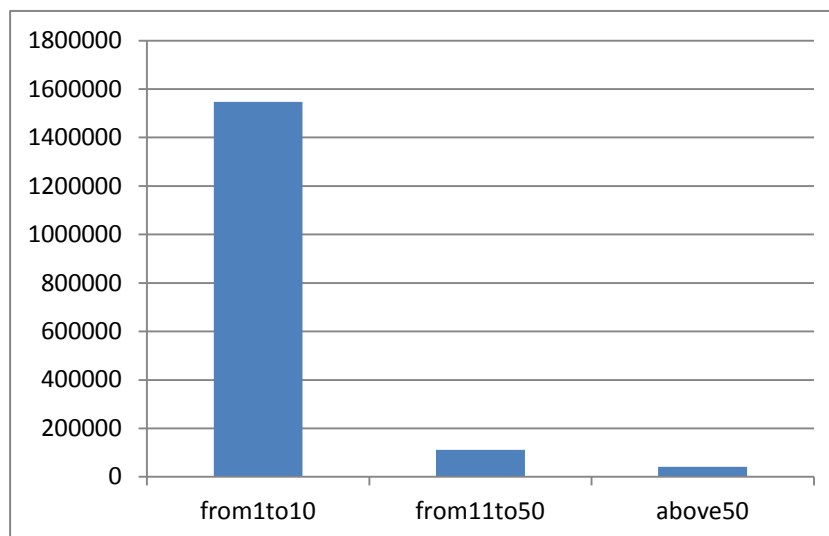


Figure 3: Major frequency ranges of the unknown words

Having created our gold standard of 2,103 unknown types, we ask a human annotator to provide the gold lemma and part of speech for each word form. In addition, the human annotator indicates a preference for whether or not to include the entry in a dictionary; that is, whether a lemmatized form makes a valid dictionary entry or not.

We noticed that the forms marked by the annotator as not fitting for inclusion in a dictionary were mostly misspelled words, colloquial words, and low frequency proper nouns.

Gold Annotation	Jellyfish2013				Attia2012 1,310 types
	Freq ≤10 701 types	Freq >10 and ≤50 701 types	Freq >50 701 types	all 2,103 types	
Valid Forms: of them	43.65%	75.46%	82.31%	67.14%	93.05%
noun_prop	70.92%	77.5%	75.74%	75.35%	48.07%
noun	15.03%	10.4%	10.4%	11.4%	21.16%
adj	11.44%	9.26%	9.88%	9.99%	20.75%
verb	2.29%	1.51%	2.25%	1.98%	4.27%
noun_fem_plural (pluralia tantum)	0.33%	0.38%	0.52%	0.42%	2.3%
noun_broken_plural	0.33%	0.38%	1.04%	0.64%	2.3%
Invalid Forms: of them	56.35%	24.54%	17.69%	32.86%	6.95%
misspelling	60.00%	65.12%	71.77%	63.39%	62.64%
not_resolved	34.68%	19.77%	13.71%	27.21%	16.48%
colloquial	5.06%	15.12%	14.52%	9.26%	20.88%
Lexicographic relevance					
Include in a dictionary	9.84%	13.12%	40.66%	21.21%	51.22%
Include in a dictionary, term not a proper noun (subset of the above)	9.70%	13.12%	16.98%	13.27%	44.35%
Do not include in a dictionary	90.16%	86.88%	59.34%	78.79%	48.78%

Table 2: Gold tag annotation of the test suite

By contrast, nouns, verbs, adjectives, and proper nouns with significantly high frequency were marked for inclusion in the lexical database. This feature of lexicographic preference helps to evaluate our lemma weighting algorithm discussed in the following section.

The POS distribution of the unknown types of our annotated data is shown in Table 2.

Table 2 compares the present gold standard, referred to as Jellyfish2013, to the gold standard presented in Attia et al. (2012), referred to as Attia2012. We observe that proper nouns comprised 48.07% of the valid forms in Attia2012, and 75.35% of the valid forms in Jellyfish2013. We also notice that Attia2012 has fewer invalid forms. Both observations can be explained by the fact that in Attia2012 data passed through filtration by a spell checker which in most cases does not accept infrequent proper nouns. As expected, most unknown words are open class words: proper names, nouns, adjectives, and, to a lesser degree, verbs. It must be noted here that morphological analyzers typically tend to include much more proper nouns than dictionaries. Ordinary dictionaries are usually interested in proper nouns only when they have frequent metonymic use such as *the White House* for ‘the US administration’ and *Westminster* for ‘the UK parliament’.

4.1 Evaluation

We conduct three sets of evaluation experiments to test three aspects of our research on acquiring new words from data: POS tagging, the lemmatization process, and lemma weighting criteria.

4.1.1 POS evaluation

In the first set of experiments we evaluate POS tagging of new words. We measure accuracy calculated as the number of correctly tagged words divided by the number of all valid words. The baseline assigns the most frequent tag (proper name) to all unknown words. In our test data the baseline accuracy stands at 75%. We notice that MADA POS tagging accuracy for unknown words is the same as the baseline, as shown in Table 3. As in Attia2012, we use Voted POS Tagging; that is, we choose the POS tag assigned most frequently by the same tagger (MADA) in the data to a lemma attested more than once. This method has improved the tagging results significantly to 81% which is higher than the baseline. It is also higher than Attia2012, though we use the same method, because of the increased ratio of proper nouns in the gold standard.

		Jellyfish2013 Accuracy	Attia2012 Accuracy
	POS tagging		
1	POS Tagging baseline	75%	45%
2	MADA POS Tagging	75%	60%
3	Voted POS Tagging	81%	69%

Table 3: Evaluation of POS tagging of unknown words

4.1.2 Lemmatization evaluation

In the second set of experiments we test the accuracy of the lemmatization process for new words. The baseline is given by the assumption that new words appear in their base form, i.e., we do not need to lemmatize them. The baseline accuracy is 65%, as

shown in Table 4. We notice that the baseline in Jellyfish2013 is higher than the baseline in Attia2012 partly due to the increased ratio of proper nouns in the new test suite.

Furthermore, lemmatization has improved significantly because of the revised matching mechanism which penalizes extra features in the guesser that have no matches in the MADA output.

	Lemmatization	Jellyfish2013 Accuracy	Attia2012 Accuracy
1	Lemmas found among corpus forms	81%	64%
3	Lemma selection baseline	65%	45%
5	Pipeline-based lemmatization	84%	63%

Table 4: Evaluation of lemmatization of unknown words

4.1.3 Evaluation of lemma weighting

We create a weighting algorithm for ranking and prioritizing unknown words in Arabic so that important words that are valid for inclusion in a lexicon are pushed up the list and less interesting words (from a lexicographic point of view) are pushed down. This is meant to facilitate the effort of manual revision by making sure that the top part of the stack contains the words with highest priority.

In our case, we have 1,698,852 unknown types. After lemmatization and basic filtration, they are reduced to 476,349 (that is a 72% reduction of the surface forms). This number is still too large for manual validation. In order to address this issue we investigate weighting criteria for ranking so that the top n number of words will include the most lexicographically relevant words. We call surface forms that share the same lemma ‘sister forms’, and we call the lemma that they share the ‘mother lemma’. The ‘combined criteria’ refers to the weighting algorithm developed in Attia et al. (2012) which is based on three criteria: number of sister forms, cumulative frequency of the sister forms, and a POS factor. The POS factor gives 50 extra points to verbs, 30 to nouns and adjectives, and nothing to proper nouns. The reason we give higher frequency for verbs is the fact that verb neologisms are usually less common.

$$\text{Word Weight} = ((\text{number of sister forms} * 800) + \text{sum of frequencies of sister forms}) / 2 + \text{POS factor}$$

We use the gold annotated data for the evaluation of the lemma weighting criteria, as shown in Table 5. In our experiments, relying on the sum of frequency of sister forms obtained the best results, giving an optimal balance between increasing the number of lexicographically-relevant words in the top one tenth of the data and reducing the number of lexicographically-relevant words in the bottom tenth.

Lexicographically-relevant words	In top tenth	In bottom tenth
relying on sum of frequency of sister forms	1032	14
relying on number of sister forms (form variation)	716	55
relying on POS factor	89	178
using combined criteria	770	12

Table 5: Evaluation of lemma weighting and ranking

In Attia2012, the combined criteria gave the best results. We notice our data has a bias towards proper nouns; therefore, it could be the case that the combined criteria will be better able to give appropriate importance to other categories, such as nouns, verbs and adjectives. Below, we list some examples of the new lemmas collected in our research.

Proper nouns: waziyrstAn وزيرستان ‘Waziristan’; mAkiyn ماكين ‘McCain’; blAkbiyrn بلاكبيرن ‘Blackburn’; guwroduwn غوردون ‘Gordon’.

Nouns: tasoyiys تسييس ‘politicizing’; AHotirAr احترار ‘warming’; maAliym معالم ‘landmarks’; tay’iys تيبيس ‘putting off’; tawziyr توزيع ‘appointing as a minister’; muhAtarap مهاترة ‘nonsense’; taDomiyyd تضميمد ‘healing’.

Verbs: taEamolaqa تعلق ‘to become gigantic’; taqAfaza تقافز ‘to jump’; xaSoxaSa خصص ‘to privatize’; AnoHa\$ara انحشر ‘to squeeze in’; tanAha\$a تناهش ‘to snatch’; \$aroEana شرعن ‘to legislate’.

Adjectives: \$aEobawiy~ شعبيوي ‘populist’; baHot بحت ‘pure’; muEawolam معولم ‘globalized’; munojaz منجز ‘accomplished’; manZuwr منظور ‘being investigated’; <ixwaniy~ اخواني ‘belonging to the Brotherhood’.

5. Flagging Obsolete Words

After a few decades in the life of any dictionary, it becomes burdened with many oddities related particularly to the preservation of obsolete words and senses. This is specifically the case with Arabic dictionaries which suffer from a lack of appropriate systematic maintenance. More than 1,300 years ago, Al-Khalil bin Ahmed Al-Farahidi compiled the first known monolingual Arabic dictionary called *Al-Ain*. Subsequent Arabic dictionaries typically included refinement, expansion, correction, or organisational improvements over previous dictionaries. These dictionaries include *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980), *al-Muheet* by al-Sahib bin 'Abbad (died 995), *Lisan al-'Arab* by ibn Manzour (died 1311), *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414) and *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791) (Owens, 1997).

Even relatively modern dictionaries such as *Muheet al-Muheet* (1869) by Butrus al-Bustani and *al-Mu'jam al-Waseet* (1960) by the Academy of the Arabic Language in Cairo were not started from scratch, nor was there an attempt to overhaul the process of dictionary compilation or to make any significant change. The aim was mostly to preserve the language, refine older dictionaries, and accommodate accepted modern terminology. In this way, Arabic dictionaries tend to preserve a fossilized version of the language with each new one reflecting the content of the preceding dictionaries (Ghazali and Braham, 2001).

The Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2004) includes 40,648 lemmas (consisting of 420 function words and 1,769 proper nouns, and the remaining 38,459 are nouns, verbs and adjectives). BAMA is widely used by the Arabic NLP research community. It is a *de facto* standard tool, and has been described as the “most respected lexical resource of its kind” (Hajič et al., 2005). The latest version of BAMA is renamed SAMA (Standard Arabic Morphological Analyzer) version 3.1 (Maamouri et al., 2010).

Unfortunately, the SAMA lexical database suffers from a legacy of heavy reliance on older Arabic dictionaries, particularly Wehr's Dictionary (Wehr Cowan, 1976), in the compilation of its lexical database. Attia et al. (2011b) estimate that about 25% of the lexical items included in SAMA are outdated. SAMA includes thousands of obsolete words that are no longer used in speaking or writing. For example, BAMA contains six obsolete words for ‘desert’ (fayfA’ فَيْفَاء, fadofad فَدْفَد, quwA’ قَوَاء, mawomAp مَوْمَاء, matolaf مَتَلَف, and sabosab سَبْسَب) which are no longer in current use.

We need to mention that a full study of the diachronic changes in a language (Lass, 1997) will include currency (words becoming obsolete), register (formal or technical words becoming unmarked), region (regional terms becoming global), syntactic behaviour (e.g. a verb acquiring a new subcategorization frame), and meaning (word meaning is changed or extended). Our focus here is only to handle the first type.

Our objective is to automatically detect and extract obsolete words found in SAMA. To do this, we use a data-driven filtering method that combines open web search engines and our pre-annotated corpus. Using frequency statistics² on lemmas from three web sites using their own search facilities (Al-Jazeera,³ Arabic Wikipedia,⁴ and the Arabic BBC website⁵), we find that 7,095 lemmas in SAMA have zero hits. On the other hand, frequency statistics from our text corpus described in Section 2.2 above show that 3,604 SAMA lemmas are not used in the corpus at all, and 4,471 lemmas occur less than 10 times. Combining frequency statistics from the web and the corpus,

² Statistics were collected in January 2011.

³ <http://aljazeera.net/portal>

⁴ <http://ar.wikipedia.org>

⁵ <http://www.bbc.co.uk/arabic/>

we find that there are 29,627 lemmas that returned at least one hit in the web queries and occurred at least 10 times in the corpus. Using a threshold of 10 occurrences here is discretionary, but the aim is to separate the stable core of the language from instances where the use of a word is perhaps accidental or somewhat idiosyncratic. We consider the refined list as representative of the lexicon of MSA as attested by our statistics.

We consider the remaining 8,832 lemmas (38,459 open-class lemmas, not including proper nouns, minus the 29,627 stable lemmas) as obsolete, and we publish them as an open-source resource⁶ to allow dictionary compilers to flag these words as outdated in their dictionaries.

6. Detecting Word Senses

The SketchEngine (Kilgarriff and Tugwell, 2002) is a tried-and-tested powerful tool for lexicographic work related to word sense discovery, based on context and significant collocates, and using partial parsing and statistical information. In this work we used a similar approach but with different techniques.

In our research we use a fully-parsed resource, the Penn Arabic Treebank (ATB) (Maamouri and Bies, 2004), to extract subcategorization frames for verbs enriched with probability scores. These subcategorization frames help in showing which word senses are more prominent than others for a given verb. We also show how word senses are tied to word forms captured in terms of co-occurrence frequencies (tri-gram frequencies) extracted from the Arabic Gigaword corpus.

6.1 Encoding of subcategorization frames

The encoding of syntactic subcategorization frames is essential in the construction of computational and paper lexicons alike. Subcategorization frames refer to the predicate argument structure. Traditional dictionaries specify whether verbs are transitive (requiring a subject and an object) or intransitive (requiring no object). Subcategorization frames, as defined by the Lexical Functional Grammar (LFG) theory (Dalrymple, 2001), have a broader coverage as they include all governable grammatical functions. The governable grammatical functions are the arguments required by some predicates in order to produce a well-formed syntactic structure, and they include SUBJ(ect), OBJ(ect), OBJ_θ, OBL(ique)_θ, COMP(lement) and XCOMP. The subcategorization requirements in LFG are expressed in the following format (O'Donovan et al., 2005):

$$\pi \langle gf_1, gf_2, \dots, gf_n \rangle$$

⁶ <http://obsoletearabic.sourceforge.net/>

where π is the lemma (predicate or semantic form) and gf is a governable grammatical function. The value of the argument list of the semantic form ensures a well-formed sentence.

For example, in the sentence {iEotamada Al-Tifolu EalaY wAlidati-hi اعتمد الطفل على والدته ‘The child relied on his mother’, the verb {iEotamada ‘to rely’ has the following argument structure: {iEotamada <(\uparrow SUBJ)(\uparrow OBL $_{>alaY}$)>. By including a subject and an oblique with the preposition $>alaY$, we ensure that the verb’s subcategorization requirements are met and that the sentence is well-formed, or syntactically valid.

Attia et al. (2011a) automatically extract the Arabic subcategorization frames (or predicate-argument structures) from the ATB for a large number of Arabic lemmas, including verbs, nouns and adjectives, as shown in Table 6.

	Verbs	Nouns	Adjectives
lemma-frame pairs in the ATB	6596	855	295

Table 6: Number of subcategorization frames in the ATB

Subcategorization frames are enriched with probability information that provides estimates of the likelihood of occurrence of a certain argument list with a predicate (or lemma). For example, Table 7 show the probability of each subcategorization frame with the verb $>abolaga$ أبغ ‘to inform’ which has a frequency of 103 occurrences in the ATB. The subcategorization frames are sorted by probability, ensuring that more frequent subcategorization frames appear on the top.

id	lemma_id	subcats	prob	sense
527	$>abolag_1$	subj,obj,comp-sbar	0.3398	to inform sb that
525	$>abolag_1$	subj,comp-sbar	0.165	to announce that
537	$>abolag_1$	subj,obj	0.1359	let sb be informed
529	$>abolag_1$	subj,obj,obj2	0.1068	communicate sth to sb
533	$>abolag_1$	subj,obj,obl-clr@bi	0.068	inform sb of sth

Table 7: Subcategorization frames with probability scores for the lemma ‘ $>abolag_1$ ’

6.2 Information on co-occurrence frequencies

In addition to subcategorization frames, the context in which words occur can provide key information on word senses, significant collocates and the various types of idioms, and multiword expressions in which the headword may occur. This is why the recording of co-occurrence frequencies in the corpus is essential.

AraComLex (Attia et al., 2011b), is a useful web application designed specifically for Arabic lexicographic work and provides, among other facilities, the ability to review

word frequencies at various levels: lemma, stem, full form, and contextual examples. Information is sorted by frequency, so that the most prominent senses occupy the top of the lists. Table 8 shows an example of the full forms and stems of the verb >bolaga أبلغ ‘to inform’.

id	index_id	full_form	stem	freq
90687	6998	>blg	>abolag	15235
1107949	6998	w>blg	>abolag	9421
31207	6998	>blgt	>abolag	7194
1191154	6998	tblg	bolig	3932
983221	6998	yblg	bolig	3523
838632	6998	wtblg	bolig	3343
492823	6998	wyblg	bolig	3277
114319	6998	>blgh	>abolag	2456

Table 8: Full form variations with frequency for the lemma ‘>abolag_1’

Furthermore, a lexicographer can go even deeper by reviewing the examples in which the words occurred, sorted according to frequency, as shown in Table 9. For practical reasons and to keep the size of the database within reasonable bound, we only keep records of the word’s tri-grams, which in most cases are enough to provide a glimpse of the context and possible collocates.

stem_id	example	freq.	translation
1107949	#وأبلغ#مصدر	263	a source informed
90687	انه#أبلغ#الى	75	that he communicated to
90687	#أبلغ#وزير	70	informed the minister of
90687	#أبلغ#اداري	17	an administrative official informed
114319	الذي#أبلغه#أنه	16	who informed him that he

Table 9: tri-gram frequencies for the lemma ‘>abolag_1’

7. Conclusion

We have developed a set of methods and techniques to equip modern dictionaries with self-updating mechanisms to allow them to discover new words, flush out (or mark) obsolete words and investigate word senses based on co-occurrence information. We automatically extract new words from a large corpus and lemmatize them in order to relate multiple surface forms to their canonical underlying representation using a finite-state guesser and a machine learning tool for disambiguation. We have developed a weighting mechanism for simulating a human

decision on whether or not to include new words in a general-domain lexical database. Out of 1,698,852 new words we created a lexicon of 476,349 lemmatized, POS-tagged and weighted entries. We have made our unknown word lexicon available as a free open source resource (<http://arabicnewwords.sourceforge.net/>).

We deal with the crucial maintenance problem faced by dictionaries in that, over time, they tend to accumulate a large subset of obsolete lexical entries no longer attested in contemporary data. We identify obsolete entries relying on statistics derived from a large pre-annotated corpus and website searches. We also provide essential lexicographic information by automatically building a lexicon of subcategorization frames from the ATB and information on co-occurrence frequencies.

8. Acknowledgements

This research is funded by the Irish Research Council for Science Engineering and Technology (IRCSET), and the Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngli.ie) at Dublin City University.

9. References

- Adler, M., Goldberg, Y., Gabay, D. and Elhadad, M. (2008). Unsupervised Lexicon-Based Resolution of Unknown Words for Full Morphological Analysis. In: Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio.
- Attia, M. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In: Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK.
- Attia, Mohammed, Pavel Pecina, Lamia Tounsi, Antonio Toral, Josef van Genabith. (2011a). Lexical Profiling for Arabic. Electronic Lexicography in the 21st Century. Bled, Slovenia.
- Attia, Mohammed, Pavel Pecina, Lamia Tounsi, Antonio Toral, Josef van Genabith. (2011b). An Open-Source Finite State Morphological Transducer for Modern Standard Arabic. International Workshop on Finite State Methods and Natural Language Processing (FSMNLP). Blois, France.
- Attia, Mohammed, Younes Samih, Khaled Shaalan, Josef van Genabith. (2012). The Floating Arabic Dictionary: An Automatic Method for Updating a Lexical Database through the Detection and Lemmatization of Unknown Words. COLING, Mumbai, India.
- Beesley, K. R. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In: The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects, Toulouse, France.

- Beesley, K. R., and Karttunen, L. (2003). *Finite State Morphology: CSLI studies in computational linguistics*. Stanford, Calif.: Csl.
- Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer (BAMA) Version 2.0*. Linguistic Data Consortium (LDC) catalogue number LDC2004L02, ISBN1-58563-324-0
- Crystal, D. (1980). *A First Dictionary of Linguistics and Phonetics*. London: Deutsch.
- Dalrymple, M. (2001). *Lexical Functional Grammar*. Volume 34 of Syntax and Semantics. Academic Press, New York.
- Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL)
- Dichy, J. (2001). On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases. ACL/EACL 2001 Workshop on Arabic Language Processing: Status and Prospects. Toulouse, France.
- Dichy, J., and Farghaly, A. (2003). Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? In: The MT-Summit IX workshop on Machine Translation for Semitic Languages, New Orleans.
- Erjavec, T., and Džerosk, S. (2004). Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words. *Applied Artificial Intelligence*, 18:17–41.
- Ghazali, S. and Braham, A. (2001). Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic. Arabic NLP Workshop at ACL/EACL. Toulouse, France
- Gottlieb, Henrik and Jens Erik Mogensen (Eds). (2007). *Dictionary Visions, Research and Practice. Selected Papers from the 12th International Symposium on Lexicography*. Copenhagen 2004. Amsterdam/Philadelphia: John Benjamins
- Grefenstette, Gregory, Yan Qu, and David A. Evans. (2002). Expanding lexicons by inducing paradigms and validating attested forms. Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Canary Islands, Spain.
- Huang, Chung-chi, Ho-ching Yen and Jason S. Chang. (2010). Using Sublexical Translations to Handle the OOV Problem in MT. in Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA).
- Hulden, M. (2009). Foma: a finite-state compiler and library. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09). Stroudsburg, PA, USA.

- Kiraz, G. A. (2001). *Computational Nonlinear Morphology: With Emphasis on Semitic Languages*. Cambridge University Press.
- Kilgarriff, Adam and David Tugwell. (2002). Sketching words *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Marie-Hélène Corréard (Ed.) EURALEX: 125-137.
- Lass, Roger (1997). *Historical linguistics and language change*. Cambridge University Press.
- Lindén, K. (2008). A Probabilistic Model for Guessing Base Forms of New Words by Analogy. In CICling-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics, Haifa, Israel, pp. 106-116.
- Maamouri, M., Graff, D., Bouziri, B., Krouna, S., and Kulick, S. (2010). LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.1. LDC Catalog No. LDC2010L01. ISBN: 1-58563-555-3.
- Mohamed, Emad; Sandra Kübler (2010). Arabic Part of Speech Tagging. Proceedings of LREC 2010, Valetta, Malta.
- O'Donovan, R., Burke, M., Cahill, A., van Genabith, J. and Way, A. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks, Computational Linguistics, pp. 329-366.
- Owens, J.: *The Arabic Grammatical Tradition*. The Semitic Languages. London:Routledge (1997)
- Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2009). Arabic Gigaword Fourth Edition. LDC Catalog No. LDC2009T30. ISBN: 1-58563-532-4.
- Roth, R., Rambow, O., Habash, N., Diab, M., and Rudin, C. (2008). Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking. In: Proceedings of Association for Computational Linguistics (ACL), Columbus, Ohio.
- Sinclair, J. M. (ed.). (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

On the Appification of Dictionaries: From a Chinese Perspective

Yongwei Gao

College of Foreign Languages & Literature, Fudan University,
220 Handan Road, Shanghai, P.R.C.
E-mail: ywgao@fudan.edu.cn

Abstract

The advent of the Internet and the rapid development of computer technology have brought great changes into the dictionary-making scene worldwide. Such changes are evident not only in dictionary-making processes but also in the way dictionaries are presented to their target users. In recent years, as the number of mobile users increases, the use of dictionary apps has experienced a sharp rise. This paper attempts to make a tentative study of currently available dictionary apps, with an emphasis on English-Chinese dictionary apps. In order to present a panoramic view of the appification scene, this paper will first briefly discuss the major English dictionary apps that are popular with mobile users worldwide, then survey the bilingual dictionary apps; subsequently, the pros and cons of appified dictionaries will be examined in depth. Finally, the paper will also touch upon the influences of the use of dictionary apps on dictionary-making.

Keywords: dictionary apps; English-Chinese dictionaries; bilingual lexicography

1. Introduction

With the advent of the Internet and the rapid development of computer technology since the late 1990s, the dictionary-making scene worldwide, monolingual, bilingual or multilingual, has undergone considerable change. Such a change is evident not only in dictionary-making processes such as the heavy reliance on the Internet for word information, the corpus-based data collection, the use of dictionary-writing or dictionary-editing systems, etc., but also in the way dictionaries are presented to their target users. The convenience that comes with the use of e-dictionaries that are presented in the form of handheld electronic dictionaries, online dictionaries, and so on, is likely to account for the change of opinion on the part of dictionary users who, generally speaking, now prefer e-dictionaries to their dead-tree editions. In recent years, as the number of smartphone and tablet users increases worldwide, the use of dictionary apps has experienced a sharp rise. In order to tap the potential of the vast global mobile market, dictionary publishers, large or 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. This paper attempts to make a tentative study on currently available dictionary apps, with an emphasis on a discussion of English-Chinese dictionary apps that are popular with Chinese learners of English. In order to present a panoramic view of the appification scene, this paper will first make a brief discussion of the major English dictionary

apps popular with mobile users, then move on to survey bilingual dictionary apps preferred by Chinese users. Subsequently, the pros and cons of dictionary apps will also be examined in depth. Finally, the paper will touch on the influences of the use of dictionary apps on bilingual dictionary-making.

2. Dictionary Apps

As a relatively new term, ‘dictionary apps’ refer to software applications that usually contain the content of at least one dictionary and are designed to run on smartphones, tablets, other mobile devices and personal computers. Prior to their emergence, mobile users were accustomed to another type of e-dictionary, namely, mobile dictionaries which are either embedded in mobile phones or downloaded online. Since Apple launched the ‘App Store’ in conjunction with the release of iPhone OS 2.0 in July 2008, applications developed by third parties have been distributed worldwide, and dictionary apps have gradually been made available by either dictionary publishers or IT companies. Paragon Software Group, for example, is the leading software developer of electronic dictionaries for mobile devices and desktop computers and a key player in the appification scene, which launched its line of dictionary apps in 2009. Available through various app distribution platforms (e.g. the Apple App Store, Google Play, Windows Phone Store, Samsung Apps etc.), dictionary apps are being downloaded by users all over the world. As the use of smartphones and tablets worldwide is increasing exponentially, the number of downloads for dictionary apps is on the rise. Certain popular dictionary apps may have been downloaded millions of times. For instance, according to its website, *Dictionary.com*, the famous dictionary aggregator, has so far been downloaded more than 60 million times. The popularity of dictionary apps can also be attested by the sheer number of such apps made available in various app stores. As of April 15, 2013, there are 3,998 and 8,303 results when the word “dictionary” is searched at the Apple App Store via an iPad and an iPhone, respectively¹. As with traditional dictionaries, dictionary apps can also be classified into different categories in terms of the language(s) being used, the subject matter, and the coverage of the vocabulary of a language, etc.

In terms of the language(s) being used, a dictionary app may contain at least a monolingual, bilingual, or multilingual dictionary. Let us take *Longman Dictionary of English* as an example. Launched as early as 2011 by Pearson Education, this Android dictionary app is a monolingual dictionary that offers 230,000 words, 77,000 audio pronunciations, and 86,000 recorded example sentences. Sometimes a dictionary app may consist of more than one language. *Collins COBUILD English/Chinese/Japanese/Korean Advanced Dictionary of American English*, for

¹ Due to the fact that dictionary apps are mostly designed for mobile devices, the same search performed via a MacBook yields only 204 results.

example, is a quadrilingual dictionary that contains over a quarter of a million translations from English to Chinese, Japanese and Korean.

In terms of the subject matter covered in dictionaries, a dictionary app may include dictionaries of virtually every subject, such as *Dictionary Business Terms* which is developed by Intersog, LLC and contains a multitude of commonly-used business terms and concepts; *Chemical Terms Dictionary* which is compiled by The CJK Dictionary Institute in Japan and contains over 243,000 entries; and *Kids Picture Dictionary* which is designed for children to learn their first words and make sentences with a fun record tool, etc.

As is often the case with traditional dictionaries, the more popular dictionary apps are always those compiled for the purpose of facilitating the learning of English. As a result, learners' dictionaries are ubiquitous, and so are dictionaries that focus on certain aspects of vocabulary, such as idiom dictionaries, slang dictionaries, etc. Let us take an idiom dictionary as an example. Dozens of English idiom dictionaries are available at distribution platforms such as the Apple App Store. *Idiom in Use—Advanced English Idioms Dictionary*, for instance, is an app that includes the 750 most-used idioms and collocations of the English language, which often feature in English tests for foreign speakers.

When it comes to the virtues of electronic dictionaries, Gilles-Maurice de Schryver discussed several kinds of “eases”, such as the electronic ease and the online ease (2003: 152–158). His views were echoed in what Jennifer Howard wrote in the *Chronicle of Higher Education* on March 11, 2013: “For dictionary makers, going electronic opens up all kinds of possibilities. It's not just that digital dictionaries can be embedded in the operating systems of computers and e-readers so that they're always at hand. They can be updated far more easily and often than their print cousins, and they can incorporate material like audio pronunciations and thesauruses.” Indeed, the electronic medium does offer dictionary-makers a number of new options that were unavailable until recent years. Such options enable lexicographers to devise features typical of this new category of dictionaries. These features may account for the popularity that dictionary apps are currently enjoying. As a matter of fact, most of these features overlap with those of online dictionaries, and include:

- A. Virtually unlimited space. Unlike traditional dictionaries that are usually encumbered by the limitation of scope, a dictionary app is usually free of such restrictions and may include a larger content. Although an app may occupy several megabytes, this is negligible in a device comprising several or dozens of gigabytes. Unlimited space offers the lexicographer a variety of choices, such as the addition of many entries, the provision of multimedia content, the listing of related words, or the inclusion of more than one language or dictionary, etc.
- B. Easy updating. The fact that updates of traditional dictionaries occur at much longer intervals than their electronic counterparts is one of the reasons for the

diminishing readerships of traditional dictionaries. Like online dictionaries, dictionary apps can in theory be updated much more easily and frequently, which might serve to keep users up-to-date with the latest changes in vocabulary.

- C. Multimedia presentation of microstructural information. The provision of audio pronunciations for headwords or even illustrative examples in online dictionaries or apps is a big plus compared to traditional dictionaries that can only offer phonetic transcriptions. For some specialized dictionary apps, the use of animation is also a unique way of presenting an entry.
- D. More search options. One such option is the ‘wild-card search’. This feature, common in online dictionaries, has also become a fixture in dictionary apps and it is particularly helpful to users who are not sure of the spelling of a word. Another option, termed here “secondary search”, allows users to select any word in a definition in a dictionary to see the desired result.
- E. Provision of additional features aimed at facilitating language learning. *Word of the Day*, for example, is a feature present in many dictionary apps and is designed to provide additional information about a chosen word each day. Other features, such as search history and favorites, can also be of some assistance to dictionary users.
- F. Easy cross-referencing. Dictionaries are intended to be an interconnected web of words. Although traditional dictionaries do offer some sort of cross-referencing, it is by no means satisfactory or complete. An ideal cross-referencing system should provide cross-references for all words with which a particular headword is connected in one way or another. For example, in *WordWeb*, cross-references are set up in several ways for the entry *black*, such as the indication of its synonyms (*African-American*, *Afro-American*, *colored*, *dark*, *dark-skinned*, *negro*, *negroid*, *non-white*) and its antonym (*white*), the provision of derived words (e.g., *blackness*, *blacken*, *blackish*, *blackly*, etc.), and the link to related words (*black and white*, *black market*, *black marketer*, *black out*, *bluish black*, *in the black*²), etc.
- G. External links to other reference works. Some dictionary apps offer links to other dictionaries when the word being sought is not present. For example, *WordWeb* has links to offline references such as *Chambers Dictionary* and *Chambers Thesaurus*, and online references such as *Wikipedia*, *Wiktionary*, *Answers.com*, etc.

Moreover, dictionary apps also offer unique features. First, some apps may provide fuzzy search for similar sounding words. Electronic dictionaries have often been criticized for their inability to be read like a book. When looking up a word in a

² This list of related words seems to be rather arbitrary and far from complete.

traditional dictionary, a user has the luxury of browsing nearby entries, which has always been lauded as a wonderful reading experience. Fuzzy search in dictionary apps might offer a different kind of browsing experience: that is, to browse entries that share similar pronunciation with the word being sought or are located adjacent to the said word. *Oxford Slang*, for instance, is a free app based on John Ayto and John Simpson's *Oxford Dictionary of Modern Slang*. Its fuzzy-search function enables users to browse entries such as *crock*, *crock of shit*, *rock of ages*, *rocket*, *rocky*, and *schooner on the rocks*, when one looks up the word *rock*. Second, some apps also allow voice search. *Dictionary.com*, for example, allows users to say the word they intend to look up. Though such a feature is by nature a bells-and-whistles one, it does offer some convenience to users.

Providing entries in multilingual languages seems to be another hallmark of a new generation of comprehensive dictionary websites, such as *Dictionary.com*, *TheFreeDictionary*, *YourDictionary.com*, *WordReference.com*, etc. Such sites usually offer “one-stop shopping” to users who intend to look up words there, and some of these dictionary aggregators have also begun to tap the potential of the apps market and launched their respective dictionary apps. As they are mostly free of charge, they are downloaded far more often than apps that do charge. As searchability rules in the world of dictionary apps, dictionary apps tend to include as much information (e.g. entries, languages, etc) as possible, thus being more inclusive in coverage seems to have become the norm.

3. English Dictionary Apps

For the classification of online or Internet dictionaries, de Schryver divided networked dictionaries into just two categories, namely intranet and Internet dictionaries (2003: 151). Pedro A. Fuertes-Olivera also proposed two main types: namely, institutional reference works and collective free multiple-language Internet reference works (2009: 103). Gao Yongwei put forth a three-type typology that includes “clicks-and-mortar” dictionaries, one-stop dictionary sites, and DIY dictionaries (2012: 423-426). The English dictionary apps scene is more or less dominated by these three types of dictionaries.

The first type of dictionary apps refers to those based on existing English dictionaries. Almost all major dictionary publishers in English-speaking countries have developed apps for their dictionaries. Oxford University Press, for instance, has developed apps not only for its learners' and general dictionaries, such as *Oxford Advanced Learners' Dictionary*, *8th edition*, *Oxford Dictionary of English*, *Australian Oxford Dictionary* and so on, but also a wide range of specialized dictionaries such as *Oxford Dictionary of Computing*, *Oxford Dictionary of Food and Nutrition*, *Oxford Dictionary of Finance and Banking*, *Oxford Dictionary of Biology*, *Oxford Concise Dictionary of Politics*, etc. As reported on *PR.com* on March 30, 2013, Oxford Dictionaries is making a greater effort to provide dictionary content to mobile users: “The number of

dictionary searches made on mobile devices and smartphones continues to increase ... In order to cater for a growing mobile audience, a fully responsive and adaptive site design is necessary as it gives Oxford Dictionaries Online users an optimized experience regardless of the device they are using to access our free dictionary content. We are always looking for ways to optimize our free online dictionary and we are confident this responsive website design will improve the functionality of Oxford Dictionaries Online on mobile devices, gaming devices, tablet devices, smartphones, and laptops.” Macquarie Dictionary Publishers has also made considerable effort in developing a series of apps for use on both Android and iOS devices, such as *Macquarie Senior Student Dictionary*, *Macquarie Complete Australian Dictionary*, *Macquarie Concise Australian Dictionary*, *Macquarie Essential Australian Dictionary*, *Macquarie Lite Australian Dictionary*, and *Macquarie Aussie Slang Dictionary*. As the majority of such dictionary apps (e.g. *American Heritage Dictionary, 5th Edition*, *Collins English Dictionary Unabridged*, *Merriam-Webster’s Collegiate Dictionary, Eleventh Edition*, *Webster’s New World Dictionary*, etc.) have the same content as existing paper dictionaries on which they are based, no further discussion will be made of them.

The second type of common dictionary apps consists mainly of existing multiple-language Internet reference works. Among them, *Dictionary.com* is undoubtedly the free dictionary app of choice. Boasting over two million words and definitions to date, this award-winning dictionary app has been a favorite with many dictionary users³. Launched in 1995 by Lexico Publishing, LLC, *Dictionary.com* now attracts more than 50 million users across the globe every month to its online English dictionary and thesaurus. As *Dictionary.com* claims on its website, it has become the world’s largest and most authoritative free online dictionary and mobile reference resource. Inspired by its goal “to empower word discovery and learning”, *Dictionary.com* teamed up with dictionary publishers such as HarperCollins and Random House to provide content for global users, so far obtaining 15 licenses from proprietary reference sources, such as *Collins English Dictionary*, *The American Heritage Science Dictionary*, *The American Heritage New Dictionary of Cultural Literacy*, *The Free Online Dictionary of Computing*, etc. Besides voice search mentioned above, this app has another unique feature - ‘Translator’ – which offers translations in various languages, such as Arabic, Chinese, French, German, and Italian. Moreover, this dictionary app also offers users the possibility to get 850,000 example sentences for less than 2\$. However, the app has not been updated as it should have been because some relatively new terms such as *bromance* and *e-shopping*, which have already been included in the online version, cannot be found there.

³ In *Incredible iPhone Apps For Dummies* written by Bob LeVitus and published in 2010, *Dictionary.com* was said to be “probably the best of the free dictionary and thesaurus apps currently available” and then it only included 275,000 definitions and 80,000 synonyms.

The third type refers chiefly to the growing number of English dictionary apps that are either developed especially for this new medium or converted from a monolingual online dictionary or lexical database. Pure online dictionaries such as *Wiktionary* and *Urban Dictionary* have been appified. For example, *Wiktionary* can be found not only in *Wikipanion*, an app mainly featuring *Wikipedia* entries, but also in *English Dictionary – Offline*, an app that includes 159,000 words from *Wiktionary* and *EN*, an app that includes 185,000 word definitions from *Wiktionary*. *Urban Dictionary* also has its app presence, which is rather commendable in terms of being up to date when it comes to its entries. The famous lexical database *WordNet* has been fully utilized by app developers, as is attested by the fact that several apps base their contents on the lexical database. For example, *WordBook XL-English Dictionary & Thesaurus*, said to be “the top-selling English dictionary on the app store since 2008” with 150,000 entries, 220,000 definitions, and 70,000 usage samples, is based entirely on *WordNet* although no claim of this kind has been made in the introduction on the part of its developer TranCreative LLC. *WordWeb Dictionary*, an app with 285,000 words, 225,000 word sense definitions, 70,000 usage examples, and 85,000 text pronunciations, was developed in a similar manner. Although there is no indication of the dictionary on which the app is based, the entries it includes are certainly taken from *WordNet*⁴. The only differences between *WordBook XL* and *WordWeb* lie in the different ordering of senses and the provision of synonyms. A case in point is the entry *à la carte*:

WordBook XL	WordWeb
n. a menu having individual dishes listed with separate prices	Adverb: By ordering items listed individually on a menu <i>we ate à la carte</i>
adj. (of a restaurant meal) having unlimited choices with a separate price for each item	Noun: A menu having individual dishes listed with separate prices ~ bill of fare, card, carte, carte du jour, menu
adv. by ordering items listed individually on a menu <i>we ate à la carte</i>	Adjective: (of a restaurant meal) having unlimited choices with a separate price for each item table d’hote

Table 1: Treatment of *à la carte*

There are other miscellaneous English dictionary apps, such as *Dictionary*⁵ and *HE Lexicon*⁶, to mention just a few.

⁴ There are two other apps that are fully based on *WordNet*, and they are *Dictionary!* and *LexicEn Lite*, which are developed by Catlin Software, LLC, and www.gogonavi.net, respectively.

⁵ It is actually a collection of offline dictionaries that include *The Collaborative International Dictionary of English*, *Webster’s Revised Unabridged Dictionary* (1913), *WordNet 3.0*, an unspecified English dictionary, and a picture dictionary.

⁶ Coincidentally, this app also bases much of its entries on *WordNet*.

4. English-Chinese Dictionary Apps

The English-Chinese and Chinese-English dictionary app scene is less crowded than its English counterpart, which may be attributed to the fact that few dictionary publishers are willing to develop app versions of their brand-name dictionaries. Their resistance to doing so can be ascribed to reasons such as being contented with the status quo and lack of innovativeness, etc. Beijing-based Foreign Language Teaching and Research Press is one of the rare few dictionary publishers in China that have developed dictionary apps. *Oxford-FLTRP English-Chinese Chinese-English Dictionary*, one of FLTRP's flagship dictionaries, now has an app presence. Although boasting about 300,000 words and phrases and 370,000 illustrative examples, however, this dictionary app has not been much favored by Chinese learners of English, partly because of its relatively hefty price tag and partly because of fierce competition from domestic developers of dictionary apps.

Similarly, in the online-dictionary scene in China, IT companies are the dominant players, including *Youdao*, *Kingsoft Power Word*, *Dict.cn*, etc. *Youdao*, better known for its desktop dictionary, has already extended its tentacles into the mobile world and developed dictionary apps for different types of mobile devices. As a multilingual dictionary app, *Youdao* has been the top downloaded dictionary among Chinese users. Its English-Chinese and Chinese-English parts include 340,000 and 330,000 entries respectively. Two English-Chinese dictionaries, namely *The 21st Century Unabridged English-Chinese Dictionary* and *Collins Comprehensive English-Chinese Dictionary*, form the backbone of *Youdao*'s English-Chinese part. A unique feature of *Youdao*'s app lies in the multiple choices one has when selecting definitions for the word being searched. A simple search will offer four types of definitions (i.e. online⁷, technical, English, and pictorial) along with the Chinese equivalents from the above-mentioned two English-Chinese dictionaries. In a similar vein, one can also choose to view different types of illustrative examples—bilingual ones, those with audio pronunciation (taken from VOA), and authoritative examples (some of which are taken from news reports). The number of examples in each type can be as many as thirty and most of them are captured online. Other features of this app include searching within an encyclopedia, real-time translation, etc.

Kingsoft Power Word, as it claims on its website, is “currently the world’s largest Learner Dictionary” as it contains more than 355,000 word articles, phrases and definitions, and selections of more than 5,000 new words and meanings. Kingsoft has long been known for its powerful desktop dictionary system that incorporates scores of dictionaries or specialized lexicons, such as *English-Chinese & Chinese-English Dictionary*, *A Glossary of Physiological Terms*, *A Glossary of Electronic Terms*, *A*

⁷ This type of definition is usually taken directly from online sources, usually including all the possible translations one can find online for the searched word.

Glossary of Terms in Chemical Engineering, A Glossary of Computing Terms, etc. Kingsoft's app, however, seems to be a watered-down version of its desktop dictionary as it comprises only a small selection of dictionaries such as *The American Heritage Dictionary of the English Language*, *WordNet*, *Collins COBUILD Advanced Learner's English-Chinese Dictionary*, a dictionary of synonyms and antonyms, and a dictionary of phrases and collocations, etc. A simple search in its no-frills version usually results in four definitions, namely, basic definition, authoritative definition, *Wiktionary* definition, and English definition. The first type is furnished with Chinese equivalents, abundant illustrative examples, antonyms and synonyms, phrases and collocations, etc. The authoritative definition is taken from the bilingualized Collins dictionary while the English one is copied from *WordNet*. The third type, though claimed to be taken from *Wiktionary*, bears no relation to the online dictionary and only provides Chinese equivalents. Besides the voice search and the provision of translation and news, this app is unique in that its in-app camera enables a user to point at a particular word to obtain its meaning.

Established in November, 2003, *Dict.cn* is a dictionary site that offers a wide range of services, such as dictionary lookup, sentence and paragraph translation, online sources, dictionary software download, etc. Its dictionary app, like its competitors, offers four types of definition—basic, bilingualized, detailed, and English. Its English definitions are also fully based on *WordNet*. Designed to be learner-friendly, the app features abundant illustrative examples, common sentence patterns, common phrases, collocations, and even quotations from classical works. Moreover, the app also includes other dictionary features such as usage notes, etymology, and the provision of synonyms and antonyms.

Eudic and *nciku* are two other dictionary apps developed by IT companies. The former, having a collection of 300,000 English-Chinese and Chinese-English entries, is notable for the laundry list of phrases and idiomatic expressions and its vast collection of miscellaneous online examples. The latter, boasting 163,000 entries, is chiefly based on two dictionaries—*Collins English-Chinese Dictionary* and *nciku*'s own comprehensive English-Chinese dictionary. Other lesser-known English-Chinese dictionary apps include *Dict Box*, *HEdictEC*, *CZ English-Chinese*, etc.

Thanks to their comprehensive coverage of the English vocabulary and a wide range of user-friendly features, the above-mentioned English-Chinese dictionary apps more or less cater to the needs of a myriad of English users in China. Nevertheless, as most such apps are merely a hodgepodge of dictionaries, monolingual and bilingual, they are deficient in many ways. First, in some apps, the selection of entries is arbitrary, and the criteria are rather loose. *Dict.cn*, for example, includes many words (mostly compound ones) which should not be recorded in dictionaries as their meanings are readily understood or they are not frequently used, such as *foodaholic*, *electronic components*, *personal information*, *warehouse management*, etc. Sometimes, the same headword may have been listed twice. *PowerWord*, for instance, includes both

mood swing and *Mood swings*, and provides different Chinese equivalents for them – “情绪波动” and “心境不稳”. Unlike their monolingual counterparts that are more or less based on name-brand general dictionaries, some of the dictionary apps mentioned above base themselves on a considerable number of minor dictionaries, mostly technical ones, and as a result, some of them include way too many technical terms, some of which being rather out of place. *CZ English-Chinese*, for example, includes *cybernetics* along with many related terms such as *cybernetic machine*, *cybernetic model*, *cybernetic simulator*, *cybernetics system*, etc. Second, some apps are riddled with awkward or inappropriate Chinese translations either of headwords or of illustrative examples, such as *cyberspeak* 虚拟对话 (literally meaning “virtual dialogue”), *cybrarian* 电脑族 (literally meaning “computer clan”), etc. The translations provided for illustrative examples are problematic in many ways. No matter whether they are taken from bilingual dictionaries or culled from the Internet, problematic translations abound, such as “Perot hoped to run another series of campaign infomercials. 佩罗期待着新一轮的竞选宣传节目的播出” and “At least in terms of bioterror attacks, I can’t imagine recommending evacuation. 起码在生物恐怖进攻时，我不能预测所谓的疏散”. Third, the illustrative examples in some apps are grammatically incorrect, thus eventually misleading English learners. Such examples include “He wrote off for information on Internet” from *Youdao*, “Internet make people more intimacy or alienation?” and “Examination result is appalling, urban chophouse, small noodle shop uses this to plant doubtful and lardy very general” from *Eudic*, “Rich as he may seem, he works in a fast-food as waiter” from *nciku*, etc.

5. The Influences of Appification on Bilingual Lexicography

Although many dictionary apps are based on existing dictionaries, traditional or online, their emergence on the dictionary-making scene will undoubtedly exert some influence upon contemporary dictionary-making, whether monolingual or bilingual. The influence upon the making of English-Chinese dictionaries is much greater as English-Chinese lexicography has been impeded by factors such as heavy reliance on monolingual dictionaries, and lack of innovation, etc. Such influences will bring about changes in the following aspects:

A. Wide coverage of the English vocabulary. Chinese learners of the English language usually expect more entries from their dictionaries as they have been accustomed to the concept of “more is better” when it comes to the number of entries in a dictionary. Traditional dictionaries such as *A New English-Chinese Dictionary* and *The English Chinese Dictionary* are known for the large vocabulary they cover. Getting their money’s worth is the prevailing mindset when people purchase a dictionary. This “more is better” concept is also prevalent when people choose electronic dictionaries. The popularity of *Lingoes*, a free desktop reference tool into which one can download a wide range of dictionaries, and *Casio*, a famous brand for a series of hand-held dictionaries that include scores of existing dictionaries, is another indication of people’s preference for reference tools with a lot to offer. In order to better cater to the

needs of Chinese dictionary users, dictionary-makers should make greater efforts in providing improved dictionary content. As *The English-Chinese Dictionary*, the largest English-Chinese dictionary now available, is trying to diversify its mode of presentation, more efforts should be made to enlarge its coverage of the English vocabulary. Though boasting 220,000 entries in its second edition, ECD has a long way to go in terms of including as many English words as possible. As a result, not only new words but also popular regional uses of English words which were rather inadequate in its current edition, should be included in the third edition, due in 2017. The fact that an app can be updated easily and frequently will also force dictionary-makers to be on the constant lookout for neologisms.

B. Provision of better Chinese equivalents. Some dictionary apps are designed in such a way that the traditional boundary of English-Chinese dictionaries and Chinese-English ones is blurred as one can search an English word as well as a Chinese one in the same dictionary database. *KTdict* is a case in point. If one searches the word *beauty* in its Chinese-English dictionary, it will offer all entries containing *beauty* in its English definitions, such as *beauty/belle* for 美人, *beauty contest* for 选美, *beauty salon* for 美发院, *a woman of unmatched beauty* for 绝世佳人, *beauty in the eye of the beholder* for 情人眼里出西施, etc. Such a bidirectional feature empowers dictionary users while laying bare the problems in the translations of some dictionary headwords. In the past, dictionary-makers seldom made comparisons between the translations for headwords with similar or related meanings, which usually resulted in inconsistencies in translation. For instance, *burn one's boats / bridges* and *cross the Rubicon* are defined differently in English, but they are often translated into “破釜沉舟” in Chinese, as is shown in the translations provided by *ECD*:

burn one's boats / bridges: 破釜沉舟, 背水布阵, 自绝退路

cross the Rubicon: 采取断然行动 (或手段); 下重大决心; 破釜沉舟

Therefore, such an app-enabled feature should prompt dictionary-makers to review their usual translation practice and then fine-tune their translations.

C. Furnishing of new features. The provision of synonyms and antonyms, though a fixture in learners' dictionaries, is absent in most English-Chinese dictionaries, partly because Chinese equivalents provided for synonymous or antonymous may not be synonymous or antonymous in Chinese. This is caused by the time-tested lexicographic practice of providing as many Chinese equivalents as possible for many English headwords. As early as 1908 when one of the first major English-Chinese dictionaries was published, lexicographers started to provide as many Chinese equivalents as possible for some headwords in order to make sure that all shades of the meaning were recorded. This practice was adopted by later dictionary-makers, and many existing dictionaries are still riddled with examples of this kind. As a result,

there might be at least three or four Chinese equivalents in some entries. Take *lack* for example. *ECD* provides three equivalents for its second sense, namely “需要；需要的东西；缺少的东西”. If we want to provide a synonym for this sense, “need” will be the first word that comes to mind. However, it differs from the definition provided by English dictionaries: “something that is lacking or is needed”. As a consequence, Chinese dictionary-makers would have to review many Chinese equivalents if features such as synonyms and antonyms are to be furnished.

More should be done if dictionary-makers in China want to convert their dictionary data into apps, such as better ways of cross-referencing, and separate listing of run-on entries, etc.

It is quite obvious now that dictionary apps, as a new way of presenting dictionary entries, will surely be here to stay as long as smartphones and tablets are used. However, the rise of dictionary apps does not necessarily spell the demise of paper dictionaries. But as more and more dictionaries are made available in app form, the traditional dictionary scene will change forever. With such changes, users’ reading habits will change accordingly. As dictionary-makers, it is high time to adapt to such changes and make special efforts to improve or even enlarge dictionary data so as to meet the changing needs of a generation of language users.

6. References

- de Schryver, Gilles-Maurice (2003). Lexicographers’ Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2): 144-199.
- Dung, Joseph (2009). Online Dictionaries in a Web 2.0 Environment. In Henning Bergenholtz, Sandro Nielsen & Sven Tarp (eds). *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp.135-164.
- Fuertes-Olivera, Pedro A. (2009). The Function Theory of Lexicography and Electronic Dictionaries: WIKTIONARY as a Prototype of Collective Free Multiple-Language Internet Dictionary. In Henning Bergenholtz, Sandro Nielsen & Sven Tarp (eds). *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp.99-134.
- Gao, Yongwei (2012). Online English Dictionaries: Friend or Foe. *Proceedings of Euralex 2012*, Oslo, 7-11 August 2012.

What should the electronic dictionary do for you – and how?

Oddrun Grønvik, Christian-Emil Smith Ore

University of Oslo, ILN, Pb 1021 Blindern 0315 Oslo
E-mail: oddrun.gronvik@iln.uio.no, c.e.s.ore@iln.uio.no

Abstract

Language is a common good and a common property. Access to information about language should be fast, easy, and intuitive. The electronic dictionary should therefore be a knowledge base with language as its access point, and with simple, yet rich access to (combinations of) linguistic and non-linguistic facts. One query frame and basic reading and writing skills must be enough to get meaningful results. This solution presupposes (1) a fine grained and systematic database format for dictionary storage and linkage to materials, and (2) a query system offering ease of access for inexperienced users. At the same time, lexicography must be able to prove itself trustworthy by offering access to sources both for usage and for normative decisions. The system described here is used for one academic multivolume dictionary and for standard monolingual students' dictionaries. It is suited to lexicographical projects where source documentation has priority. The focus is on dictionaries integrated with other language resources and produced for the Web.

Keywords: electronic dictionary, relation database, database linking, database entry format, the Meta Dictionary, full form register, indexing source materials, linking source materials to product.

1. Introduction

Electronic lexicography and language analysis is moving from the research and experimentation stage to becoming mainstream. In this setting, attempts are made to work out and present generic solutions. Our argument is that while important steps forward have been made, the present models for generic solutions are too limited, and in particular fail to take into account the importance of documentation as a method for building trust and consensus around lexicographic products.

The issues discussed in this paper are based on our experience with the electronic formats and solutions developed for *Norsk Ordbok* (NO) and the standard one-volume monolingual dictionaries *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB). We also draw on experience from projects aimed at promoting monolingual lexicography for African languages¹.

¹ the ALLEX Project (1991–2006) which dealt with the African Languages of Zimbabwe, and the CROBOL Project 2006–2011, which dealt with cross border languages involving Zimbabwe, Mozambique and South Africa.

A model for lexicography encompassing

- collecting materials
- analysing materials
- writing dictionary entries
- supervising flow
- presenting the finished product in an optimally accessible fashion is enough in a language community where
- the written standard is fixed and has been more or less unchanged for a long time
- there are plenty of materials documenting the standard through a long time span
- the community is used to using dictionaries
- the community is used to trusting its dictionaries (and there are plenty of them for comparison)

The model above is in short a sufficient model for language communities where there is general agreement on what the written standard looks like and how it is used. Dictionary making can then build on a general consensus concerning the object to be described, which is the lexicon of the language in question.

This is the situation for many of the world's major languages, especially for an important group of European languages.

A recent and very good lexicographical handbook, the *Oxford Guide to Practical Lexicography* (Atkins and Rundell, 2008) presents a model for dictionary projects suitable for dictionary making language communities of this kind. A similar understanding of lexicographical needs underlies Schryver (Schryver, 2011) in his presentation of *TshwaneLex*.

We argue that in many of the world's language communities, this model is insufficient, because it assumes trust, instead of including mechanisms that pre-empt distrust. There are plenty of dictionaries that look trustworthy, and could be produced following this model to the last letter, but are skewed in their selection of materials and lemmata, are incomplete in their presentation of orthography and word senses, and so on.

The reasons for skewed lexicography may be ideological (promoting one particular world view) or in favour of a certain language variant (presented as valid for the whole language community). It may also have to do with the ease of production (imposing a standard and omitting variants for languages which do not have a written standard). The result can very easily be general distrust, not of a certain dictionary,

but of dictionaries and reference works in general. So, as language is so important to people, lexicographers need to be trusted – not as missionaries of a particular cause, but as providers of facts of life.

The only way of dealing with distrust and building trust in linguistic reference works, is to take suspicion and the need for external control for granted, by integrating access to the raw materials (for the whole, and for each entry and sense) into the dictionary model itself. Access to the lexicographical sources has to be easy to obtain and easy to understand, from the Web.

We therefore propose a model encompassing the following stages:

1. Collecting and preparing materials (including referencing and marking)
2. Indexing materials to collect variant forms
3. Generating entries from indexed materials, with a link to the materials
4. Analysing linked materials
5. Generating entry head from a separate full form register
6. Writing dictionary entries, linking materials to each sense
7. Supervising flow
8. Presenting the finished product in an optimally accessible fashion
9. Using a staged search system that first searches the headword register, then other fields

This model is an ideal. In the following we will base our argument on the collective experience of the Norwegian Dictionary 2014 project. Most of the examples are taken from this project². *The Norwegian Dictionary* NO aims at providing a scholarly and exhaustive account of the vocabulary of Norwegian dialects from 1600 to the present and of the written standard Nynorsk since 1853.

A common challenge in editing historical and dialect dictionaries is the heterogeneity of the source material. Since NO covers sources for speech and writing through 400 years, this heterogeneity must be handled both diachronically and synchronically. The source material spans from modern texts, via traditional paper slips to local dialect dictionaries and word lists dating back to the 17th century. The interpretation and use of these materials call for explicit referencing and preferably linking to the source material so that users can check the basis for the editors' conclusions.

² The Project *Norsk Ordbok 2014* (The Norwegian Dictionary) to be completed in 12 volumes in 2014.

2. Collecting and preparing materials

In all modern introductions to lexicography the text corpus is presented as the chief electronic source. In our case, the digital sources are of several kinds³. Materials in electronic form can include images of for instance manuscript pages, and their transcripts. For languages with a weak standardization or with several orthographies it is not a trivial task to build a lemmatized and POS tagged corpus. To be able to include all texts in a homogeneous corpus one has to encode the text at three levels: The original word form, a standardized word form and a lemma form. The two latter have to be taken from an orthographical standard chosen for the entire corpus. This process is hard to computerize and is therefore very resource demanding. For reference, check the Menota guidelines for medieval Nordic texts (www.menota.org). Norwegian orthography has been thoroughly revised several times during the last 150 years. A POS-tagger developed for modern Norwegian has a very low success rate for text from the first half of the 19th century. Therefore, only the modern part of our text corpus⁴ is lemmatized and given a POS mark-up. This is clearly not a problem confined to Norwegian. This is a problem in creating corpora for all languages with changing orthography over time or for weakly standardized languages.

A second challenge is source material which is not running text, e.g. slip archives and older dictionaries and word lists. Including already synthesized information in the source material of a dictionary project obviously requires great caution, and deep philological expertise. The editorial text of old dictionaries may not be written in the language to be documented, e.g. in our case the editorial texts are in Danish or occasionally Latin. When the running text of these sources is made available electronically, the sources are not included as corpus text, but stored and referenced to the indexing system for the electronic language collections, see below.

3. Indexing materials to collect variant forms

For highly standardized languages like the major modern European languages, a lemmatized and POS-tagged text corpus stored in a standard corpus system gives an excellent and coherent access to the source material. For the less standardized languages with many heterogeneous sources a common indexing system is needed to group variant forms according to the standard that will be used for the headword of a dictionary. This is equally important whether the task is collating forms in ancient manuscripts or attempting to standardize a language for the first time.

In the case of NO, a common indexing system called *Metaordboka* (the Meta

³ The Norwegian language collections, dating back to the 1930s, were computerized in the 1990s.

⁴ Texts published after 1938 comprise the modern part of the corpus, about 85 % of the total 90 mill.

Dictionary) (MO) was designed (see Ore, 1999 and Ore & Ore, 2010). The original motivation was to create a common web-based interface to the huge lexicographic materials digitized in the 1990s. MO was later redesigned to become a pivot in the combined source database, text corpus and editing system for NO. An index entry in the MO can be seen as a folder containing pointers to (possibly commented) samples of word usage and word descriptions found in the linked sources. Each entry is labelled with a normalized headword, POS information and the source word form. The linked sources cover the ground from glossaries compiled for the Danish state administration in the 17th and 18th centuries to modern dialect surveys and local dictionaries. The MO has proved itself a very useful tool in the practical editing of NO, as well as an invaluable tool in managing the Norwegian standard language Nynorsk.

For NO the task of collating variant speech and written forms to index forms in the MO includes adding POS information, so that identically-spelt lemmata with different POS get separate entries. Index forms of compounds are marked to show joins, very important in dealing with a compounding language like Norwegian:

headword	POS	Status	Nr
fisk*e*saks	noun fem	recent	1
fisk*e*sal*s*lag	noun masc	OK	4

Figure 1: The Meta Dictionary - normalization categories.

The join marks facilitate searching for end and middle parts of compounds, to keep an eye on productivity, semantic developments etc.

MO is an independent system component that can be linked to many different lexicographical projects. It has in itself become a valuable repository. The old and the local dictionaries are kept in their original form as individual works expressing the language view of their time and author. The bidirectional linking in the system makes each headword in a source an entry point to the entire system (including NO), thus enabling dialect users a unique opportunity to see their dialect in the larger context.

All the collections coordinated under MO as the source material index are searchable in themselves. Some have the standard form of their lemmata as part of their original information, as mentioned above. Many do not, and are standardized only through their link to the MO. Both synchronic variation and diachronic heterogeneity can be a challenge, as shown below:

kjiru, kjuru, kjury, kjære, tjere, tjære tjøre, tjyru, tjörru
--

Figure 2: The Meta Dictionary - headword forms found in directly indexed materials for the noun *tjøre*, 'tar'.

The language collections coordinated through MO are under constant maintenance. One index entry can have several thousand items connected to it. In the standardization frame, index entries show standardization level by their status, cf. the Status column shown in Figure 1. Items can be moved from one index entry to another using “cut” and “paste”. The MO is a very flexible tool, and looking after it is a specialized skill, closely allied to work with language standardization in general. MO is an important source of information for the Language Council in Norway, the state agency that deals with language issues⁵, and is accessible on the Web for the general public.

4. Generating entries from indexed materials, with a link to the materials

An important aspect in trustworthy dictionary databases is that it should not be possible to create entries with no source bound materials showing form and usage. The dictionary databases of the Norwegian language collections do not permit the generation of a new entry unless it is linked to an index entry with adequate materials behind it.

If editors encounter unedited and undocumented lemmata that should be included in the dictionary, they first have to collect and register the documentation in MO, as a corpus text or as one or more electronic excerpts.

In the NO2014 bibliography⁶, sources are marked for genre and other qualities. The marking is used to generate advice to editors on whether a lemma merits an entry. If an entry in the MO f.i. is documented only in one work of fiction (a literary hapax), the advice will be not to include it. If it is documented only in older standard dictionaries, the advice will be the same. The editor can overrule this advice, or change it by adding better materials to MO.

5. Analyzing linked materials

We agree with Atkins and Rundell that the linguistic information contained in the documentation for each entry needs to be analyzed, and that the analysis needs to be conserved for future (re)use (Atkins and Rundell, 2008: 98 f.). We do not agree that analysis should be a separate task from editing. The editor needs to do both. This is of particular importance if language standardization is a permanent task. In NO, many lemmata are described in a dictionary entry for the first time.

If the dictionary source material is a giant corpus, ensuring at least 500 usage

⁵ See <http://www.sprakradet.no/>

⁶ Yet another independent but linked database, drawing its bibliographical information from the Norwegian National Library

examples of each lemma qualifying for entry (Atkins and Rundell, 2008), running a statistical analysis on them all is an obvious course of action.

This is something we would like to try, but only for a very small part of the 300,000 lemmata to be edited in NO. Since the language we deal with, Norwegian, is a compounding language with a medium rich inflection system, the section of the language collections occurring 500 times or more is much smaller than for English, be it word forms or lemmata. In a corpus of 90 million tokens, only about 1% of tokens occur 500 times or more, and well over 50% of tokens are single occurrences (hapax forms). Of more than 570,000 entries in the MO, fewer entries (i.e. lemmata) than 1:1000 have 500 or more items of documentation, while roughly 50% are (as yet) hapax forms. Many of the hapax forms culled from older materials require careful analysis in themselves, to decide their status and possible affiliation to already identified vocabulary.

What we do have is a corpus function that will give us real numbers of occurrences, with concordances and expanded text excerpts. A search argument like this:

"sus.*"

will produce a frequency sorted list of all word forms starting with *sus-* plus the two following words. It is a very useful function⁷, even if numbers are small:

sus i serken	16
sus og dus	14
suset frå pisserenna	10
sus i lufta	9
suste inn i	8
susar av garde	7

Figure 3: *Nynorskkorpuset* - Search result.

Our current solution for analysing data is a database, called “the sorter”. It is separate from, but linked to both MO and NO. In what it offers, it is a great deal less sophisticated than a lexical profiling tool (Atkins and Rundell, 2008: 91–92 and 107 f.), but is undergoing improvement. In the sorter, the editor generates a list of links to all instances linked to the MO entry, served up in a spreadsheet. The instances can be annotated and sorted, spread on several work sheets etc. The sorter has proved suitable as a note block for dealing with fringe materials (old, rare or poorly documented word forms). A sorter can have as many work sheets as the editor wants. The sorters are saved and stay linked to their entries. Sorters (with lists of instances) can also be moved to other entries, if materials are found to be misplaced.

⁷ The work of Dr. Daniel Ridings, who is in charge of *Nynorskkorpuset*.

Once sorted, documentation items can be linked directly to the relevant piece of information in the entry, be it dialect form, back up for definition or usage example (comprising both generic examples showing f.i. valence, and full citations). See Figure 4.

6. Generating entry heads from a separate full form register

In a dictionary offering information on spelling and inflection, entry heads traditionally present this information in a condensed form with extensive use of codes and abbreviations. Norwegian is a compounding language, as are most Germanic languages, with a medium rich inflection system⁸. In most Norwegian paper dictionaries compounds have no POS and inflection information since a compound has the same POS and inflection as the final part of the compound. It is assumed that all native Norwegian speakers can analyze compounds. This assumption has proved useful, given the space limitations of a printed dictionary. In an electronic dictionary space is not a problem – nor is it true that all Norwegian speakers can analyze compounds.

However, full inflection tables in the entry head as a first option are not a good idea. They should be shown on request.

The information on POS and inflection has to be accurate, complete and in accordance with school requirements. In the Nordic countries, publicly funded Language Councils are tasked with providing this mass of detail in a comprehensible fashion. Due to the complex spelling rules of Norwegian, with a large number of alternative forms and frequent spelling adjustments, this has been a daunting task. A complete, detailed overview of official standard Norwegian spelling (including all inflected forms) was a by product of the first edition of NOB and BOB. Today, a quality checked database, a word bank, exists for both written standards.

The Word Bank is based on an extension of a spellchecker made by IBM in the 1980s (Eng, 1993). The central idea is to link each lemma to one or more inflection patterns which in turn produce all possible forms. This process will cause the generation of possible but undocumented word forms. These forms are useful for the POS-tagger in which they are used, but not for human users. To avoid generating spurious forms and also to ensure that each set of inflected forms is in accordance with official orthography, additional information is added to the links between a lemma and inflection paradigms. For each link, validity level (unknown, variant form, norm) and the time span for this status, is listed.

⁸ Nouns for instance have four forms, eight if genitive forms are included: more than English, less than German.

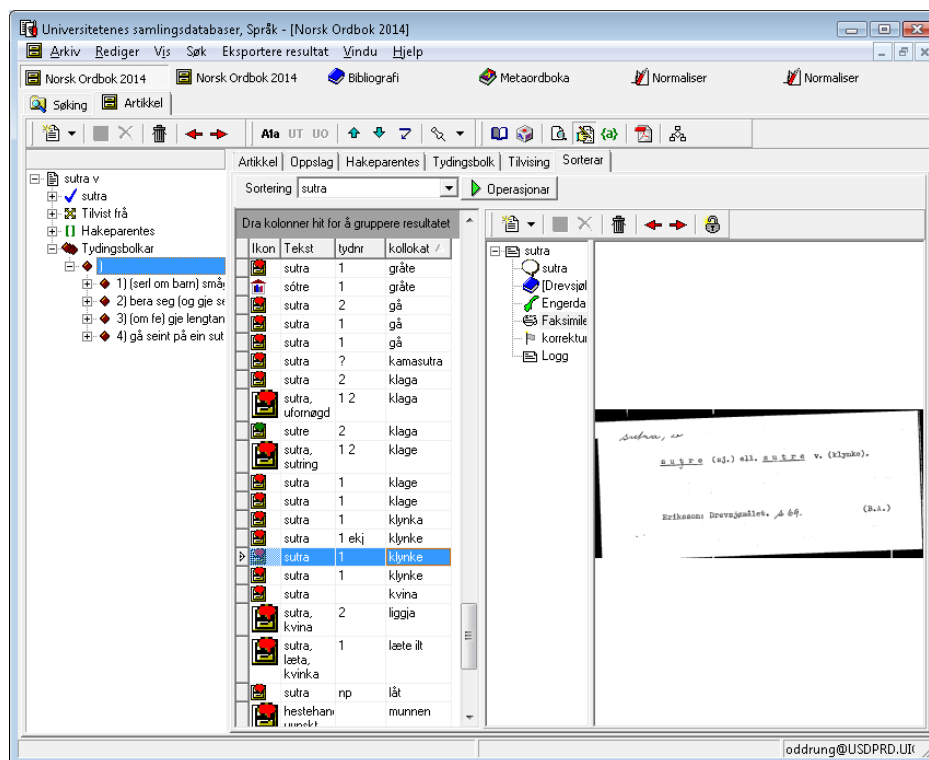


Figure 4: The sorter with list of sources, with entry in tree form to the left and image from slip archive to the right.

Currently, the Word Bank contains information for the time period 1996 to the present. It is possible to generate a valid orthography for any year in this period. An important feature of this system is that it can be used to “wash” lemma lists. The Word Bank has f.i. been used to check the Norwegian part of an Icelandic to Danish, Norwegian and Swedish web dictionary. This exercise turned out to be very useful.

The Word Bank can be used to generate the entry heads of a dictionary. No two Word Bank lemmata have the same set of paradigms and the same status history, but they are not separated with respect to homonyms beyond this point. Separate homographs have a strong tradition in Scandinavian lexicography. Thus one single lemma in the Word Bank may be linked to several lemmata in a dictionary.

Below, we show three examples of how POS and inflection was shown in a standard paper dictionary of Nynorsk from 2005:

rope v1 el. v2

II skru el. **II skrue** v1 el. *-r, -dde, -dd* el. *-tt* el. **II skruve** v1

I søkje el **søke** *-r, -kte, -kt*

Orthographic information in the form of codes and abbreviated forms is no longer acceptable in teaching, and the Web has freed the editors from the need to save space at every turn.

In the new web edition, all headwords of the two standard orthographic dictionaries BOB and NOB are linked to the entries in the Word Bank. The entry head of (web version) is now generated from the Word Bank, in schemas shaped according to school and Language Council requirements. The entry is shown with the headword followed by POS information. A click on the POS information opens a new window with a schema showing the inflection pattern(s) for the word in question (Figure 5).

rope **v1 v2 v3** (truleg frá ty jamfør norr *hrópa* 'baktale')
 bruke sterk røyst,; skrike, kalle;
 varsle med visse ord
rope om hjelp / rope hurra / rope på nokon
/ rope opp (namn, nummer på ei liste) / rope noko ut /
*som ein roper i skogen får ein svar, sjå **skog (1)***

Figure 5: NOB new website - the entry *rope* v with POS plus codes for inflection.

Bøying i samsvar med 2012-rettskrivinga:

rope	Infinitiv	Presens	Preteritum	Presens perfektum	Imperativ
v1	å ropa å rope	ropar	ropa	har ropa	rop
v2	å ropa å rope	roper	ropte	har ropt	rop
v3	å ropa å rope	ropar	ropte	har ropt	rop

rope	Perfektum partisipp				Presens partisipp
	Hankjønn/hokjønn	Inkjekjønn	Bunden form	Fleirtal	
v1	ropa	ropa	ropa	ropa	ropande
v2	ropt	ropt	ropte	ropte	ropande
v3	ropt	ropt	ropte	ropte	ropande

Figure 6: NOB - form showing inflection paradigms for *rope*.

This solution was launched last autumn and has proved a success with users.⁹ It is clear that it is complete for each lemma or lemma variant, and it encompasses the entire vocabulary in the dictionary in question. This solution for presenting inflection data can be implemented for any dictionary that is linked to the index MO. As a general feature this solution would be a great improvement for learner dictionaries on the Web.

⁹ The evidence for this statement is twofold: The feature is frequently used, and correspondence with users through Ordvakta

7. Writing dictionary entries, linking materials to each sense

Once the materials for a headword are analyzed, the entry gets written. The editorial interface shows the entries in three formats, (1) a tree structure (to the left), (2) a viewer showing the entry as xml text, and (3) a set of forms for editing the entry and managing the MO materials ('entry administration', 'entry head', 'form information', 'sense unit', 'cross reference' and 'sorter').

The sense unit form is where defining and entering usage examples happens. This form also has links to the bibliography and the location register, fields for cross referencing, etc. A particular feature is the compound table which allows editors to give instances of compounds where the sense shown in the definition is applicable. Compounds included in the compound table are linked to MO, which means that their usage is documented.

The sorter is linked to the entry and can be made searchable from the Web. However, it is also possible to link individual items of documentation directly to any node in the entry tree. In Figure 7 a link has been added to the synonym "drynja" (see arrow and boxes). A click on the "Belegg" icon leads straight to the image of the original slip. Currently these pointers to the material are mostly inserted for the benefit of colleagues, and typically added to convince doubters or as aids to the editors' memory. However, there is nothing to stop general access to such links.

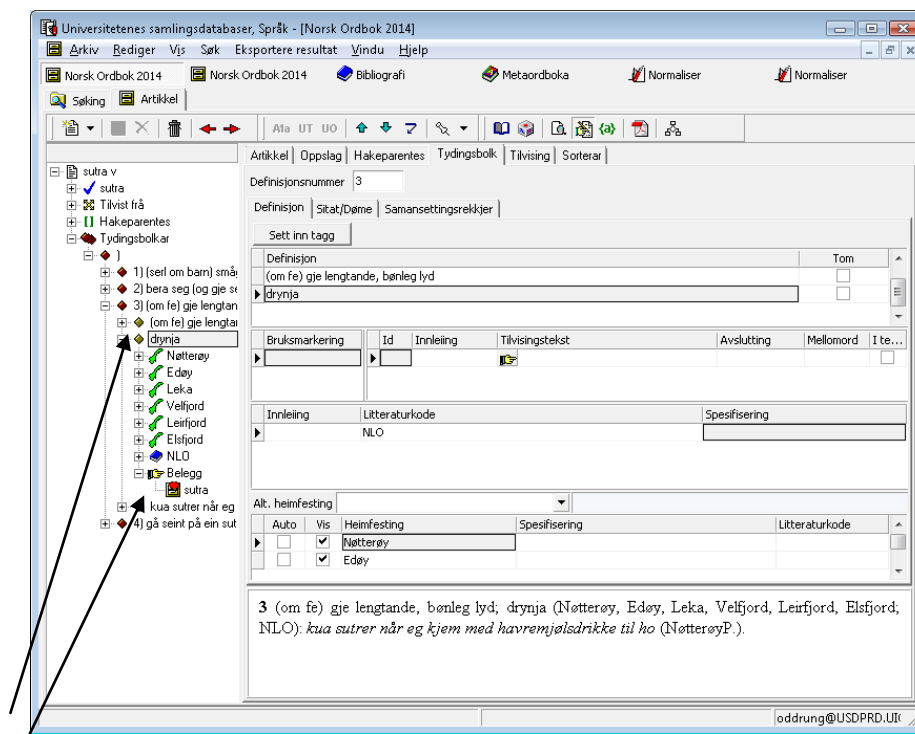


Figure 7. NO editor interface, sense unit form. Arrows show word in definition and its link to the source ("Belegg", red ikon).

Why is it important to have this possibility? Editing a historical dictionary based on materials from Norwegian dialects and Nynorsk, a written standard with a short history, is bound to cause discoveries that break with general preconceptions about language. We mention one in particular: Words associated with “slang”, “street language” and other frowned-upon innovations from young people in urban areas often turn out to be dialect variants of words well known in vernacular Norwegian from wide tracts of the country, or standard derivations from such words¹⁰. Some of them are attested back to Old Norse. When the hoodie turns out to be a preserver of old lexical items, one needs easy access to sources to be believed. Our experience in codifying languages with limited literary documentation and presenting them in dictionaries, has shown us that people very often believe their dialect forms to be unique to their own area. They never use these word forms away from home and will not be aware of their being part of the general vocabulary in the country. In such cases, easy access to documentation is essential.

8. Supervising production flow

Dictionary production is to a large extent a matter of managing time and money. There is no reason why a dictionary project should have poorer progress management than any other kind of project. For ease of administration, the system for supervising production flow is inbuilt in the database package set out in the introduction.

The management devices built into the administrative system is in part a result of what has been known to go wrong in previous large dictionary projects, partly a result of new possibilities when NO in 2003 moved to a digital platform. We will here comment on the management of size, status and storage.

The standard failing of older, paper-bound projects is that entries get longer and longer, and also take longer to produce, so that while manuscript production rockets, alphabet progression grinds to a halt. Our system for supervising size is therefore geared towards ensuring alphabet progression, and proper distribution of entry length within alphabet sections. Editorial work is measured in a given amount of finished manuscript per month. When an entry is generated, a maximum size is suggested, based on the amount of documentation available at the time of generation. Real size is measured against maximum size of the entry throughout editing. The editor can overrule the maximum size for individual entries, but the size of the alphabet section is fixed.

Data concerning production flow is shown in connection with each entry in the form “artikkel” (‘entry administration’). Figure 8 shows the subform dealing with size management, with the maximum number of lines and the present line count of edited text outlined.

¹⁰ Examples are verbs *loka* ‘hang (aimlessly) around’ and *kødda* ‘joke, “take the mickey”’.

All change in the dictionary database is logged with name, date and status change. The project management draws out reports every month to see manuscript progress, and while individual progress is always a matter between editor and management; the whole staff knows the exact state of progress per volume in moving manuscript along from draft through several control and correction stages to finished, publishable text. This supervision system combined with the possibility of generating a print version in PDF, promotes both efficiency and job satisfaction, since it is easy to see both from reports and from the dictionary database itself exactly how much one does. As work on NO also counts as scientific production for each editor in the University of Oslo crediting system, an exact count of lines and pages is very important.

The third point concerns the vulnerability of a project as large as NO, where one lost day means the loss of 1.5 man months, and where processed detail can be hard to recapitulate. Dictionary manuscript is stored in the database. Backups are taken every night, and stored. This ensures the project against production losses bigger than that of one working day, but it also means that it is possible to take care of the long version of an entry that needs to be shortened, or reinstatement entry that got deleted by mistake. The XML and HTML presentation of entries is synchronized with the editing. From the XML version, proofs with the correct typesetting are produced as PDF documents.

9. Presenting the finished product in an optimally accessible fashion

The dictionaries BOB and NOB have been searchable as a free web service since 1994. The website was thoroughly upgraded in 2009, with a view to making it visually appealing, especially for school use. The database solutions were thoroughly upgraded in 2012–2013. NO appeared on the Web in March 2012, as a by-product of the printed dictionary. This was possible on a tight budget because the databases have XML-presentation of entries built into the standard production format.

The finished product is the entry as it is presented on the Web, and web lay-out should be as clear as possible. This includes presenting the information most often sought up front, and hiding less popular items behind icons or codes. At the NO website, information on language variants is hidden behind a row of icons above the sense units. The dictionaries BOB and NOB are built on the language collections, but are not directly linked to them. Every entry is, however, directly linked to the Word Bank, and when users look up grammatical information, they are looking into the Word Bank full form lists for that particular lemma.

At present it is not possible to go directly to sources from the web presentation of NO. Source reference detail (bibliography, location) appears to be fixed to the right of the dictionary text.

This does not mean that the sources are inaccessible. In the case of NO, the language collections had been accessible on the Web for more than a decade before the dictionary itself appeared there, and links to the different sources are to be found on the home page of NO. The collections are well known amongst professional linguists and interested amateurs, and represent an important channel to public interest.

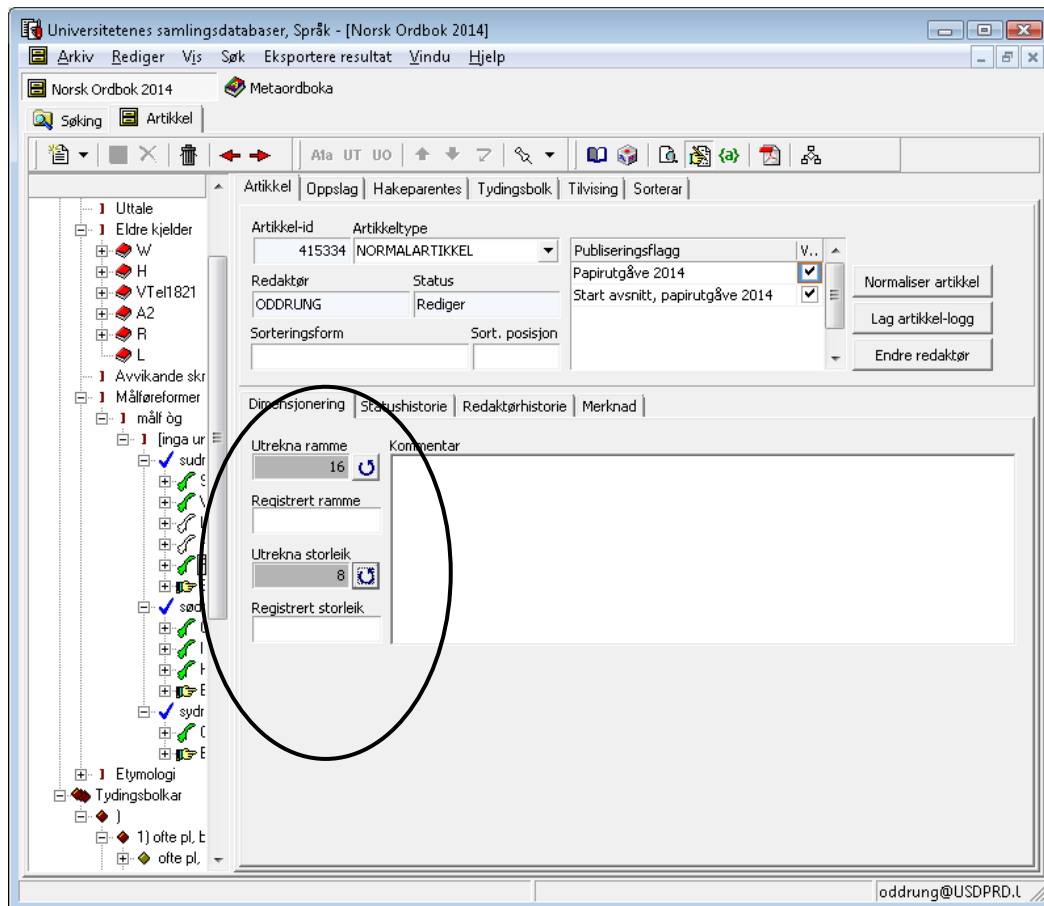


Figure 8: Administration page in the NO editing interface, with maximum number of lines allowed and the estimated number of lines required indicated, cf. section 8 above.

10. Simple search systems for complex databases

A great deal has been written about user-friendly access to web resources, including dictionaries and other language resources. A sort of scale seems to have emerged for solutions. At one extreme one finds the Google-type box where the user writes his search argument and then goes on to refine it, depending on the results. At the other end one finds solutions which require user profiling as a first step¹¹, so that the database can direct its user to the supposedly most relevant results. In between there are endless possibilities and combinations.

¹¹ Please note that this kind of user profiling is not the same as planning what sort of user one expects a dictionary to have, as in Atkins and Rundell, 2008 p. 486 f.

The user profiling approach has been promoted by Bergenholz and Tarp [see f.i. Fuertes-Olivera, 2009 p. 132 f.) in connection with their functional theory of lexicography. The idea is to create a lexicographical database as a multifunctional dictionary, with sophistication and detail in the entry increasing according to active choices in self profiling made by the user. They see the lexicographical database as a knowledge base containing multiple dictionaries from which virtual dictionaries, specialized according to the user's (self-described) profile and assumed needs, can be queried.

In the age of Google one may ask if this is a good idea. Our impression is that people tend to use Google and other search tools as data mining tools. A general search is iteratively narrowed until the required information is found. Under this assumption an electronic dictionary should be wide open to Google and other search engines. It is important that when a Google hit is clicked, the user reaches a web page which make the context clear and which offers the user a more detailed search in the dictionary.

On the other hand an electronic dictionary should offer its own search interface. We have seen that complex search forms scare away users. A simple search field should be standard. One can, however, include advanced search strategies in a simple field.

For the two standard monolingual dictionaries BOB and NOB a four step search strategy is implemented. First of all, an auto-complete function is attached to the search field. This gives a quick overview of possible headwords. Combined with wild-characters (truncated searches) this serves as an excellent tool for crossword and Scrabble. Multiword expressions (treated as sub-entries) are included in the headword search. If a headword search does not produce results, the search continues to the full form lists in the Word Bank. If there are no hits there, the search continues to the full text of the dictionary.

We think that queries going through several set stages could be useful in searching NO as well. One possible combination would be 1 headword field, 2 definition field, 3 usage example field (comprising both standardized examples and citations). Another possibility, for advanced searches, would be to extend the search to the source material linked through MO.

Active editors of the NO system have access to the whole of the category system in the linked databases, can put together their own searches, and store results as lists or export them as excel workbooks. The editors have had this possibility since 2003 and they use it actively in support of editorial work, or other information needs. However, this would be beyond the needs of the average dictionary user.

11. One database format - several dictionaries

The database system created for NO was in 2011–2012 utilized for the one volume standard dictionaries BOB and NOB, without any adaptations to the software. This

was not only possible, but completely painless, because the database for NO was created as a maximum format, catering for all the documentation and verification needs of a large academic dictionary with the task of working its way through heterogeneous language collections for the first time, and with a high academic standard to its referencing system, dealing with both written and spoken sources.

Before designing this maximum format the project tried going the other way, i.e. using and expanding existing software designed for a smaller dictionary. It didn't work because the framework was too cramped. We learnt from this experience for instance that speed in a very large and rich database system has to be planned for right from the start, as keeping the highways free is an important aspect of information architecture.

The NO database has four types of entry: standard, prefix, suffix, and cross reference. In addition there is an entry format for multiword expressions, for use within the standard entry. The smaller dictionaries did not have these types of entries, but they could be identified by text criteria (suffix entries having head words starting with a hyphen etc.).

The fact that the database system already had well defined, different formats for different types of entry, simplified the work with NOB and BOB. Two examples: (1) Affix entries do not have usage examples. What they do have are little lists of derived words demonstrating the use of the affix in question. Those derived words now exist in the dictionary database as a sort of minimal entry: they were picked out, got their full form entry in the Word Bank and are linked to the affix entry. (2) In between the usage examples of the NOB, there were also a number of multiword expressions masquerading as usage examples with a comment added. All usage examples with explanations attached were picked out and about 5000 selected for the multiword expression type of entry, with minimal textual adjustments.

12. Some comments on information architecture

When computers and ICT in general were introduced into lexicography several decades ago, computer specialists, as well as many lexicographers, started to talk about dictionaries as databases or knowledge systems. This is not really true, since dictionaries are written as structured texts for human users. Lexicographers used these terms metaphorically while the ICT-specialists saw the potential of extracting information into a relational database from what appeared to be highly structured texts.

The introduction of SGML and later XML technology represents a compromise. The use of XML in dictionary writing systems requires that every dictionary entry has to have a tree-like structure defined by a formal grammar. This is handy for most new dictionaries, but in order to fit older dictionaries into such a structure, a thorough

editing and restructuring of the text may be required. This fact was borne out by the revision process necessary in order to move NO on to a digital platform in 2003 (Grønvik, 2005).

It is often argued that the XML approach is superior to relational databases. This is in many ways a false debate. Most dictionary writing systems (DWS) are a mix. The entries are stored as XML-documents in a relational database and edited in an XML-editor. This gives flexibility, and it is easy to store many different dictionaries in a single system. XML is a format for manipulating and storing structured texts. It is not designed for active linked data. Thus, in the case of NO, where one has a set of heavily interlinked resources, the XML-approach is not sufficient. It is better and easier to decompose the entry text into a relational table structure to ensure data integrity. It is easy to produce XML from a relational database and in the versioning system the entries are stored as XML-documents. XML technology is also used for publishing PDF and HTML for the Web.

13. Conclusion

Everyone must be in favour of generic solutions for dictionary making, provided that the generic solution really covers every need. But a generic DWS must take into account the need to link dictionary text to sources through the database system itself. The need for control and verification is general, and in many cases essential, in showing that the dictionary really is the consensus product its editors set out to make it.

Once done, source linking is also very labour-saving. A click on the screen replaces a trip to the library or searching through archives and bookshelves. In Norway, the Word Bank is freely available for download. With a full form register and a truly generic DWS that can stay linked to its sources, many dictionary writers should find themselves in clover, and dictionary users will be able to see what their own dictionary is built on.

14. Acknowledgements

It should be emphasized that the very rich information architecture for lexicography described above has been shaped in response to input from a large working environment of lexicographers at the University of Oslo, and from important external users, all of whom are hereby thanked and acknowledged. All software development has been done by the Unit of Digital Documentation at the University of Oslo (EDD).

15. References

- Atkins, S.B.T. and Rundell, M (2008): *The Oxford Guide to Practical Lexicography* Oxford; Oxford University Press
- BOB = Wangensteen, B. et al. *Bokmålsordboka. Definisjons- og rettskrivingsordbok*. (1986-. 4. paper ed. 2006) Oslo; Universitetsforlaget. New web edition 2013. <http://www.nob-ordbok.uio.no/>
- Engh, J. (1993): Linguistic Normalisation in Language Industry. Some Normative and Descriptive Aspects of Dictionary Development. In: *Hermes, Journal of Linguistics*. no. 10 p. 53-64. <http://download2.hermes.asb.dk/>
- Fuertes-Olivera, P.A. & Bergenholtz, H. (red.) (2011): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum.
- Grønvik, O. (2005): Norsk Ordbok 2014 from manuscript to database - Standard Gains and Growing Pains. In Papers in *Computational Lexicography Complex 2005*. Budapest: Linguistics Institute, Hungarian Academy of Science. s. 60-70
- Menota guidelines for medieval Nordic texts* (www.menota.org).
- MO = *Metaordboka* (The Meta Dictionary)
<http://www.edd.uio.no/perl/search/search.cgi?tabid=571&appid=7>
- NO = Hellevik, A. et al. (eds) *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet* (1950-) Oslo; Det norske Samlaget (Web edition: <http://no2014.uio.no>)
- NOB = Hovdenak, M. et al. *Nynorskordboka. Definisjons- og rettskrivingsordbok*. (1986-. 4. paper ed. 2006.) Oslo; Det norske samlaget (New web edition 2012. <http://www.nob-ordbok.uio.no/>)
- Nynorskkorpuset*
http://www.muspro.uio.no/NO2014nynorskkorpus/conc_enkeltok.htm
- Ore, C.-E. Metaordboken - et rammeverk for Norsk Ordbok In *Nordiska studier i leksikografi 5. Rapport från Konferens om leksikografi i Norden, Göteborg 27-29 maj, 1999*. Göteborg: Nordiska föreningen för leksikografi.
- Ore, C.-E., Ore E., *Re-linking a Dictionary Universe or the Metadictionary Ten Years Later* Presentation at *Digital Humanities 2010*, King's College London, UK.
<http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-786.html>
- Schryver G.M. de (2011): *Why Opting for a Dedicated, Professional, Off-the-shelf Dictionary Writing System Matters*.
<http://tshwanedje.com/publications/asialex2011.PDF>
- Setelarkivet* (The Norwegian language collections - Nynorsk)
<http://www.edd.uio.no/perl/search/search.cgi?tabid=436&appid=8>
- TLex Lexicography and Terminology Software <http://tshwanedje.com/tshwanelex/>

The Woordenbank van de Nederlandse Dialecten (Wordbase of Dutch Dialects)

Jacques Van Keymeulen, Veronique De Tier

Ghent University
Jacques.vankeymeulen@ugent.be, Veronique.detier@ugent.be

Abstract

Three major regional dialect dictionaries, the *Dictionary of the Brabantic Dialects*, the *Dictionary of the Limburgian Dialects* and the *Dictionary of the Flemish Dialects* inventory the vocabulary of the southern Dutch dialects, i.e. the dialects spoken in Dutch-speaking Belgium, Northern Brabant and Limburg in the Netherlands and French-Flanders in France. The three dictionaries are onomasiologically arranged, according to the lexicographic ideas of A. Weijnen. Because of their arrangement, the dictionaries cannot convey detailed semantic information. They are to be considered atlases, rather than dictionaries. Therefore, in order to get a complete overview of the lexicon of the southern Dutch dialects, professional lexicography has to call in the help of ‘amateur’ lexicography, i.e. the regional and local dialect dictionaries, made by non-professional lexicographers.

In this presentation a project is described which aims at the creation of a digitized lexicographical database for the alphabetical amateur lexicography of the (southern) Dutch dialects, including both the old alphabetical tradition of the end of the 19th / beginning of the 20th century and the new tradition, rooted in the so-called dialect revival of the 70s and afterwards. The project is still in progress. First, we present the aim of the project; next we go into the details of the database structure and the search engines.

Keywords: database, dialect dictionaries, Dutch, Flanders

1. Introduction

The Dutch language has excellent digital dictionaries, thanks to the *Instituut voor Nederlandse Lexicologie*¹ (Institute for Dutch Lexicology) at Leiden University. The INL is financed by the *Nederlandse Taalunie* (Dutch Language Union), an organisation two-thirds subsidized by the Dutch government and one-third by the Flemish regional government. The main project of the INL was the *Woordenboek der Nederlandsche Taal* (WNT, Dictionary of the Dutch Language), a gigantic dictionary covering the modern Dutch period (i.e. 1500–1976) in 43 volumes (supplement included), totalling over 49,000 pages (95,000 main entries); it is the most comprehensive and detailed dictionary in the world. It has been digitized and is available in open access (see iWNT, the on-line *Woordenboek van de Nederlandse Taal* / Dictionary of the Dutch Language); in recent years other historical dictionaries have been linked to the WNT: the *Vroegmiddelnederlands Woordenboek* (VMNW,

¹ For more information on the INL, see <http://www.inl.nl>

Dictionary of Early Middle Dutch, 1999), the *Middelnerlands Woordenboek* (MNW, Middle Dutch Dictionary by Verwijs and Verdam) covering the 1200–1500 period and the *Oudnerlands Woordenboek* (ONW, Dictionary of Old Dutch, 2009). In 2011 the Frisian Dictionary was added. When the ANW² is completed, all historical periods of the Dutch language will be covered.³ All the aforementioned historical dictionaries are, for obvious reasons, based on written sources.

Many words, however, only survive orally in the traditional dialects. In this paper, we use the term *dialect* in its continental sense, i.e. as a term for a *geographically determined language variety*. We even use it in its narrowest – traditional and nowadays old-fashioned – sense: the geographically determined language variety typical for Trudgill’s ‘NORM’-informant, the Non-mobile Old Rural Male, a person who has as its urban counterpart the unskilled blue-collar factory worker. Traditional dialects are disappearing or have already disappeared, especially since the 60s of the last century, due to changes in the modern world: increase of social and geographical mobility, increase of schooling level and introduction of the mass media.

In what follows, we will discuss the dictionaries for the geographically differentiated vocabulary of traditional Dutch dialects. In section 2, we consider the dictionaries for oral language traditions, both the geographically oriented *onomasiological* ones (§2.1.) and the local/regional *semasiological* ones (§2.2.). It will be made clear that thematically arranged dictionaries should be supplemented with semasiologically arranged ones. The latter dictionaries should be brought together in a database: in section 3 an outline for such a database is presented with the project *Woordenbank van de Nederlandse Dialecten* (WND, Wordbase of the Dutch Dialects). Section 4 is devoted to the structure of the database; section 5 comprises the conclusion.

2. Oral language dictionaries: dialect lexicography⁴

Dialect vocabulary is mainly an oral vocabulary and it is geographically differentiated: both aspects should be documented. The first characteristic requests that the data are to be collected by way of fieldwork; the second that the fieldwork should be conducted in a place-to-place manner in order to be able to draw word maps. To this, one may add that the task is urgent. Since the 60s of the last century,

² Contemporary vocabulary (i.e. post 1970), finally, will be accounted for in the current INL-project *Algemeen Nederlands Woordenboek* (ANW).

³ With some reservations with regard to the 14th–15th centuries, which are in principle covered by the Middle Dutch Dictionary (MNW). The MNW, although it certainly is a major achievement, has some flaws due to its relatively restricted text basis when compared to both the chronologically preceding (= Early Middle Dutch Dictionary: 13th century) and following dictionary (Dictionary of the Dutch Language: 1500–1976)

⁴ For the most recent history of Dutch dialect lexicography see Goossens & Van Keymeulen (2006) and Taeldeman & Hinskens (*Language and Space: Dutch*, Mouton - De Gruyter, in press).

traditional dialects have been affected by both a functional and a structural reduction: they are spoken by ever fewer people in ever fewer situations. The vocabulary – the least stable language component – disappears first, due to the pressure of the standard Dutch language and the disappearance of the referents themselves, as is for instance the case for traditional agriculture, traditional crafts and trades and many aspects of modern life in general. A large part of the dialect lexicon has indeed already become a historical vocabulary.

Since the end of the 19th century, a considerable amount of work has been carried out in collecting dialect words, with different motivations: romantic and nostalgic feelings for the agrarian past; interest in historical linguistics and language reconstruction in the Neogrammarian paradigm; and recently an interest in cognitive semantics, etc. The linguistic interests of the public and scientists, of course, vary widely. Yet, it will be shown that in the case of lexical dialect research, a fruitful cooperation for both groups can be brought about.

2.1 Onomasiological dialect dictionaries and dialect atlases

Scientific Dutch dialect lexicography is not carried out at the Institute for Dutch Lexicology, as one might expect, but in fact began at the Catholic University of Nijmegen (today: Radboud University) by A.A. Weijnen, professor in Dutch, Indogermanic linguistics, Dialectology and Onomastics. Weijnen introduced the systematic, onomasiological arrangement of Dutch dialect lexicography. In doing so, he applied the lexicographical ideas of his predecessor Van Ginneken. According to Van Ginneken, the macrostructure of a dictionary should be presented in such a way that the clustering of semantically related words reveals the everyday life of the dialect speakers: this position resulted in a thematic arrangement. Next to the choice for a thematic arrangement, the lexicographical initiatives of Weijnen were also indebted to the dialectological paradigm of his day: i.e. word geography.

In the 1960s, Weijnen started the *Woordenboek van de Brabantse Dialecten* (WBD, Dictionary of the Brabantic Dialects) and the *Woordenboek van de Limburgse Dialecten* (WLD, Dictionary of the Limburg Dialects).⁵ In 1972, prof. W. Pée started the *Woordenboek van de Vlaamse Dialecten* (WVD, Dictionary of the Flemish Dialects),⁶ along the lines set out by Weijnen. The WBD was completed in 2005; the WLD in 2008. The Dictionary of the Flemish Dialects is still being compiled. The three dictionaries combined (and together with the ‘mappable’ WZD)⁷ cover the

⁵ In the 1990s an editorial board for the two dictionaries was opened at the KULEUVEN (Catholic University of Louvain).

⁶ The term ‘Flemish’ is often misunderstood. The Flemings have Dutch as their standard language. Flemish is used as a colloquial term for ‘Belgian Dutch’; in dialectology it denotes a dialect group in the west of Dutch-speaking Belgium.

⁷ The *Woordenboek der Zeeuwse Dialecten* (WZD, Dictionary of the Zeeland Dialects, 1964) by Mrs. H.C.A. Ghijsen is the first regional dictionary which gave detailed geographical information for every single dialect word.

whole of the southern Dutch language area, i.e. French-Flanders (France), Dutch-speaking Belgium, and the three southern provinces of the Netherlands (Zeeland, Northern Brabant and Limburg) (see Figure 1 below for the dialect landscape of the southern Dutch dialects).

All the dictionaries of Weijnen's school⁸ are arranged in the same way: every fascicle deals with a certain 'theme' (e.g. 'birds', 'the miller') pertaining to one of three parts: I. Agricultural Vocabulary; II. Technical and Crafts Vocabulary; III. General Vocabulary. Every fascicle is onomasiologically arranged and consists of a row of concepts, headed by a standard Dutch 'title', followed by a description and the heteronyms⁹ which can be used to refer to the concept. Every dialect word is followed by (general) indications as to frequency and location¹⁰ (see Figure 2 below for an example).

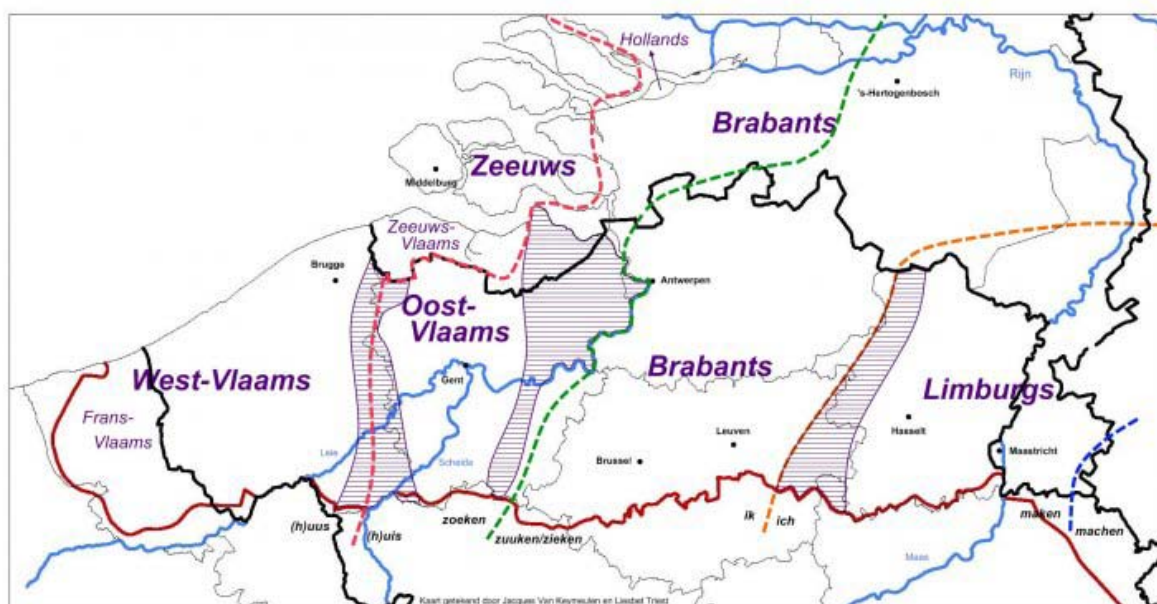


Figure 1: Dialect landscape of the southern Dutch dialects (according to Taeldeman, 2001: 8)

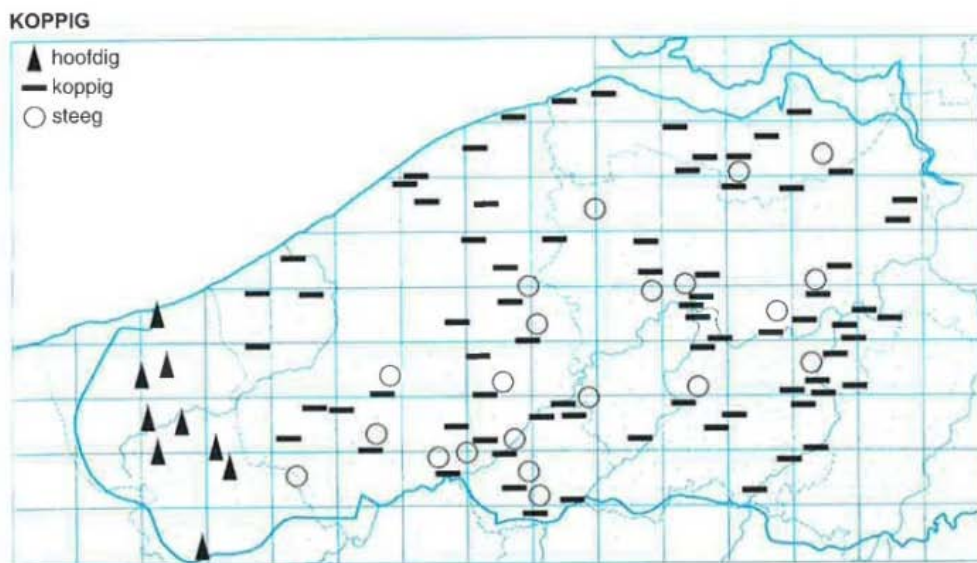
The printed texts of the three above-mentioned dictionaries for the southern Dutch dialects were based on automated databases from the late 80s onwards. For WBD and WLD the program Filemaker was used; for the WVD a specialized program (a relational database under Oracle) was developed by InfoService Belgium (ISB), a Ghent software company. The databases are provided with cartographic tools. The Filemaker database was already used for sophisticated geographical analyses (see

⁸ The WBD/WLD also served as an example for the WALD, WOD and WGD afterwards.

⁹ Weijnen coined the useful term *heteronym* for (dialect) words which mean the same in different (dialectal) language systems.

¹⁰ See Van Keymeulen & De Tier (2010) and Kruijssen & Van Keymeulen (1997) for more details.

Foldert de Vriendt, 2012 for the D² project). A combination of the databases¹¹ of the three dictionaries will hopefully start in the near future.



KOPPIG

Niet van eigen wil of inzichten af te brengen.

De Bo vermeldt voor 'koppigheid' **kopachtigheid** i.v. kop.

WVD 129 (2001), 75. ZND 1 (1923), 50e; ZND 28 (1938), 31; Willems (1886), 145.

bokkerig [*bukkerig*]: ♦Veurne.

bokkig [ook: *bukig*]: Lovendegem.

♦Schoonaarde. Wdb: Loquela: **bukig**, Kortemark.

dikkoppig: ♦Heist-aan-Zee.

driekante: ♦Beveren (IJzer).

eenwillig: Wdb: Lievevrouw-Coopman.

hoofdig: freq. FV.

♦spor. FV, ook Nazareth.

keikoppig: Roeselare.

kopachtig: Wdb: De Bo; Teirlinck: kopachteg i.v. kop.

koppig: freq. WV, ZV en OV.

♦alg. WV en OV, ook Steenvoorde, Aardenburg en Zuiddorpe. Wdb: Teirlinck: koppeg i.v. kop.

kopziende: Sint-Michiels.

obstinaat [ook: *opsternaat*]: Oostende, Kooigem, Nieuw-Namen.

Wdb: De Bo; WZD: opste(r)naot, opste(r)naet, obstinaot: WZV (Breskens, Groede, Retranchement), L. v. Hulst.

obstinatig [ook: *opstenadig*]: ♦Elsegem. Wdb: Desnerck: opsenoadig, opstenoadig.

opienig: Geraardsbergen.

Wdb: De Bo: opijnig; Teirlinck: opieneg i.v. opiene.

ossekoppig: ♦Heist-aan-Zee.

steeg: freq. WV zuidoost; spor. str. v. Ieper-Poperinge, ZV grensstr. en OV.

Wdb: Joos.

stijf: ♦Sint-Goriks-Oudenhove.

stijfhoofdig: ♦Wachtebeke.

stout: ♦spor. OV zuid, ook Koksijde.

KOPPIG PERSOON

Iemand die niet van eigen wil of inzicht af te brengen is, die tegen alle rede en goede raad in aan zijn eigen mening vasthoudt.

Voor een koppig kind geeft WZD **steenvreter** (*stêênfreter*) in WZV (Retranchement) en L. v. Axel (Axel, Hoek). De Bo

Figure 2: A page from WVD III, 4: 195 'stubborn'

¹¹ All three dictionaries have printed fascicles dating from the pre-computer era. Their data still have to be digitized.

The geographical aspect of the southern Dutch dialect landscape is covered by the three dictionaries of Weijnen's school. The semantic aspect of the dialect vocabulary, however, is poorly accounted for. Because of their onomasiological arrangement, the dictionaries cannot render detailed semantic descriptions. Overall, they have a very poor microstructure. Only in a dictionary where semantic detail (together with other microstructural information) is added to a headword (normally alphabetically arranged), is it possible to describe 'meaning', since in order to do so one needs to start from the lexeme (and not from the concept as is the case in onomasiological dictionaries).¹² Thus, the onomasiological dictionaries of Weijnen's school are to be considered as highly-structured geographical inventories of word usage: they are – notwithstanding the names of the publications – atlases, not dictionaries.¹³

2.2 Semasiological dialect dictionaries: two traditions

2.2.1 The old alphabetical tradition

The last half of the 19th century / first half of the 20th century witnessed some important works in the Netherlands, such as Molema (1887) for the province of Groningen and Boekenooen (1897) for the dialect of the Zaan district (North Holland). The major regional alphabetical dialect dictionaries in Dutch-speaking Belgium were compiled because of romantic motivations and the search for linguistic identity by the Flemings.¹⁴ De Bo (1873), for instance, collected West-Flemish words in order to contribute to a Flemish standard language of its own (a plan that did not work out). All those regional dictionaries, which sometimes cover whole provinces, have one important flaw: they do not exactly locate the dialect words. The geographical scope of the words is not indicated.

The above-mentioned dictionaries were not produced by an institute; hence we have called them 'amateur' dictionaries. This qualification perhaps does an injustice to the high scientific quality of many of them. The *Zuid-Oostvlaandersch Idioticon* (Southeast Flemish Dictionary) of Teirlinck (1908–1924), for instance, is a marvellous example of completeness and semantic detail. Some amateurs of the 'old school' certainly deserve admiration for the scientific excellence of their work.

2.2.2 Dialect 'revival' and amateurs

The standard Dutch language has spread to large sections of the population since the 1960s, hence using a dialect is nowadays regarded as a matter of 'choice' and not as a sign of backwardness. The new positive attitude towards these 'endangered' varieties, considered as 'immaterial heritage', has resulted in the production of amateur dialect dictionaries, meant for a local population.

¹² For a discussion of this matter: see Weijnen 1961, 1963, 1967 and De Tollenaere 1960, 1968.

¹³ The WZD proves that geography can have its proper place in an alphabetical dictionary, together with detailed information as to meaning.

¹⁴ The most important ones being: De Bo (1873), Joos (1900), Tuerlinckx (1886), Teirlinck (1908–1922), Rutten (1890), Cornelissen-Vervliet (1899–1903).

The table below (Table 1), taken from Oosterhof & Van Keymeulen (2009), visualizes the production of amateur dictionaries from 1835 until the beginning of the 20th century. The lexicographic effect of dialect revival becomes clear after the 1980s.

	FLANDERS					THE NETHERLANDS												
	LI	AN	VB	OV	WV	GR	FR	DR	OV	GD	FL	UT	NH	ZH	ZL	NB	LB	TOT
1835-39																1		1
1850-54																	1	1
1880-84																	1	1
1895-99									1									1
1900-04														1				1
1910-14																	1	1
1915-19																	1	1
1925-29															1		1	2
1935-39			1				1									1		3
1940-44										1								1
1945-49														1				1
1950-54				1														1
1955-59									1							1	1	3
1960-64																		0
1965-69							1		1					1			1	4
1970-74					1					1						1	2	5
1975-79			1	1			1			1	1					1	2	8
1980-84	3		1	1			1			5			1	2		2	3	19
1985-89	3	1	2	4	1					3				2		4	7	27
1990-94	3	1	4	3	2		1		2	5	2		1			8	9	41
1995-99	5	2	5	7	5		2	1	2	4		2	1	2		7	3	48
2000-04	6	2	2	3	6		2		5	3		1	4	2		8	8	52
2005-		1	1	4	1				1	3			1	4		7	11	34
TOT	20	7	17	24	16	0	9	1	13	26	3	3	8	15	1	41	52	256

Explanation of abbreviations: LI = Limburg (Belgium); AN = Antwerp; VB = Flemish Brabant; OV = East Flanders; WV = West Flanders; GR = Groningen; FR = Friesland; DR = Drenthe; OV = Overijssel; GD = Gelderland; FL = Flevoland; UT = Utrecht; NH = North Holland; ZH = South Holland; ZL = Zeeland; NB = North Brabant; LB = Limburg (the Netherlands)

Table 1: Local dialect dictionaries by lustrum and province (Van Keymeulen & Oosterhof, 2009)

Many things may go wrong when amateurs engage in writing dictionaries. There is indeed a huge variety, both in quantity (i.e. number of entries) and quality (notably of the semantic descriptions) between amateur dictionaries, and in the macro- and microstructural options. It goes without saying that the best of such dictionaries are made by amateurs with a linguistic schooling, although some authors with no training command admiration for their perseverance in detailed observation and the ensuing lexicographical result.

Nearly all amateur dictionaries copy, to a certain extent, the macro- and microstructural options of the standard language or bilingual dictionaries with which they are acquainted. Few amateurs know that there is such a thing as professional dialectology or professional lexicography. Nearly all local dictionaries are alphabetically arranged, mostly using a home-made dialect spelling for the headwords. Few amateurs are aware that this way of presenting the macrostructure of a local dialect dictionary will frustrate the intended user.

The interpretation of an entry word, rendered in a home-made dialect spelling, indeed presupposes a good knowledge of the dialect, on both lexical and phonological levels, in order to be able to look up a word. The form of the headwords is very often in blatant opposition to the needs of the user as envisaged in the introduction of many a dictionary: namely the user in a dialectless future. The example set by alphabetical standard language dictionaries seems to be very strong indeed, and inventing a spelling system for a dialect is a codifying activity with a high symbolic value.¹⁵ Many authors are inclined to appropriate the dialect by creating spelling for it. Members of the local population buy the dictionary not to use it, but to ‘possess’ it, as a symbol of their local identity.¹⁶

3. Towards a *Woordenbank van de Nederlandse Dialecten*¹⁷ (WND, Wordbase of Dutch Dialects)

A comprehensive collection of the dialect vocabulary of a vast area is a gigantic task, often beyond the financial reach of an institution. Although amateur lexicography obviously has its flaws, there are many amateur lexicographical products of good quality. They give at least a word form and an indication as to locality or region. Since they are alphabetically arranged, they are in principle able to describe meaning. Many of them also include collocations of all types: idiomatic expressions, proverbs, etc. Many contain example sentences in which meaning is illustrated. In short, they come as very welcome additions to the onomasiological dictionaries of the WVD-type, which are not capable of conveying semantic details and are altogether very poor in microstructure in general.

In September 2009, a pilot project¹⁸ was launched at Ghent University, funded by the Flemish Ministry of Culture, which envisages the creation of a digital database for

¹⁵ Van Keymeulen (1993) and Cajot (1995) are lexicographical manuals, meant for amateurs.

¹⁶ Delarue (2009) devoted a master thesis to the evaluation of a number of amateur dictionaries.

¹⁷ We thank J. Kruijssen for his suggestion to name the database *Woordenbank* instead of *Woordenboek*.

¹⁸ The first tentative results of the pilot study were reported in Van Keymeulen & De Tier (2010).

alphabetically arranged regional and local dialect dictionaries.¹⁹ The project consists of two phases: 1) creating a digital database on the basis of the dictionary texts; and 2) annotating the database in order to be able to perform efficient search operations. Software is being developed by the firm Info Service Belgium (Ghent). The overall purpose of the project is to combine the products of the ‘amateur’ dialect lexicography, and organize them in a digitized database in such a way that many types of automated searches can be done. In what follows, the present state of affairs of the *Woordenbank* (demo on www.woordenbank.be) will be presented. In 2012 the idea was copied by the Meertens Institute in Amsterdam; it was agreed that Ghent University would take care of the amateur lexicography in Dutch-speaking Belgium and the province of Zeeland; the Meertens Institute would collect the material of the dictionaries in The Netherlands.

3.1 The old alphabetical tradition

The dictionaries of the ‘old alphabetical tradition’ (end 19th / beginning 20th century) have a relatively bad printing quality, which creates problems for OCR-procedures. Much time is needed to correct the text files, and prepare texts for the input. Luckily, the project can rely on a number of volunteers to do the job.

Old dictionaries usually cover fairly large areas. Since they had to cope with a wide geographical differentiation in matters phonological, they were obliged to ‘lift’ the orthographical form of the headword to a relatively high level, i.e. they were obliged to normalize dialect word towards standard Dutch. In the West-Flemish dialects, for instance, standard Dutch *sch-* [sx-] as in *school* can be represented by [sk]-, [ʃ]-, [ʃx]-, [sʔ]. All this variation is implied and summarized in the normalized form <sch> in the headword *school* in the West-Vlaams Idioticon of De Bo (1873). The dialectal headwords are, so to speak, ‘dutchified’, i.e. written as if the word were a standard Dutch word. For the older dictionaries, the headwords should be modernized in spelling (e.g. *bosch* > *bos* ‘wood’; *keeren* > *keren* ‘turn’; *roozewied* > *rozewied* ‘cornflower’). Since this activity is relatively easy, it can be carried out by volunteers.

3.2 The dictionaries of the ‘dialect revival’

As explained above, a few hundred amateur dialect dictionaries have been written since the 1980s. Since the late 90s, many amateurs have created lexicographic databases, and have produced digital texts. Many of them were willing to contribute a copy of their final texts, and even volunteered to prepare them for input in the WND. In many cases, however, corrections of the proofreading still had to be transferred to the textfile. As the dictionaries were very often sold out in their printed form, most authors were willing to cooperate with the WND, without requesting financial or other compensation.

¹⁹ Financial support was given to the organization ‘Varieties vzw. Koepelorganisatie voor Dialecten en Oraal Erfgoed in Vlaanderen’, based at Ghent University.

Most amateur dictionaries are meant for a local population. Even in ‘local’ dictionaries, phonological (and even lexical) differentiation may occur. The dictionary of Cools (2000) on the dialect of Beveren-Waas (East Flanders), for instance, tries to account for seven former villages (Doel, Haasdonk, Kallo, Kieldrecht, Melsele, Verrebroek, Vrasene) which were joined with the town of Beveren in the 1970s. Because of this geographical scope, however small it may be, the author was forced to normalize the headwords towards standard Dutch. The word *lopen* ‘to run’, e.g., is pronounced in three different ways ([ly.°pm], [lu.°pm] and [li.°pm]) in the different villages. So as to avoid frustrating a dialect-speaking community, the ‘dutchified’ headword *lopen* is used, thus avoiding having to make an unwanted choice between the dialects. Since the normalizing of dialect words towards standard Dutch presupposes etymological insights, the dutchifications by amateurs without linguistical training cannot always be trusted. They should be evaluated and, if necessary, corrected before input into the WND.

Most dictionaries of the dialect revival have a very small geographical focus and use home-made spellings. Dutchification is applied on the basis of etymology and the correspondence rules between standard Dutch and the dialect. This is performed by volunteers, who receive in-job training, and is corrected afterwards by linguistically skilled persons.

A few examples of normalization / dutchification of headwords are:

dialect spelling	>	dutchification
<i>buttersjhuute</i> (‘butterfly’)	>	<i>boterschuit</i>
(Desnerck, 1972)		
<i>kraaisj</i> (‘cross’)	>	<i>kruis</i>
(Pletinckx, 2003)		
<i>trênink</i> (‘training’)	>	<i>training</i>
(Pletinckx, 2003)		
<i>ouverttoegd</i> (‘convinced’)	>	<i>overtuigd</i>
(Wellekens, 1994)		
<i>(h)euneenk</i> (‘honey’)	>	<i>honing</i>
(Pieters, 1995)		

The normalization of the original spelling of the dialect word is undoubtedly the most essential addition to the database.²⁰ In this way, the word collections will be opened up for non-dialect speakers and for scientific research. As a rule of thumb, the orthography of the headwords (and variants of them) in the WNT will be taken as a guide, since the WND will eventually be linked with iWNT in the Integrated Language Database.

²⁰ Normalization is not always easy, hence the database allows for the input of more than one normalized form.

At present, records in the database contain fields for: ‘original headword’, ‘dutchified headword’ and ‘search term in standard Dutch’ (boiling down to a one-word translation). It is evident that many more additions or refinements are possible: the different elements of the microstructure may be inputted (or added) into different fields, in order to facilitate many types of research. The fully-fledged enrichment of the database (including the addition of audio files) is postponed until a later phase of the project. The present record structure is as follows:

Original dictionary article:

Beddezièèkre, (nen) A.w.v. Paardebloem. (*Taraxacum officinale*). Samengesteldbloemigen met een gele bloem en gepluimde zaadjes. Men zegt ook wel Nen beddepissere.²¹

Field structure in the WND-database with annotations:

Beddezièèkre (= <i>original headword</i>)
beddezeiker, beddezeker (= <i>dutchification of the headword</i>)
paardenbloem (= <i>search term in standard Dutch</i>)

(example taken from Clinckemaillie, 1996)

4. Procedures

4.1 Software

The software of the Wordbase of Dutch Dialects operates under an Oracle platform. The software has been developed by the firm Info Service Belgium (Ghent). The database has been compiled from existing amateur dialect dictionaries on paper (see §2.). The first step is to digitize these dialect dictionaries by scanning and ocr-ing, if they are not available in a digitized version.²²

When a dictionary is scanned and ocr'd, a Word file is used as output. In most cases this text is not without mistakes, especially if the headwords are written in a self-made dialect spelling that is not recognizable by a computer. The Word files of these dictionaries need to be corrected, in our case by volunteers. At this moment a new campaign has started to increase the number of volunteers. Most volunteers are not skilled lexicographers. They are mostly male persons between 50 and 70 years of age, willing to do some cultural work.

²¹ Translation in English: ‘dandelion’.

²² The final text file version of a dictionary is normally in the editor’s possession (and not the author’s).

In order to import the Word file into the database, the text must comply with some standards. Each dictionary article has to begin with a bold headword (it being the original headword) and end with two hard returns. Once a Word file is corrected, and once the headword is put into bold, and two hard returns are inserted after each dictionary article, the file is ready for input in the database (see Figure 3).

Fleur(e)pietje [ˈflø:r(ə)pitʃə], zn. o. 1 Het allerbeste, allermooiste, allersterkste. - 2 Goede tol (speelgoed).

Fleures [ˈflø:rəs], zn. o. Longontsteking, pneumonie. Bij Kiliaan al: *pleuris/pleurisje*. Uit *pleuris/pleuritis* met wisseling van *pl/fl*, zoals in *flereijn*, Wvl. *prut/frut* ‘cichorei’, *Plutol/Flutol*, *perplex/perflex*. Ook *vliegend fleures*. Zie ook *waterfleures*.

Fliflouder, zie Fifouter.

Figure 3: Word file corrected by volunteers as input for the database
(example from Debrabandere, 1999)

4.2 Input in the database

The Word document is converted into a standard XML-file by means of a script. The XML-file can be imported into the database by means of a purpose-built application for this database. For some dictionaries, it is necessary to write an adapted custom script, which generates the standard XML-file for the application. These scripts will generate the XML-file by means of the already mentioned typographical conventions. Once the XML-files are uploaded, the database of the Wordbase of Dutch Dialects can be made. Each bold headword + the microstructure of it²³ will be put into the database of the Wordbase of Dutch Dialects as one entry (see Figures 4 and 5).

4.3 Annotation

The editors or volunteers (most of whom speak, or are acquainted with, the dialect of the specific dictionary) may then annotate the database with dutchifications, search terms in standard Dutch, comments and (later on) thematic markers and location (see the example in Figure 6). Sometimes it is possible to add some of these annotations automatically (e.g. when the headword is already a dutchification in the original work), but mostly the editor has to do this part half automatically or manually.

Because volunteers are helping to add the dutchifications and the search terms in standard Dutch, we decided two years ago to introduce a second, easier, way to put the data into the wordbase. While volunteers are correcting the ocr'd text, they can

²³ Different dictionaries vary widely as to their microstructure (semantic definitions, phonetic notation, grammatical information, example sentences ...).

already insert the dutchifications and the search terms in the Word file itself. This is performed by putting codes at the end of the dictionary article. The dutchification is placed between \$\$ \$\$; the search term is put between ££ ££ (see Figure 4).

<p>Fleur(e)pietje [ˈflø:r(ə)pitʃə], zn. o. 1 Het allerbeste, allermooiste, allersterkste. - 2 Goede tol (speelgoed). \$\$fleurepietje, fleurpietje\$\$ ££allerbeste; tol££</p> <p>Fleures [ˈflø:rəs], zn. o. Longontsteking, pneumonie. Bij Kiliaan al: <i>pleuris/pleurisje</i>. Uit <i>pleuris/pleuritis</i> met wisseling van <i>pl/fl</i>, zoals in <i>flerecijn</i>, Wvl. <i>prut/frut</i> ‘cichorei’, <i>Plutol/Flutol</i>, <i>perplex/perflex</i>. Ook <i>vliegend fleures</i>. Zie ook <i>waterfleures</i>. \$\$fleuris\$\$ ££longontsteking££</p>
--

Figure 4: Word file corrected and annotated by volunteers as input for the database (example from Debrabandere, 1999)

	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fientig	fijntig	tenger, frêle
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fiesshow	fichau, fisjauw	bunzing; sluwe rappe persoon
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fiester'n	fisteren	zeuren, zaniken
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fietjefatjerieë	fitjefatjerie	waardeloze spullen; charcuterie
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fiette	fietje	sofietje; dommevrouw; ziezo
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fikke	fikje	klein, levendig kind/persoon
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fikkel'n	fikkelen	onhandig werken/snijden
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	fleet'aor	fleet haar	dun sluijk haar
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	flêitre	fleiter	oorveeg
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	flîèëws	fleeuws	flauwe fietse smaak
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	flieflotte	flieflotter	vlinder
	Het dialect van Midden-West-Vlaanderen - J. Clinckemaille Import: F-G (421)	flikk'n	flikken	werkje opknappen; bedriegen

- 1) ▲ button to visualize the lemma ▲ 2) importfile ▲ 3) original headword ▲ 5) search term
 ▲ 4) dutchification

Figure 5: List of lemmata uploaded in the database (by clicking the button on the left, one accesses the specific dictionary article) (example from Clinckemaille, 1996)

Macros are employed to change typical dialect or orthographic characteristics in order to make the adding of dutchifications and search terms a bit more comfortable for the volunteers. In this way, it is possible to suggest dutchifications or search terms. The volunteer knows what alterations are made by means of the macro and when correcting the text file and reading the annotations, he may correct and complete suggestions simultaneously. These macros are of course not always the same for each dictionary, so for each new dictionary the macrostructure has to be analyzed.

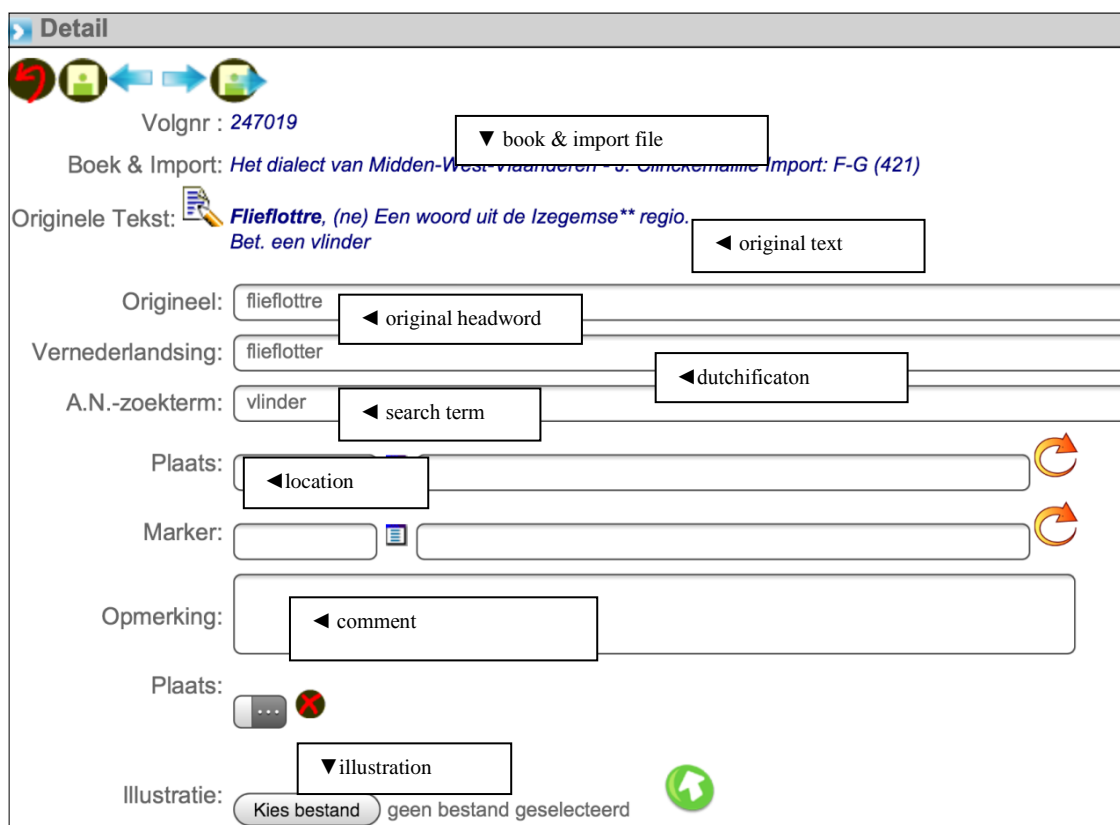


Figure 6: Example of the lemma *flieflottré* ('butterfly') (*flieflottré* = original headword) in the Wordbase of Dutch Dialects, annotated with dutchification (= vernederlandsing) (*flieflotter*) and search term (= A.N.-zoekterm) (*vlinder*) (example from Clinckemaille, 1996)

This procedure is easier for volunteers, because they do not have to use the online database for correction. The work can thus be performed on a stand-alone computer or without internet. The input text for the computer program is the corrected Word file, with the addition of dutchification and search terms between codes (see Figure 4). This coded text is then converted into a standard XML file by means of a script, and this file can be imported into the database by a purpose-built application. The fields with dutchification and search terms will therefore already feature in the database and do not need to be added online after the import.

4.4 Front and back matter

Both the front matter (such as copyright page, imprint, foreword, list of abbreviations, user's guide etc.) and the back matter (lists, index, etc.) of each dictionary are incorporated in the database as pdf-files. Next to that, the bibliographical reference is added (together with a pdf for title page and cover) and the state of affairs with regard to correction and annotations. The illustrations are imported either at the level of the dictionary article, or at a higher level. Addition of sound samples is not yet possible (partly because of the high time investment it would require).

4.5 Website

The database is connected to a website with search facilities and the following search terms: 'original headword', 'dutchification', 'search term in standard Dutch' and 'word in article'. The dutchifications and general search term 'word in article' really disclose the word collection as a whole. In each search facility, one can use wildcards.

The list of dictionaries is growing. Most users are looking for a particular word in one region. That is why there is also an option to search in all or some dictionaries or in one dictionary. One can also choose the dialect dictionaries of one province.

Recently a contact form has been added. With this form, users can let us know if they see mistakes, or if they want to give a new explanation which is not yet in the dictionary. In the next addition we intend to visualize a comment-line. Therefore, when there are obvious mistakes or questions, when an author wishes to add something, or when the editors wish to provide additional information, adding and visualizing these comments will be possible. Currently, the database itself includes a comment possibility. This is intended for the volunteers to offer questions or comments. The comment-field is not displayed on the website. A list of the dictionaries that are completely or partially annotated and imported is also available. Under the information button, one can find information on the state of affairs for each dictionary.

4.6 Preliminary results

The search facilities in the database are already operational although the dictionaries in the database have not been completely imported and annotated as yet. The search facilities still have to be evaluated and maybe adapted to the needs of the user. To evaluate the usability, it would be good to start with a selected user group in the near future.

The import of some dictionaries has been completed (e.g. the *Kortrijks Woordenboek* of Debrabandere, 1999). In the case of other dictionaries, it remains to import the front and back matter and the annotations. We rely heavily on volunteers, but hope to be able to present a database within a few years which will contain at least 30 dictionaries. The total number of entries (dictionary articles from all dictionaries incorporated in July 2013) is 226,788. Twenty-six dictionaries have been (partly or completely) imported in the database, but the regional distribution has to be improved in the near future.

The demo version of the website with the first results can be seen on www.woordenbank.be (see a list of search results in Figure 7).

Origineel trefwoord: flieflo(e)ttre **Vernederlandsing:** fliefloeter, flieflotter, flieflouter **A.N.-zoekterm:** vlinder; losbol
 flieflo(e)ttre (ne) : (1) vlinder, (2) losbollig meisje met de ene vrijer na de andere

Debrabandere, F. (1999). *Kortrijks Woordenboek*. Kortrijk: De Ieiegouw; Brugge: Van de Wiele, 553 p. 

Origineel trefwoord: fifouter **Vernederlandsing:** fifouter **A.N.-zoekterm:** vlinder; vleermuisbrander


Fifouter ['fɛf(l)otr / 'fɛflotr], zn. m. 1 Vlinder. Het dialectwoord kan op een lange geschiedenis bogen, maar gaat snel achteruit. Ohd. *fifaltra*, Os. *fifoldara*, Mnl. *viveltere*, *vivoudere*, D. *Folter*. Lit.: J.L. PAUWELS, *De vlinder. Hand. KCTD 9* (1935), 329-382. - 2 Vroeger ook wel: gasbrander zonder kous, vertaald uit Fr. *bec à papillon* : vleermuisbrander.



Origineel trefwoord: fliflouter **Vernederlandsing:** fliflouter **A.N.-zoekterm:** vlinder
 Fliflouter, zie Fifouter.

Pieters, M. (1995, 2008²). *Woordenboek van het Lokers Dialect*. Lokeren: Oelbrandt, 487 p. 

Origineel trefwoord: bottevijvere **Vernederlandsing:** botvijver **A.N.-zoekterm:** vlinder
 bottevijvere: vlinder, schubvleugelig insect (orde der Lepidoptera). Ook *mottevijvere*.

Taeldeman, J. (2011). *Woordenboek van de Oosterzeelse Dialecten*. Gent: Neveland Graphics, 300 p. 

Origineel trefwoord: kuuëlwitse **Vernederlandsing:** koolwitje **A.N.-zoekterm:** koolwitje, vlinder
 kuuëlwitse < zn., o > koolwitje

Clinckemaillie, J. (1996). "Ool koett'n en ool doen". *Het dialect van Midden-West-Vlaanderen*. Aartrijke: Uitgeverij Emiel Decock, 206 p. 

Origineel trefwoord: flieflottre **Vernederlandsing:** flieflotter **A.N.-zoekterm:** vlinder
 Flieflottre, (ne) Een woord uit de Izegemse** regio. Bet. een vlinder

Figure 7: Search results for search term = vlinder ('butterfly') in the Wordbase of Dutch Dialects

5. Conclusion

Compilation of the *Woordenbank van de Nederlandse Dialecten* is under way, thanks to volunteers, and coordinated by Veronique De Tier at Ghent University. Mrs Silvia Weusten collaborated for approximately one year, especially for Limburg dictionaries (and the Ghent dictionary). The workload is indeed gigantic, because of the lexical and semantic richness of the dialects and because of the sheer number of good dictionaries. Thanks to the different search terms, both the semantic and encyclopedic information of the dictionaries are easily accessible. The results of a search term in the microstructure reveal not only lexicographic data, but also show that the *Woordenbank* can be used as an ethnographic database.

6. References

- Boekenoogen, G.J. (1897). *De Zaaansche volkstaal. Bijdrage tot de kennis van de woordenschat in Noord-Holland*. Leiden: Sijthoff.
- Cajot, J. (1995). *Hoe maak ik een dialectwoordenboek. Een handleiding voor Limburgers en anderen die dialectwoorden willen spellen, verzamelen en beschrijven*. Mededelingen van de Vereniging voor Limburgse Dialect- en Naamkunde nr. 78/79. Hasselt: Vereniging voor Limburgse Dialect- en Naamkunde.
- Clinckemaillie, J. (1996). *Ool koett'n en ool doen. Het dialect van Midden-West-Vlaanderen*. Aartrijke: Decock.
- Cornelissen, P.J. & Vervliet J.B. (1899-1903). *Idioticon van het Antwerpsch dialect (stad Antwerpen en Antwerpsche Kempen)*. Gent: Siffer.
- Debrabandere, F. (1999). *Kortrijks Woordenboek*. Kortrijk/Brugge: De Leiegouw/Uitgeverij Van de Wiele.
- De Bo, L. (1873). *Westvlaamsch Idioticon*. Brugge: Gailliard.
- Delarue, S. (2009). *Van woordenlijst tot woordenboek. Lexicografische beschrijving en evaluatie van de Vlaamse amateurdialectlexicografie in de 19de, 20ste en 21ste eeuw*. (unpublished master thesis, Ghent University).
- De Tier, V. & Van Keymeulen, J. (2010). Software demonstration of the dictionary of the Flemish Dialects and the Pilot Project Dictionary of the Dutch Dialects. In: Dykstra, A & T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress*. Fryske Akademy, Leeuwarden. 620-627 (issued on CD-ROM).
- De Tollenaere, F. (1960). *Alfabetische of ideologische lexicografie? Bijdragen tot de Nederlandse Taal- en Letterkunde*. Leiden: Uitgegeven vanwege de Maatschappij der Nederlandse Letterkunde I.
- De Tollenaere, F. (1968). Problemen van het dialectwoordenboek. Theorie en praktijk. In: *Tijdschrift voor Nederlandse Taal- en Letterkunde* 84. 197-212.
- De Vriendt, F. (2012). *Tools for computational analyses of dialect geography data*. Radboud University Nijmegen (Ph.D. dissertation).
- Frisian Dictionary = *Woordenboek van de Friese Taal / Wurdboek fan de Fryske Taal*. Leeuwarden : Fryske Akademie.
- Goossens, J. & Van Keymeulen J. (2006). De geschiedenis van de Nederlandse dialectstudie. In: *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 78. 37-97.
- Joos, A. (1900). *Waasch idioticon*. Gent/St.-Niklaas: Siffer/Strijbol.
- Kruijssen J. & J. Van Keymeulen (1997). The Southern Dutch Dialect Dictionaries. In: *Lexikos* 7. 207-228.
- Molema, H. (1887). *Woordenboek der Groningsche volkstaal*. Winsum: Mekel.
- Pieters, M. (1995), *Woordenboek van het Lokers Dialect*. Lokeren: Uitgeverij Oelbrandt.
- Pletinckx, L. (2003). *Woordenboek van het Asses. Bijdrage tot de studie van de West-Brabantse streektaal*. Asse: Koninklijke Heemkring Ascania.

- Rutten, A. (1890), *Bijdrage tot een Haspengouwsch Idioticon*. Antwerpen, Zuidnederlandsche Maatschappij van Taalkunde; Boucherij.
- Rys K. & Van Keymeulen J. (2009). Intersystemic correspondence rules and headwords in Dutch dialect lexicography. In: *International Journal of Lexicography* 22. 129-150.
- Taeldeman, J. (2001). De regenboog van de Vlaamse dialecten. In: Devos, M., J. De Caluwe, & J. Taeldeman (eds.). *Het taallandschap in Vlaanderen*. Wetenschappelijke Nascholing UGent.
- Taeldeman, J. & F. Hinskens (in press). *Language and space: Dutch*. Berlin : Mouton - De Gruyter.
- Teirlinck, I. (1908-1924). *Zuid-Oostvlaandersch Idioticon*. Gent: Siffer.
- Tuerlinckx, J.F. (1886). *Bijdrage tot een Hagelandsch Idioticon*. Gent: Zuidnederlandsche Maatschappij van Taalkunde.
- Van Dale = *Van Dale Groot Woordenboek van de Nederlandse Taal* (2005).
- Van Keymeulen, J. (2003). Dialectwoorden verzamelen. Een praktische handleiding. In: *Handelingen van de Koninklijke Commissie voor Toponymie en Dialectologie* 75. 383-506.
- Van Keymeulen, J. (2003). Compiling a dictionary of an unwritten language. A non corpus-based approach. In: *Lexikos* 13. 183-205.
- Van Keymeulen, J. (2004). Trefwoorden en lexicale varianten in de grote regionale dialectwoordenboeken van het zuidelijke Nederlands (WBD, WLD, WVD). In De Caluwe J., G. De Schutter, M. Devos & J. Van Keymeulen (eds.). *Taeldeman, Man van de Taal, Schatbewaarder van de Taal*. Gent: Vakgroep Nederlandse Taalkunde / Academia Press. 897-908.
- Van Keymeulen, J. (2009). Volkslinguïstiek en dialectlexicografie in de zuidelijke Nederlanden. In: *Lexikos* 19. 314-339.
- Van Keymeulen, J. & A. Oosterhof (2009). Local dialect dictionaries and a proposal for a dictionary of the Dutch dialects. In: Gooskens, C., A. Lenz & S. Reker (eds.). *Low Saxonian dialects across borders: Synchrony and diachrony*. Stuttgart: Franz Steiner. 109-124.
- Van Keymeulen, J. & V. De Tier (2010). Pilot Project: A Dictionary of the Dutch Dialects. In: Dykstra, A. & T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress*. Fryske Akademy, Leeuwarden. 754-763 (issued on CD-ROM).
- Van Keymeulen, J. & V. De Tier (2010). Towards the completion of the Dictionary of the Flemish Dialects. In: Dykstra, A. & T. Schoonheim (eds.), *Proceedings of the XIV Euralex International Congress*. Fryske Akademy, Leeuwarden. 764-773. (issued on CD-ROM).
- WALD = Schaars, L. (1984 -). *Woordenboek van de Achterhoekse en Liemerse Dialecten*. Doetinchem: Staringinstituut.
- WBD = Weijnen, A. e.a. (1967-2005). *Woordenboek van de Brabantse Dialecten*. Assen/Maastricht: Van Gorcum; Groningen/Utrecht: Gopher.
- Weijnen, A. (1961). De semantische en syntactische problematiek van het dialectwoordenboek. In: *Tijdschrift voor Taal- en Letterkunde* 78. 81-95.

- Weijnen, A. (1963). Het dialectwoordenboek. In: *Woordenboek en Dialect. Bijdragen en Mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam 24*. 34-52.
- Weijnen, A. (1967). De waarde van een dialectwoordenboek. In: *Mededelingen van de Nijmeegse Centrale voor Dialect- en Naamkunde 6*. 5-11.
- Wellekens, W. (1994), *Diksjonêr van 't Leives*. Academie voor het Leuvens Dialect: Leuven.
- WGD = Giesbers, C./Scholtmeijer, H. (2005 -). *Woordenboek van de Gelderse Dialecten*. Utrecht: Matrijs.
- WLD = Weijnen, A., J. Goossens e.a., (1983-2008). *Woordenboek van de Limburgse Dialecten*. Assen/Maastricht: Van Gorcum; Groningen/Utrecht: Gopher.
- WOD = Scholtmeijer, H. (2000-). *Woordenboek van de Overijsselse Dialecten*. Kampen: Stichting IJsselacademie.
- WVD = Devos, M. e.a. (1972-). *Woordenboek van de Vlaamse Dialecten*. Gent-Tongeren: Michiels.
- WVD III, 4 = De Pauw, T. & M. Devos (2005). *Woordenboek van de Vlaamse Dialecten. Algemene Woordenschat. Aflevering 4: Karakter*. Gent-Tongeren: Michiels.
- WZD = Ghijsen, H.C.M. (1964). *Woordenboek der Zeeuwse Dialecten*. Den Haag, Van Goor. [Fraanje, K. e.a. (2003), *Supplement Woordenboek der Zeeuwse Dialecten*. Krabbendijke: Van Velzen]
- WNT = *Woordenboek der Nederlandsche Taal* (1864-1998). See: <http://gtb.inl.nl/?owner>

Mining a parallel corpus for automatic generation of Estonian grammar exercises

Antoine Chalvin, Egle Eensoo, François Stuck

Institut national des langues et civilisations orientales (INALCO)

65 rue des Grands-Moulins, 75013 Paris, France

E-mail: antoine.chalvin@inalco.fr, egle.eensoo@inalco.fr, francois.stuck@inalco.fr

Abstract

The aim of our research is to develop a system to generate Estonian grammar exercises for French-speaking learners, based on a large lemmatised parallel corpus (<http://corpus.estfra.ee>) and on the data of the Comprehensive French–Estonian Dictionary (<http://www.estfra.ee>). We concentrate on exercises on nominal and verbal morphology. Although the corpus is not syntactically tagged, we also explore the possibilities of generating some types of syntax exercises. The system generates on demand exercises consisting of a specified number of Estonian sentences, in which relevant word forms are replaced by their lemmas. The learner has to construct the right form and can check his or her answers. Sentences are accompanied by their French translation. In this article, we concentrate on the problems related to the definition and tuning of sentence selection criteria. Exercises can be generated at three levels of difficulty. Relevant sentences are picked up in the corpus according to their length and the “frequency” of the lemmas they contain, i.e. the presence of the lemmas in one of the four subsets of headwords specified in the data of the dictionary: basic vocabulary (4000 words), small dictionary (10 000 words), lower-medium dictionary (15 000 words), and upper-medium dictionary (40 000 words).

Keywords: parallel corpora; readability; e-learning; Estonian as a foreign language; grammar exercises

1. Background and objectives

Since the 1990s there has been a growing interest in using corpora for language learning purposes (see Boulton, 2008; Huang, 2011). One of the key approaches in this field is ‘data-driven learning’ (DDL), which has been described as an “attempt to cut out the middleman” and to give the learners “direct access to the data” (Johns 1994: 297). In practice, the DDL, which focuses on the use of corpus concordances in the classroom, still supposes the guidance of a teacher. A more effective way to really “cut out the middleman” is to develop systems that use corpora as a source to generate self-correcting tests. An impressive number of test generation systems have been developed in the field of EFL (English as a Foreign Language), mainly to generate vocabulary tests in multiple-choice format (e.g. Coniam, 1997; Gao, 2000; Mitkov & Ha, 2003; Hoshino & Nakagawa, 2005; Brown et al., 2005; Liu et al., 2005; Sumita et al., 2005; Kilgarriff et al., 2010), and more rarely grammar tests (Chen et al., 2006; Lee & Seneff, 2007; Hoshino & Nakagawa, 2008). For French, the GramEx system developed by Beltrachini, Gardent & Kriszewski (2012) is not based on corpora, but on a grammar-based sentence generation process.

The aim of our project is to develop a system to automatically generate fill-in-the-blank Estonian grammar exercises consisting of authentic sentences. Fill-in-the-blank exercises are widely used in foreign language learning to help build grammar proficiency. One of their drawbacks is that they usually consist of specially designed sentences, which do not necessarily reflect real language use. The other drawback of manually designed exercises is that, since their creation is very time-consuming, textbooks and learning environments usually propose a limited number of them, which is not sufficient for the learner to acquire full proficiency on the specific points dealt with in the exercises. Our idea is that the automatic generation of exercises from a corpus of authentic language material could remedy these drawbacks and offer the learner the possibility to continue building his/her grammatical proficiency after he/she has completed all the exercises in his/her textbook. The system we want to develop is thus conceived as complementary to traditional language learning materials. It may address the needs of elementary, intermediate or advanced learners, but probably not those of complete beginners. Its implementation is complicated by a number of difficulties related to the quality of the corpus and the definition of complexity (readability) criteria for sentence selection. Our main concern, in the first stage of the project, is not so much pedagogical as computational: we want to determine how to process a large corpus of real unmodified texts in order to make it a suitable source for generating L2 grammar exercises. In other words: how to extract from a general language corpus a specific subcorpus more fitted to the needs of foreign language learning? And what kind of grammar exercises is it possible to create on the basis of a morphologically tagged corpus?

2. The Estonian-French parallel corpus

Our system is based on the Estonian-French parallel corpus (CoPEF: <http://corpus.estfra.ee>) compiled by the French-Estonian Lexicography Association (Prantsuse-eesti leksikograafiaühing, Tallinn). The corpus was designed primarily to address the needs of lexicographers compiling a comprehensive Estonian-French dictionary of 110 000 entries (GDEF: <http://www.estfra.ee>). Considering this specific purpose and the relatively limited number of available bilingual texts, the main principle followed in the compilation of the corpus was to attain the critical mass needed for lexicographical work, and not to produce a balanced corpus. The whole corpus contains 65 million words and is subdivided into seven subcorpora:

- Estonian literature (3.85 million words),
- French literature (4.09 million words),
- Estonian non-fiction (132 000 words),

- French non-fiction (990 000 words),
- European Union legislative texts (26.3 million words),
- Debates of the European Parliament (28.2 million words),
- Bible (1.4 million words).

The corpus is lemmatised and morphologically tagged. Estonian texts were tagged with Estmorf (cf. Kaalep 1996, 1998) and disambiguated with Tahmm (Tahmm, 1998). But the result is not 100% reliable. Tahmm does not always choose the right variant. In some cases it is not able to disambiguate and results in several variants. This occurs especially when the variants refer to the same grammatical form and differ only in their lemmas (Tahmm, 1998). Potential mistakes in morphological analysis will have to be taken into account when designing the exercises. In order to reduce their impact, it is necessary to avoid exercises based exclusively on specific forms that Tahmm has difficulty identifying. For example, we will not propose specific exercises on the formation of singular genitive, because some of the “genitive” forms that the learner would have to build could be in fact the singular partitive or singular nominative of the same word (homography between these three forms is quite frequent). We can propose instead more global exercises on nominal morphology, including genitive and partitive forms, but without specifying which of these cases is concerned in each question.

Sentence-level alignment of the corpus was made at different periods with different tools, either automatically (for EU texts) or semi-automatically (for other subcorpora). In the latter case, alignments with a low probability index were controlled and corrected manually. A few literary texts were aligned fully manually. The reliability of alignments was not precisely estimated, but there are obviously mistakes, which might cause problems in the exercises by giving wrong French translations to Estonian sentences.

For exercise generation purposes, we decided to exclude the EU legislative subcorpus, which contains a high proportion of long sentences, repetitive formulae and technical vocabulary. We also excluded the Bible, from which the Estonian and French translations included in the corpus are stylistically marked and do not represent standard contemporary language. However, the remaining subcorpora also contain many sentences which could be difficult to understand for language learners. Generating “good” grammar exercises thus implies selecting sentences fitted to the proficiency level of the learner, which means evaluating the readability of the sentences.

3. Selection of sentences, readability criteria

3.1 Previous work

Works on readability started in the early 40's (Dale & Chall, 1948; Flesch, 1948), mainly to improve native learners' reading skills. They used surface textual features, such as the average number of words or sentences, or the proportion of words not belonging to the basic vocabulary, combined through a linear regression model to set out simple readability formulae. Although this approach gave some acceptable results, it was criticised for its simplicity. Later works (Kintsch & Vipond, 1979; Redish & Selzer, 1985; Meyer, 1982) introduced more complex features, such as text cohesion, information density or macrostructure, but in fact for little gain. During the last fifteen years, with the progress and spread of corpus and NLP techniques, such as automatic classification, works on readability have been renewed (Collins-Thompson & Callan, 2004; Feng et al., 2010; François & Fairon, 2012). More and more complex features covering various linguistic fields (lexical, syntactic, semantic, discursive) are now implemented and evaluated for various languages. As for Estonian, work has been done since the 70's on the readability of textbooks for native speakers. A readability formula was proposed by Mikk (1980, 1991), based on two criteria: average length of independent sentences and abstraction level of repeated nouns.

Beyond its technical aspect we should not forget that the very notion of readability has several meanings, and most of them concern whole texts. For example, one can assess the readability of a text by testing its global understanding through the ability of writing an abstract or answering questions.

Moreover, the works on readability often differ when targeting the mother tongue (L1) or a foreign language (L2). Some works deal with French as a second language (Henry, 1975; Richaudeau, 1979; Daoust et al., 1996; François & Fairon, 2012). We are not aware of any similar work dealing with readability of Estonian as a second language.

Being concerned more, in this study, by short text segments or sentences than whole texts, our point of view on readability will follow that of Kilgarriff: "intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting names, anaphoric references or other deictics which cannot be understood without access to the wider context. We call this its 'readability'" (Kilgarriff et al., 2008).

So we will define readability as the ability for a learner to understand the constituents and the structure of a sentence, sufficiently to modify or complete it.

It is known that cultural knowledge and familiarity with the domain facilitate the comprehension process. Nevertheless, as we are working with a bilingual corpus of general language and can provide the translation of any text segment, we assume, in this study, that the impact of world knowledge on readability, as we defined it above,

is largely neutralised and that the readability of a sentence, for a foreign language learner, depends mainly on two characteristics: its syntactical complexity and its lexical complexity.

3.2 Syntactical complexity

The intuitive meaning of the notion of syntactical complexity at sentence level can be defined in formal terms as the number of nodes in the parse tree of the sentence. In practice, this criterion is not applicable to large corpora, because identifying and counting nodes generally requires manual coding (Szmrecsányi, 2004: 1033).

A more automatable approach could consist in counting certain types of surface units which qualify as good indicators of structural complexity, such as subordinating conjunctions and relative pronouns, or commas in languages where they function mainly as clause separators (for Estonian, see e.g. Kerge, 2002). The drawback of this method is that it is language-specific: subordinating units are different in each language, and this type of units might not be pertinent for languages in which subordination is not materialised by specific words or in which complexity can be achieved by means other than subordination.

Another criterion of complexity which has been widely used is sentence length (i.e. the number of words of the sentence). It has the advantage of being language-independent and very easy to implement. It seems also quite pertinent. A comparison conducted on 50 English sentences suggests that counting words gives almost the same complexity rankings as counting the nodes or calculating a complexity index based on the number of subordinating units, verbal forms and noun phrases (Szmrecsányi, 2004). It seems indeed quite logical that long sentences are structurally more complex than shorter ones, even if there may be exceptions. Since counting words is the most economical method and gives very consistent results, we decided to adopt this criterion to evaluate the syntactical complexity of the sentences. We intuitively defined three length ranges: up to 10 words, from 11 to 15 words, and from 16 to 29 words. For a language such as Estonian, which uses fewer function words than English or French (it has no article and 14 declension cases which notably reduce the use of pre- or postpositions), adding five words to a sentence generally results in a significant increase in syntactical complexity.

If excessively long sentences are difficult to understand by language learners, sentences that are too short can also cause problems, because they are understandable only within a larger context. Three words seemed to be a minimum for an Estonian sentence to constitute a sufficiently clear and autonomous message. We thus excluded sentences shorter than three words.

3.3 Lexical complexity

Since the corpus is not balanced, we could not take as a criterion for evaluating lexical

complexity, the frequency of the lemmas in the corpus. Neither did we find reliable external data on the frequency of Estonian words. The first frequency dictionary of contemporary Estonian (Kaalep & Muischnek, 2002) is not fully satisfying, as it was made from a very small corpus (1 million words) and contains only 10 000 words. A newer frequency list, based on a larger corpus (15 million words), was recently released (<http://www.cl.ut.ee/ressursid/sagedused1/>). Although much more comprehensive (40 000 lemmas), it still contains some oddities (from a pedagogical point of view), such as the presence of very specific terms among the most frequent words, or very different rankings of words belonging to the same semantic series. We thus decided to evaluate the lexical complexity of sentences on the basis of manually compiled or checked word lists, i.e. the subsets of the GDEF.

The GDEF is divided into four subsets of entries: basic vocabulary (4000 words), small dictionary (10 000 words), lower-medium dictionary (15 000 words), and upper-medium dictionary (40 000 words). These headword lists have been established by GDEF lexicographers, who used as a basis the above mentioned frequency dictionary as well as entry lists compiled by the Institute of Estonian Language for an Estonian Fundamental Dictionary (Eesti keele põhisõnastik) and for a general bilingual dictionary base with Estonian as a source language (Eesti-X sõnastikupõhi). These lists compiled for lexicographical purposes appeared more consistent and better suited to pedagogical purposes than automatically calculated frequency lists. A reason for that is probably the fact that entry selection principles followed by lexicographers compiling small or medium dictionaries are somewhat similar to those followed by authors of language textbooks (priority given to concrete notions and words of everyday life, consistency of semantic series, etc.). The four subsets of the GDEF give us four levels of lexical complexity.

3.4 Global sentence complexity and its relationship with language proficiency

Combined with the three levels of syntactical complexity, the four levels of lexical complexity give us 12 categories. This classification is obviously too complex to be understandable by the learner. It has to be reduced to a limited number of proficiency levels. One has to determine which combinations of lexical and syntactical complexity give sentences that can be understood without too much effort (and with the help of the translation) by learners of each level. A quick evaluation led us to the following table of equivalences, which remains a working hypothesis and needs to be confirmed by a more comprehensive assessment. Proficiency levels are expressed according to the categories of the Common European Framework of Reference for Languages.

LC \ SC	1	2	3
1	A2	B1	B2
2	B1	B1	B2
3	B1	B1	B2
4	B2	B2	B2

Table 1: Sentence complexity and language proficiency
(LC: lexical complexity; SC: syntactical complexity)

3.5 Sentence selection process and results

The bitexts of the CoPEF corpus are aligned at a so-called segment level. A segment is usually a sentence, but not always. It can also be a set of sentences or a sentence chunk (see Table 2 below).

Before applying any complexity selection on the corpus segments, a filtering is made to keep only the valid ones. The segment validation process follows the rules here below.

	multi-sentence	single sentence	sentence chunk
Estonian literature	4,980	80,006	40,296
French literature	4,297	115,021	46,019
Estonian non-fiction	85	1,906	301
French non-fiction	973	16,573	4,264
European Parliament	26,506	532,630	63,279
TOTAL	11,279	497,511	561,116

Table 2: Types of segments and their number per subcorpus

1. The segment must not be a sentence chunk, but a set of one or more “well-formed” sentences, i.e. it must start with an upper-case letter and end with a strong punctua-

tion; it must contain at least one finite verb; it must contain more than two words but fewer than thirty.

2. The segment must contain only acceptable words, i.e. words which are either a supposed proper nouns or an entry in one of the four subsets of the GDEF dictionary.

The resultant set of valid segments is then broken up into twelve subsets combining the four lexical and the three syntactic complexity levels (Table 3).

A final step reduces them to three segment sets according to the patterns of Table 1. They correspond to the three desired proficiency levels.

The numbers of segments for each level are as follows: A2: 22 558; B1: 21 758; B2: 10 862. As can be seen from the table below, the percentage of selected segments is quite low (5.9% of the total). It is significantly lower for the European Parliament subcorpus than for the other subcorpora, and, among the latter, significantly higher for French literary texts. This reflects, on the one hand, the higher lexical complexity of European Parliament debates (more technical terms) and, on the other hand, the lesser complexity of Estonian literary translations, as compared with Estonian original texts.

		Estonian literature	French literature	Estonian non-fiction	French non-fiction	European Parliament	TOTAL
Corpus total size		125 282	165 337	2 292	21 810	622 415	937 136
LC1	SC1	3 247	6 454	25	443	12 389	22 558
	SC2	304	308	4	43	1 725	2 384
	SC3	52	45	5	13	413	528
LC2	SC1	1 793	3 662	35	376	5 039	10 905
	SC2	422	486	22	95	1 851	2 876
	SC3	134	128	3	35	751	1 051
LC3	SC1	639	1 287	14	150	2 099	4 189
	SC2	174	228	6	42	954	1 404
	SC3	60	77	7	30	546	720
LC4	SC1	843	1615	14	180	2 759	5 411
	SC2	288	371	17	83	1 334	2 093
	SC3	109	145	8	38	759	1 059
Total number of selected segments		8 065	14 806	160	1 528	30 619	55 178
% of selected segments		6,4	9,0	7,0	7,0	4,9	5,9

Table 3: Number of segments at different complexity levels in the corpus (LC: lexical complexity; SC: syntactical complexity)

4. Converting sentences into exercises

4.1 Types of exercises

Taking into account the main difficulties of learners of Estonian as a foreign language, we generate two types of exercises, aimed at developing two types of language competence: 1) morphological competence (constructing forms), and 2) syntactical competence (choosing the appropriate form in a given context).

Morphological exercises present the user with sentences in which one inflected verb or substantive has been replaced by a textbox containing the corresponding lemma. Each exercise deals only with one type of form (e.g. partitive plural or indicative present), so the user knows which case and number or tense and mood has to be used and his/her task consists only of constructing the form and typing it in the text box. We generate this type of exercise for all declension cases (except singular nominative) and for the main verbal forms (present indicative, simple past indicative, present conditional, present imperative). For verbal forms, we give an additional hint after the lemma that tells the user which person has to be used, because there are many sentences in which the person cannot be predicted from the context. The French translation can help the user to disambiguate in many, but not all, cases. Performing separate exercises on each person would be too monotonous for the learner.

Syntax exercises are more difficult to generate, because the corpus is tagged only morphologically. It is still possible to imagine some types of syntax exercises relying only on morphological tags. The most obvious topic that can be dealt with is the use of declension cases: the user is presented sentences in which various case forms are replaced by textboxes with the corresponding lemmas. He/she must find which case has to be used in the context and construct the inflected form. Exercises can either mix all cases indifferently or concentrate on a certain subset of cases which can be used for similar syntactic purposes (e.g. nominative, genitive and partitive, which in Estonian can all be used to mark the object, depending on the context, or the so-called local cases, which are used to form adverbials of place or direction). For successfully performing this type of exercise, the learner needs to see the translation, otherwise many forms are impossible to predict unequivocally. An alternative possibility is to provide at the beginning the list of all inflected forms which have to be placed in the different sentences.

Another syntax topic on which we can generate exercises is the use of adpositions (postpositions and prepositions). In each sentence an adposition is replaced by a textbox. The user has to find the adposition fitting to the context (adpositional reaction of a verb or a nominal) and/or to the meaning of the sentence (here also translation is necessary). The list of adpositions which have to be placed in the blanks can be given or not in the beginning of the exercise.

We also consider the possibility of generating exercises on particle verbs, taking as a basis the list of verbs identified as such in the GDEF (1411 particle verbs combining one of 460 simple verbs with one of 67 adverbial particles). The user would be asked to identify in a list the appropriate particle (or the appropriate couple verb-particle) to fill the blank(s) in a sentence. A specific problem for generating that type of exercise is the fact that the particle can be placed either in the left context of the verb (with infinitives and participles) or in the right context (with finite forms). In the latter case, it is often separated from the verb by other constituents. Furthermore, many particles can also be used as adverbs, in which case they do not form a lexical unit with the verb. On the whole, automatically identifying particle verb constituents in order to create exercises seems possible, but rather tricky. We identified possible solutions, but left their implementation as a direction for further work.

4.2 Generation process

4.2.1 Exercise definition and configuration

Through an HTML form (Fig. 1), the user is asked to define the type of the desired exercise, i.e.:

- its class (e.g. nominal or verbal morphology, use of cases, adpositions, particle verbs);
- its precise content (e.g. case and number for nominal morphology, mood and tense for verbal morphology).
- The user must then specify the source of segments from which the exercise items are to be generated. He or she will define:
- the set of subcorpora to be used,
- the proficiency level.

Figure 1: Screenshot of the exercise generator

Some hidden parameters, automatically set, help control item generation and exercise layout.

4.2.2 Exercise generation and display

The generation process first selects candidate-items. To do so, it obtains the list of tagged Estonian segments of the desired level from the chosen subcorpora. Then it parses them at both morphological and syntactical level to filter out any segments that do not fit the specified type of exercise, or that would lead to some identified ambiguities (e.g. we filter out verbal forms ending with the emphatic particle *-gi/-ki*, which is not tagged).

Among the candidate items, a very limited number are selected to be ‘blanked out’ and become part of the exercise, according to the following principles:

- one blank per item (or more than one for the advanced level, if the sentence length allows it);
- a similar lemma will never be reused as a blank within the current exercise (this is necessary to avoid over-representation of very frequent words, such as the verb *olema* ‘to be’ in verbal morphology exercises);
- items are chosen randomly.

The French translation is then retrieved and associated to the item. A complementary feature could consist of linking each lemma of the item to the corresponding article of the GDEF. This would assist the learner in developing his/her lexical knowledge and overcoming possible comprehension difficulties due to loose translation of the segment (quite frequent in literary texts). The implementation of this feature will become relevant when at least one subset of the GDEF is fully available, which is not yet the case.

The requested exercise is generated as an XML document describing, on one hand, the different items (Estonian blanked out text, French translation, answer), and, on the other hand, the various generation and layout parameters. An XSL style-sheet transforms it into a dynamic HTML document.

The exercise generator provides the user with an HTML fill-in-the-blank exercise (Figure 2) with classical functionalities, like “answer evaluation”, “reset”, “answers” and various help modes (lemma in the blank, list of possible answers, no help at all).

comitatif singulier
Écrivez dans les cases la forme qui convient.

Niveau A2

Mode d'emploi
Évaluer
Recommencer
Solution

Sciences humaines
le 25/10/2013

1. Täidame puhta **süda** oma igapäevast kohust.
Accomplissons, en conscience notre tâche quotidienne.
2. Mõne **nädal** kadusid need täielikult..
En quelques semaines, celles-ci disparurent complètement...
3. **kohtuotsus** ei tahetud kuidagi leppida.
La sentence est mal acceptée.
4. Ta tahtis seda kõigest väest, kogu oma **olemus**.
Il le voulait de tout son désir, de tout son être.
5. Piirdun paari **näide**.
Quelques exemples seulement.
6. Mitte sellest maailmast **riik** ei olnud Martin Lutheril asja.
Ce n'était pas du royaume de ce monde que Martin Luther avait à s'occuper.
7. Esmalt teda suure **tõenäosus** ei usuta.
D'une part, il a toute chance de ne pas être cru.
8. Neil pole **vaim** midagi tegemist.
Ils n'ont rien à voir avec l'esprit.

Figure 2: Screenshot of an exercise on comitative singular

4.3 Results and evaluation

In the last stage of the project, it will of course be necessary to have all types of exercises evaluated by learners of Estonian as a foreign language at different proficiency levels. At the present stage, we evaluated the linguistic and pedagogical relevance of 991 automatically generated exercise items, selected randomly among the 6454 A2-level (LC1-SC1) segments of the French literature subcorpus (and also for adposition exercises in the LC2-SC2 and LC3-SC3 segments of the same subcorpus). This preliminary evaluation was made by Antoine Chalvin, in the light of his 15 years' experience of teaching Estonian grammar to French students. It appeared that the overwhelming majority of items were linguistically pertinent (the form in the blank corresponded to the topic of the exercises) and pedagogically appropriate (blanks were possible to fill with the help of hints, the context and/or the translation). Exercises on verbal morphology had the highest reliability rate (97%), followed by exercises on case forms other than genitive and partitive singular (91%). Exercise on these last two forms contained, as expected, a significant number of errors (only 77% of the items were adequate). Exercises on adpositions were the least reliable (67%).

The detailed analysis of exercises revealed several types of problems, which made some items difficult or disconcerting for the learner.

A first category of problems was caused by errors in lemmatisation or morphological analysis. At this stage, we were unable to solve this problem, because identifying and correcting errors in the corpus would have been very time consuming. In the

exercises we generated, we discovered a few recurrent errors which could be searched and corrected semi-automatically in the corpus. For example, several verb forms ending in *-ta* (factitive derivational suffix or infinitive ending) were wrongly analysed as nouns in the abessive case (the abessive suffix is *-ta*), several active past participles (in *-nud*) were analysed as plural nominative of substantives in *-nu* (which is a far less common form), several postpositions or adverbs ending in *-l* were analysed as adessive forms of substantives (suffix *-l*), etc. If correcting errors in the corpus proves too difficult, another way to solve the problem would be to generate a list of ambiguous forms and exclude them from exercises in which a confusion is possible (e.g. in an exercise on the translative case, never create a blank on the form *peaks*, which, though analysed as the translative singular of *pea* ‘head’, could in fact be the conditional present of the verb *pidama* ‘have to’).

A pedagogical problem which affected mainly exercises on adpositions was the possibility of multiple correct answers, either because the translation was not sufficient to specify the meaning of the sentence, or because, although the meaning was clear, several synonym adpositions could be used, but only one of them being recognised as correct by the automatic correction system. This could be frustrating and disconcerting for the learner. A possible way to reduce the impact of this problem could be to make a list of synonym adpositions (such as *saadik* and *peale* ‘since’, *seas* and *hulgas* ‘among’) and instruct the system to accept them as correct variants.

The problem of multiple answers also affects exercises dealing with plural forms of substantives, because Estonian has two plural paradigms. The so-called *i*-plural, usually very rare, nonetheless occurs rather frequently for certain words as a variant of the more common *de*-plural (*aastail* vs. *aastatel* ‘in the years’; *päevil* vs. *päevadel* ‘in the days’). The morphological tags in the corpus do not distinguish these variants. However, in the 991 items analysed, we found very few *i*-plural forms.

A third problem affects morphology exercises combining several forms (e.g. several persons in verb exercises, or several cases in multi-case exercises), namely, the excessive predominance of certain forms in the questions. One of the forms dealt with in a given exercise could be much more frequent in the corpus than the other forms. If exercise items are picked up randomly in the corpus, this particular form has chances to be more present also in the exercise, leaving little space for the others. This is the case, for example, in our conjugation exercises, where the third person singular concerns at least 60% of the items. To reduce monotony and maximise the usefulness of these exercises, it will be necessary to find a way to balance the representation of the forms.

The last (minor) problem is excessive easiness. In exercises on nominal and verbal morphology, many forms are very easy to construct, because the stem serving as a basis (singular genitive for substantives, indicative present stem for verbs) is easily predictable from the lemma. In order to make exercises more interesting and more

useful for the learner, we should find a way to over-represent problem words, i.e. words whose radical is not predictable from the lemma. Lists of such words could be easily generated with the aid of morphological data included in the GDEF.

5. Conclusion

By applying sentence readability criteria to a large real language corpus of around 940 000 segments, we generated a ‘readable’ corpus of 55 000 segments. We showed that, on the basis of such a corpus, it is possible to generate a very high number of fill-in-the-blank grammar exercises that can serve as a useful training material for learners of Estonian, without it being necessary to submit these exercises to prior manual control and filtering by a language teacher. On the whole, generated exercises have a surprisingly high degree of pertinence and reliability. Residual problems, such as lemmatisation errors, possibility of multiple answers, monotony of questions and excessive predictability of answers, do not seem insurmountable and will be addressed in a second stage of the project. Once operational, the system will be made freely available on the Internet.

A possible further development, on the basis of the same corpus, could be a French grammar exercise generator for Estonian learners. This would probably be even easier to implement, due to the lower frequency of morphological homography in French as compared with Estonian.

The general methodology of our project and large parts of the program could also be applied to other language pairs for which a reliable morphologically tagged parallel corpus of general language is available.

6. References

- Beltrachini, L., Gardent, C. & Kruszewski, G. (2012). Generating Grammar Exercises. In *The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT Workshop 2012*. Montreal, Canada.
- Boulton, A. (2008). Esprit de corpus: promouvoir l’exploitation de corpus en apprentissage des langues. *Texte et Corpus*, 3, pp. 37-46.
- Brown, J. C., Frishkoff, G. A. & Eskenazi, M. (2005). Automatic Question Generation for Vocabulary Assessment. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Pp. 819-826.
- Chen, C.-Y., Liou, H.-C. & Chang J. S. (2006). FAST – An Automatic Generation System for Grammar Tests. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney: Association for Computational Linguistics.
- Collins-Thompson, K. & Callan, J. (2004). A language modeling approach to

- predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. Boston, pp. 193-200.
- Coniam, D. (1997). A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Cloze Tests. *CALICO Journal*, 2-4, pp. 15-33.
- Dale, E. & Chall, J.S. (1948). A formula for predicting readability. *Educational research bulletin*, 27(1) pp. 11-28
- Daoust, F., Laroche, L. & Ouellet, L. (1996). SATOCALIBRAGE: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1), pp. 205-234.
- Feng, L., Martin Jansche, M., Huenerfauth, M., Elhadad, N. (2010). Comparison of Features for Automatic Readability Assessment. In *Proceedings of Coling 2010 (Poster Volume)*, Beijing, pp. 276-284.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3) pp. 221-233.
- François, T. & Fairon, C. (2012). An AI readability Formula for French as a Foreign Language. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, Jeju, South-Korea, pp. 466-477.
- Henry, G. (1975). *Comment mesurer la lisibilité ?* Bruxelles: Labor.
- Hoshino, A. & Nakagawa, H. (2005). A Real- Time Multiple-Choice Question Generation for Language Testing: A Preliminary Study. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Ann Arbor, Michigan, pp. 1-8.
- Hoshino, A. & Nakagawa, H. (2008). A Cloze Test Authoring System and Its Automation. Advances in Web Based Learning. In *ICWL 2007 : 6th International Conference Edinburgh, UK, August 15-17, 2007*. Berlin/Heidelberg: Springer, pp. 252-263.
- Huang, L.-S. (2011). Corpus-aided language learning. *ELT Journal*, 65(4), pp. 481-484.
- Johns, T. (1994). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (ed.). *Perspectives on Pedagogical Grammar*. Cambridge: Cambridge University Press, pp. 293-313.
- Kaalep, H.-J. (1996). ESTMORF, a Morphological Analyzer for Estonian. In H. Õim (ed.) *Estonian in the Changing World*. Tartu, pp. 43-98.
- Kaalep, H.-J. (1998). Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. *Keel ja Kirjandus*, 1/1998, pp. 22-29.
- Kaalep, H.-J., Muischnek, K. (2002). *Eesti kirjakeele sagedussõnastik*. Tartu: TÜ kirjastus.
- Kerge, K. (2002). *Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX*

- sajandil*. TPÜ eesti keele osakonna veebitoimetised, *Lingvistika* 1.
<http://digar.nlib.ee/digar/contentpdf?key=637f6b16470041ae9dof91a6ofde1410&group=2> Accessed 27 August 2013.
- 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 XIII EURALEX International Congress*, Barcelona: Universitat Pompeu Fabra, pp. 425-433.
- Kilgarriff, A., Smith, S. & Avinesh, P.V.S. (2010). Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.
- Kintsch, W. & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L.G. Nilsson (ed.) *Perspectives on Memory Research*. Hillsdale NJ: Lawrence Erlbaum, pp. 329-365.
- Lee, J. & Seneff, S. (2007). Automatic Generation of Cloze Items for Prepositions. In *Interspeech 2007*, vol. 3, pp. 2173-2176.
- Liu, C.L., Wang, C.H., Gao, Z.M., & Huang, S.M. 2005. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items, In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pp. 1-8, Ann Arbor, Michigan, 2005.
- Meyer, B.J.F. (1982). Reading research and the composition teacher: The importance of plans. *College composition and communication*, 33(1), pp. 37-49.
- Mikk, J. (1980). *Teksti mõistmine*, Tallinn: Valgus.
- Mikk, J. (1991). Studies on teaching material readability. In *Papers on education II: Problems of textbook effectivity*, Tartu, pp. 34-50.
- Mitkov, R. & Ha, L.A. (2003). Computer-Aided Generation of Multiple-Choice Tests. In *Proceedings of the HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing*, Edmonton, Canada, May, pp. 17-22.
- Redish, J.C. & Selzer, J. (1985). The place of readability formulas in technical communication. *Technical communication*, 32(4), pp. 46-52.
- Richaudeau, F. (1979). Une nouvelle formule de lisibilité. *Communication et Langages*, 44, pp. 5-26.
- Szmrecsányi, Benedikt M. 2004. On Operationalizing Syntactic Complexity. In : *JADT 2004 : 7es Journées internationales d'Analyse statistique des données textuelles*, pp. 1031-1038.
- Tahmm (1998) = Morfoloogiline ühestaja (beetaversioon).
<http://www.eki.ee/keeletehnoloogia/projektid/tahmm/tahmm.html>. Accessed 10 April 2013.

***Kommunikationsverben* in OWID:
An Online Reference Work of German Communication
Verbs with Advanced Access Structures**

Carolin Müller-Spitzer, Kristel Proost

Institut für Deutsche Sprache, R 5, 6-13, D-68161 Mannheim, Germany
E-mail: mueller-spitzer@ids-mannheim.de, proost@ids-mannheim.de

Abstract

Kommunikationsverben, an online reference work on German communication verbs and part of the dictionary portal OWID, describes the meaning of communication verbs on two levels: a lexical level, represented in the dictionary entries and by sets of lexical features, and a conceptual level, represented by different types of situations referred to by specific types of verbs. These two levels have each been implemented in special types of access structures. A first explorative access to the conceptual level provides the user with a list of the main classes of communication verbs, the subclasses of each of these, and the lexical fields pertaining to each subclass. Lexical fields are presented together with a characterisation of the situation type to which the verbs of that field are used to refer. Information about the conceptual level is additionally accessible by an advanced search option allowing the user to combine components of the characterisation of situation types to “create” any kind of situation and search for the verbs that correspond to it. Information about the lexical level of the meaning of communication verbs is accessible via the dictionary entries and by another advanced search option allowing the user to search for verbs with particular lexical features or combinations of these.

Key words: communication verbs, lexical fields, online dictionary, access structures, advanced search options

1. Communication Verbs

This contribution deals with the different types of information offered by *Kommunikationsverben*, the online version of the *Handbuch deutscher Kommunikationsverben* (cf. Harras et al., 2004; Harras, Proost & Winkler, 2007), which has recently been integrated into the dictionary portal OWID (‘Online-Wortschatz-Informationssystem Deutsch’ www.owid.de) of the Institut für Deutsche Sprache (IDS). *Kommunikationsverben* contains about 800 verbs, 241 of which are lemmatised and appear with an entry of their own. All other verbs are listed as synonyms of the verbs lemmatised and differentiated from these in extensive synonymy commentaries included in the entry of the corresponding lemmatised verb.

In *Kommunikationsverben*, communication verbs are understood to be verbs used to refer to situations in which a speaker (henceforth: S) utters something to a hearer (henceforth: H). In the default case, the speaker’s utterance also contains a

proposition (henceforth: P). Examples of German communication verbs are *sagen* ('to say'), *sprechen* ('to speak'), *behaupten* ('to assert'), *bestreiten* ('to deny'), *mitteilen* ('to inform'), *versprechen* ('to promise'), *auffordern* ('to request'), *loben* ('to praise'), *klagen* ('to complain'), *schreien* ('to shout'), *unterbrechen* ('to interrupt'), and *mailen* ('to mail'). The term "speech act verbs" is used to refer to the much smaller set of verbs lexicalising speaker attitudes including the speaker's propositional attitude, i.e. the attitude of the speaker to the proposition of his/her utterance, the speaker's intention, and the speaker's presuppositions (cf. Proost, 2006: 65; 2007: 8–9). Of the communication verbs mentioned above, only *behaupten* ('to assert'), *bestreiten* ('to deny'), *mitteilen* ('to inform'), *versprechen* ('to promise'), *auffordern* ('to request'), *loben* ('to praise'), and *klagen* ('to complain') are speech act verbs. *Kommunikationsverben* focuses on speech act verbs.

Following a distinction made in two-levels-semantics (cf. Bierwisch & Lang, 1989; Bierwisch & Schreuder, 1992; Lang, 1994), *Kommunikationsverben* describes the meaning of German communication verbs as comprising two levels: a conceptual level, represented by different types of situations referred to by specific types of speech act verbs, and a lexical level, represented in the dictionary entries. As will be shown below, these different levels have each been implemented in special types of access structures.

2. The Conceptual Level of the Meaning of Communication Verbs

2.1 The General Resource Situation Type

All situations referred to by communication verbs are characterised by the presence of four features or situational roles: a speaker, a hearer, a set of speaker attitudes, and an utterance (mostly) containing a proposition. Since these four elements are part of any situation referred to by communication verbs, they constitute the unifying feature of the meaning of these verbs (cf. Verschueren, 1980: 51–57; 1985: 39–40; Wierzbicka, 1987: 18; Harras et al., 2004: Introduction; Proost, 2006: 651). The type of situation referred to by all speech act verbs is therefore called the 'general resource situation type'.

2.2 Specifications of the Role of the Speaker Attitudes and of the Propositional Content

Two of the roles of the general resource situation type, the role of the speaker attitudes and that of the utterance, may be specified in different ways. The role of the speaker attitudes may be specified as consisting of the speaker's attitude to the proposition of his/her utterance, the speaker's intention, and the speaker's presuppositions. The speaker's propositional attitude may be further specified as S taking P to be true, S knowing P, S wanting P, S evaluating P positively or negatively,

and so on. Specifications of the speaker's intention include S's intention to make H recognise S's propositional attitude (for example, to make H recognise that S knows P or takes P to be true) or to get him/her to do something. The speaker's presuppositions may concern an attitude of H (whether H takes something to be true, whether he/she knows something), the interests of S and H concerning P (whether P is in the interest of S or in the interest of H), or properties of P (for example, whether P is the case). The role of the utterance is specified by properties of the propositional content. These include the event type of P (whether P is an action, event, or state of affairs), the temporal reference of P (whether P precedes, coincides with, or follows the time of S uttering P) and, in the case that P is an action, the agent of P (S, H, S & H, and so on).

2.3 Methods Used

Following a procedure proposed by Baumgärtner (1977: 260–264), the different specifications of the role of the speaker attitudes and the role of the utterance as well as the lower-level specifications of each of these, are obtained from a comparison of sentences containing speech act verbs. The well-formedness of some of these and the ill-formedness of others show which elements are relevant to the meaning of the verbs they contain. For example, a comparison of the sentences in (1) and (2) shows that *to order* lexicalises the values 'future', 'action' and 'hearer' for the specifications of the temporal reference, the event type and the agent of P, respectively, while *to promise* lexicalises the values 'future', 'action' and 'speaker', respectively, for these specifications:

- (1) a. I *order* you_i to PRO_i leave the room.
 b. *I *order* you_i to PRO_i have left the room.
 c. *I *order* you_i for me_j to PRO_j leave the room.
- (2) a. I_i *promise* you to PRO_i leave the room.
 b. *I_i *promise* you to PRO_i have left the room.
 c. *I_i *promise* you_j to PRO_j leave the room.

The introspective analysis exemplified in (1) and (2) has shown that the higher-level specifications of the speaker's propositional attitude, the speaker's intention, the speaker's presuppositions and the propositional content, are essential aspects of the meaning of speech act verbs. These four aspects correspond to five of the seven components of illocutionary force which Searle & Vanderveken (1985: 12–20) and Vanderveken (1990: 103–136) have argued to determine the conditions under which a particular type of speech act is both successful and non-defective. Particularly, the aspect of the speaker's propositional attitude corresponds to the component of the sincerity conditions, the aspect of the speaker's intention to the component of the illocutionary point, the aspect of the speaker's presuppositions to the components 'mode of achievement of the illocutionary point' and 'preparatory conditions', and the aspect of the propositional content to the component of the propositional content

conditions (cf. Harras, 2001: 26–31, Proost, 2006: 654–655).

While the higher-level specifications of the speaker’s propositional attitude, the speaker’s intention, the speaker’s presuppositions and the propositional content are obtained from the type of analysis exemplified in (1) and (2), the lower-level specifications of each of these are calculated systematically, i.e. irrespective of any existing lexicalisations. For example, the specification ‘temporal reference of P’ is assumed to have the specifications ‘past’, ‘present’ and ‘future’, the specification of the event type of P, the specifications ‘action’, ‘state’ and ‘event’, and so on. The question of which values are lexicalised by a particular verb was decided on the basis of examples from the Mannheim German Reference Corpus DeReKo (“Deutsches Referenzkorpus”). Methodological issues are dealt with in detail in the introductions to both volumes of the *Handbuch deutscher Kommunikationsverben* (cf. Harras et al., 2004; Harras, 2007), which are also available in the online version.

2.4 Special Resource Situation Types

Different combinations of specifications of the different types of speaker attitudes and of the properties of the propositional content constitute special resource situation types. These are referred to by distinct types of verbs. For example, verbs like *behaupten* (‘to assert’) and *auffordern* (‘to request’) are used to refer to the situation types characterised by the specifications listed in Tables 1 and 2, respectively:

Special Resource Situation Type: Representatives.Assertives.behaupten	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S takes to be true: P
Intention	S wants: H recognise: S takes to be true: P
Presuppositions	H does not know: P

Table 1: Situation type referred to by *behaupten* (‘to assert’)

The combinations of the specifications of the speaker attitudes and of the properties of the propositional content lexicalised by *behaupten*, *auffordern*, and *mailen*, respectively, may also be conceived of as the concepts lexicalised by these verbs. Thus, *behaupten* (‘to assert’) lexicalises the concept of a verbal action performed by a speaker who takes P to be true and assumes that H does not know P with the intention that the hearer recognise that he/she (i.e. S) takes P to be true, P being an

action, event or state of affairs preceding, co-occurring with or following the time of S's utterance.

Special Resource Situation Type: Directives.Request.auffordern	
Propositional Content (P)	
Event Type	action
Temporal Reference	future
Agent	H
Speaker Attitudes	
Propositional Attitude	S wants: P
Intention	S wants: H do: P
Presuppositions	in the interest of S: P

Table 2: Situation type referred to by *auffordern* ('to request')

2.5 Lexical Fields

Verbs which are used to refer to the same special resource situation type constitute a "paradigm" or lexical field. For example, a situation in which a speaker who takes P not to be true and assumes that H does not know P tells a hearer that he/she takes P to be true, may be referred to not only by verbs like *lügen* ('to lie') and its prefixed forms *anlügen* ('to lie to sb. '), *belügen* ('to lie so sb. '), *erlügen* ('to lie about sth. '), *rumlügen* ('to tell lies') and *vorlügen* ('to lie to sb. about sth. '), but also by verbs like *schwindeln* and *flunkern* (both 'to fib') and the prefixed forms of these (*anflunkern*, *anschwindeln*, *beschwindeln*, *rumflunkern* etc.). The situation type referred to by these verbs may be characterised as follows:

Special Resource Situation Type: Representatives.Assertives.lügen	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S does not take to be true: P
Intention	S wants: H recognise: S takes to be true: P
Presuppositions	H does not know: P

Table 3: Situation type referred to by *lügen* ('to lie'), *schwindeln* and *flunkern* (both 'to fib') and their prefixed forms

3. The Lexical Level of the Meaning of Communication Verbs

Verbs that differ from each other with respect to their lexical meaning appear with an entry of their own. Lexical meanings were differentiated on the basis of examples from the IDS-corpora of written German. All other verbs are listed as synonyms of the verbs lemmatised. With respect to the *lügen*-field, this means that *lügen* ('to lie') and *schwindeln* ('to fib') each appear with a separate entry. These verbs differ from each other in that *schwindeln* but not *lügen* expresses an evaluation by a discourse situation speaker, i.e. a speaker who uses this verb to comment on the utterance of the resource situation speaker. Particularly, a speaker who uses the verb *schwindeln* to refer to the resource situation speaker's act of lying thereby indicates that he/she does not consider S's act of lying to have serious consequences for H. In *Kommunikationsverben*, this difference in the lexical meaning of *lügen* and *schwindeln* is reflected by the meaning paraphrases of these verbs in their respective entries. Since the evaluation expressed by *schwindeln* is an evaluation by a discourse situation speaker, it is not an element of the resource situation referred to by this verb. Hence, within the framework of *Kommunikationsverben*, it is not part of the conceptual component of its meaning. Rather, it is an essential part of the lexical component of the meaning of this verb.

3.1 Information about Lemmatised Verbs

Apart from meaning paraphrases, the dictionary entries list additional information for each of the lemmatised verbs in the following units:

- (i) FELDZUGEHÖRIGKEIT ('FIELD POSITION'). This unit provides information about the special resource situation type referred to by the verb and its synonyms as well as the position it occupies within the hierarchy of resource situation types. This information is reflected by the name of the special resource situation type (e.g.: "Representatives.Assertives.behaupten" is meant to indicate that *behaupten* ('to assert') belongs to the group of assertives, which is a subclass of the class of representatives).
- (ii) LEXIKALISCHE BEDEUTUNG ('LEXICAL MEANING'). This section of the dictionary entry includes a colloquial paraphrase of the lexical meaning of the verb as well as a paraphrase which explicitly makes reference to the elements of the corresponding special resource situation type. The meaning paraphrases given for *lügen* ('to lie'), for example, are: 'to say something which one does not believe to be true' and 'speaker S addresses one or more utterances with a propositional content P to a hearer H with the intention that H recognises that S takes P to be true; S does not take P to be true.'
- (iii) VERWENDUNGSSPEZIFIK ('SPECIFICS OF USAGE'). This unit lists the pragmatic properties of the lemmatised verb and includes information on whether it

belongs to a particular stylistic or regional register (whether it may be used performatively), as well as its contextual restrictions (whether the roles of S, H and P are realised optionally or obligatorily and whether they may be realised at all, the way in which these roles are realised syntactically, typical modifiers of the verb in question, its collocates etc.). For *lügen*, the section VERWENDUNGSSPEZIFIK lists the following information:

- The role of H may be realised by an adpositional phrase with *gegenüber* ('in front of', 'to') (*jemand hat jemandem gegenüber gelogen* ('someone lied to somebody')).
- The role of P cannot be realised.
- *lügen* is often used in semi-idiomatic expressions like *lügen wie gedruckt* (literally: 'to lie as if it were printed'), *lügen, dass sich die Balken biegen* (lit.: 'to lie until the beams bend') and *das Blaue vom Himmel (her)unterlügen* (lit.: 'to lie the blue down from heaven').
- *lügen* cannot be used performatively.

(iv) SYNONYME ('SYNONYMS'). This section lists all verbs and fixed multiword expressions which may be used as synonyms of the lemmatised verb. Multiword expressions are mentioned in round brackets. For example, verbs mentioned as synonyms of *lügen* are *anlügen* ('to lie to somebody'), *belügen* ('to lie to somebody'), *vorlügen* ('to lie to somebody about something'), *rumlügen* ('to tell lies') and *erlügen* ('to lie about something').

(v) ANTONYME ('ANTONYMS'). In this unit, antonyms of the lemmatised verb are listed where present. Since there are no verbs with the meaning 'to tell the truth' (cf. **wahrsagen*, **wahren*) in German, no antonyms are mentioned for *lügen*. The entry for *loben* ('to praise'), for example, mentions *tadeln* ('to reprimand') as an antonym of *loben*.

(vi) KOMMENTAR ('COMMENTARY'). This section provides information about the restrictions on the range of contexts the non-lemmatised verbs may occur with. The section KOMMENTAR in the entry for *lügen*, for example, mentions the following context restrictions for the prefixed verbs *anlügen*, *belügen*, *vorlügen*, *erlügen* and *rumlügen*:

- *anlügen*, *belügen* and *vorlügen* differ from *lügen* in the syntactic realisation of their arguments: *anlügen* and *belügen* obligatorily realise the role of H as an NP in the accusative case; *vorlügen* realises the role of P obligatorily either as an NP in the accusative case or as a finite subordinate clause. With the exception of the differences in their argument structures, these four verbs may be used as synonyms as illustrated by the following examples:

- Der Ministerpräsident hat vor dem Untersuchungsausschuss gelogen. ('The prime minister lied to the commission.')
- Der Ministerpräsident hat den Untersuchungsausschuss angelogen/belogen. ('The prime minister lied to the commission.')
- Der Ministerpräsident hat dem Untersuchungsausschuss vorgelogen, dass er mit dem Fall nichts zu tun habe. ('The prime minister lied to the commission, telling them that he did not have anything to do with the affair.')
- *rumlügen* is often used in utterances like *Lüg hier nicht so rum!* ('Don't go about telling lies!'), which express a discourse speaker's criticism of the verbal behaviour of a resource situation speaker. It is also frequently used with reference to situations in which a speaker tells several lies to several hearers.
- *erlügen* is usually used in the perfect tense as in *Diese Geschichte ist erlogen* ('This story is a lie').

(vii) BELEGE ('EXAMPLES'). This unit contains a selection of the examples from DeReKo which served as the empirical basis of the information in the dictionary entries.

3.2 Lexical Features

Each of the lemmatised and non-lemmatised speech act verbs (representatives, directives, commissives and expressives) and each of the communication verbs expressing a particular mode of speaking is characterised as having or lacking the following features: (i) the possibility of the realisation of the thematic roles of H and P¹, (ii) the syntactic realisation of the thematic roles, (iii) the possibility for the verb to be used in the passive voice, (iv) resultativity, (v) lexicalisation of an evaluation by a discourse situation speaker, (vi) polysemy, (vii) the possibility for the verb to be used performatively, and (viii) stylistic register. Information about lexical features is presented in the form of tables which the user may access by selecting the name of one of the resource situation types listed under the menu item "Wortartikel/Paradigmen" ('entries/lexical fields'). The screenshot in Figure 1 shows the lexical features of *lügen* ('to lie') and its synonyms:

¹ The situational roles of the speaker, the hearer and the propositional content correspond to the thematic roles 'Speaker', 'Hearer' and 'Propositional content' used in *Kommunikationsverben* to describe the argument structure of communication verbs. These thematic roles are similar to the roles of the Speaker, the Addressee, and the Message used to describe the meaning of communication verbs in FrameNet (cf. Boas 2010: 61–65). The roles of the Speaker, Hearer and Propositional content may be taken to be special instances of the more general roles 'Source', 'Target' and 'Theme', respectively.

Lexikalische Merkmale

Verben	Merkmale							
	Seman- tische Rollen	Argu- ment Struktur	Passiv	Resulta- tivität	Bewer- tung	Poly- semie	Performa- tivität	stilistische Markiert- heit
lügen	H (block)		+	-	-	-	-	-
	P (block)							
anlügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
belügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
erlügen	H (block)		+	-	-	-	-	-
	P (obl)	NP<Akk>						
rumlügen	H (block)		+	-	-	-	-	+
	P (block)							
vorlügen	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE Inf						

Fig. 1: Lexical features of *lügen* ('to lie') and its synonyms

The argument structure properties of the verbs in Figure 1 are illustrated by the following examples from DeReKo (the verbs' arguments are indicated by square brackets, their syntactic realisations by round brackets; S: Speaker, H: Hearer, P: Propositional Content):

- (3) Anwalt Gregory Craig sagte, in der Anklageschrift gebe es keine konkreten Beweise, daß [der Präsident]_{S(NP-nominative)} *gelogen* habe. [Berliner Zeitung, 22.01.1999]

Attorney Gregory Craig stated that there was no concrete evidence in the indictment sheet that the President had lied.

- (4) Auch 2010 werden [die Politiker]_{S(NP-nominative)} [uns]_{H(NP-accusative)} wieder *anlügen* und uns Geschenke machen, die wir selbst bezahlen. [Mannheimer Morgen, 23.01.2010]

In 2010 too, politicians will once again belie us and give us presents that we have to pay for ourselves.

- (5) Tagelang *belog* [er]_{S(Pro-nominative)} [Trainer Erik Gerets]_{H(NP-accusative)} und bestritt seine Anwesenheit in dem Club. [Braunschweiger Zeitung, 28.03.2008]

For days he had lied to trainer Erik Gerets and denied his presence in the Club.

- (6) [Ich]_{S(NP-nominative)} habe gar keinen Vorteil davon, [diese

Behauptung]_{P(NP-accusative)} zu "*erlügen*". [Diskussion:Mikojan-Gurewitsch MiG-105, In: Wikipedia-URL:http://de.wikipedia.org/wiki/Diskussion:Mikojan-Gurewitsch_MiG-105: Wikipedia, 2011]

I gain no advantage from contriving an untruth.

- (7) Kerstin Brandt braust auf. »Dann *lüg* hier nicht die ganze Zeit *rum!*«

[Die Zeit (Online-Ausgabe), 29.11.2001; Der Prozess [S. 74]

Kerstin Brandt flared up. "Then don't be lying the whole time."

- (8) ..., daß [die Frau]_{S(NP-nominative)} [den Ärzten]_{H(NP-dative)} im Krankenhaus] *vorlog*, [im Haus eines Bekannten "einfach mal ausgeholfen" zu haben]_{P(infinitival clause)}.

[Frankfurter Allgemeine, 11.07.2001; Schwarzarbeit im Haushalt rächt sich nicht immer Razzien vor allem auf Baustellen und in Gaststätten / Bis zu 90 000 illegale "Dienstmädchen" in Hessen]

..., that the woman lied to the doctors at the hospital, saying that she only helped out in the house of an acquaintance.

4. Degrees of Synonymy

Verbs which are listed in *Kommunikationsverben* as synonyms of other verbs may be synonymous with these to different degrees. Verbs which are used to refer to the same special resource situation type such as, for example, *lügen*, *schwindeln*, *flunkern* and their prefixed forms are considered to be synonyms in a broader sense. Verbs which may be substituted in specific contexts such as, for example, *lügen*, *anlügen*, *belügen* and *vorlügen* (see section 3) are regarded as synonyms in a narrower sense.

5. Explorative Access

A first explorative access to *Kommunikationsverben* via the menu item "Wortartikel/Paradigmen" ('entries/lexical fields') provides a clustering of German communication verbs by main verb classes. These include the general communication verbs, the five main types of speech act verbs (representatives, directives, commissives, expressives and declaratives), and the different classes of communication verbs (verbs expressing a particular mode of speaking, verbs expressing a communication medium, verbs referring to conversational structure, ...).

By selecting one of the main classes, the user is presented with a window showing the different types of verbs subsumed under the larger class, for example, "Assertive" ('Assertives') and "Informationsverben" ('information verbs') for the class of representatives, "Auffordern" ('request'), "Verbieten" ('forbid'), "Erlauben" ('allow'),

“Fragen” (‘ask’) and “Raten” (‘recommend’) for the class of directives, “Lautstärke” (‘sound intensity’), “Artikulation” (‘articulation’), “Intonation” (‘intonation’), “Stimmqualität” (‘quality of voice’), “Rhythmus” (‘rhythm’) and “Iterativität” (‘iterativity’) for verbs expressing a particular mode of speaking, and so on. The different types of verbs of a larger class are shown together with characterisations of special resource situation types. These are the types of situations to which verbs of that type are used to refer. They are listed together with the corresponding lexical fields. The class of directives of the type “Auffordern” (‘request’), for example, is presented together with the special resource situation types to which directives of that type may be used to refer. Figure 2 shows the information presented to the user for directives of the type “Anleiten”:

The screenshot shows a web interface titled "Paradigmenübersicht" under the heading "Kommunikationsverben". On the left, there is a vertical list of verb types: Allgemein, Repräsentativ, **Direktive**, Kommissiv, Expressiv, Deklarativ, Gesprächs- und themenstrukturierende, Redesequenzverben, Modale, Mediale, Eröffnende, and Abschließende. A dropdown menu is open over "Direktive", listing "Auffordern", "Verbieten", "Erlauben", "Fragen", and "Raten". The main content area is divided into two sections for "Paradigmen" and "Verben". The first section, titled "Bezugssituationstyp: Direktive.auffordern.anleiten", lists: Propositionaler Gehalt: Mitteilungsgehalt: P; Geschehenstyp: Handlung; Zeitbezug: zukünftig; Rollenbezug: Hörer; Kommunikative Einstellung von S; Propositionale Einstellung von S: S will: P; S kennt: korrekte Ausführung von P; Sprecherabsicht: S will: H tut korrekt: P; Vorannahmen von S: Im Interesse von H: korrekte Ausführung von P. Below this, it lists related verbs: anleiten · anweisen · instruieren; einweisen · anlernen · anweisen · einarbeiten · einführen · instruieren; unterweisen · beibringen · unterrichten. The second section, titled "Bezugssituationstyp: Direktive.auffordern.anordnen", lists: Propositionaler Gehalt: Mitteilungsgehalt: P; Geschehenstyp: Handlung; Zeitbezug: zukünftig.

Fig. 2: Resource situation type “Direktive.auffordern.anleiten” in the online presentation

The explorative access to verb classes makes *Kommunikationsverben* a useful instrument for university students interested in speech act theory and/or speech act verbs.

6. Extended Search Options

Apart from the explorative access via the list of main verbs classes, *Kommunikationsverben* provides its users with two more advanced search options: a search for situation types and the verbs matching them as well as a search for verbs with particular lexical features. Both search options are provided via the menu item “Erweiterte Suchen” (‘extended search options’).

6.1 Searching for Situation Types

By selecting the option “Paradigmen” (‘lexical fields’) under the menu item “Erweiterte Suchen” (‘extended search options’), the user is presented with an input mask, which he/she may use to “create” any situation type he/she can think of and search for the verbs which may be used to refer to it. For example, to create a situation type in which a speaker tells a hearer that he/she does not approve of a past action of that hearer, the following values for the specifications of the different types of speaker attitudes and of the properties of the propositional content may be entered:

Propositional Content (P)	
Event Type	action
Temporal Reference	past
Agent	H
Speaker Attitudes	
Propositional Attitude	S considers: P bad
Intention	S wants: H recognise: S considers: P bad
Presuppositions	P is the case

Table 4: Search for verbs used to refer to situations in which S tells H that he/she disapproves of a past action of H.

After activating the search, the user is presented with the *vorwerfen*-Paradigma, i.e. the lexical field comprising the verbs *vorwerfen*, *vorhalten* (both: ‘to reproach’/‘to blame’) and *zurechtweisen* (‘to reprimand’).

6.2 Searching for Verbs with Particular Lexical Features

Verbs with specific lexical features may be searched for by selecting the option “Lexikalische Merkmale” (‘Lexical features’) under the menu item “Erweiterte Suchen” (‘extended search options’). A user interested in the use of, say, communication verbs in the double object construction may select the options ‘H: optional/obligatory’ and ‘P: optional/obligatory’ from the section “Semantic Roles”, and the options ‘H: NP<dative>’ and ‘P: NP<accusative>’ from the section “Argument Structure” in the input mask to search for communication verbs which

appear in constructions of that type. A list of corresponding verbs appears to the right of the input mask.

Any of the lexical features mentioned in 3.2 or any combination of them may be searched for by selecting the relevant features from the input mask.

The searches for situation types and for lexical features may prove to be particularly interesting to linguists interested in lexicalisation phenomena (lexicalisation patterns, lexical gaps) or issues related to argument structure, respectively. Because of the inclusion of these two advanced access structures, *Kommunikationsverben* is an example of how the possibilities of the digital medium may be used to extend and accelerate access to the information provided by the print reference work. It is also likely to be of interest to university students learning German as a foreign language. These potential users may employ *Kommunikationsverben* to find out which verbs may be used to refer to a particular type of situation in German as compared to their native language, and to learn about the argument structure properties of these verbs from a contrastive perspective.

7. References

- Baumgärtner, K. (1977). Lexikalische Systeme möglicher Performative. *Zeitschrift für Germanistische Linguistik*, 5, pp. 257-276.
- Boas, H. C. (2010). The syntax-lexicon continuum in Construction Grammar: A Case study of English communication verbs. *Belgian Journal of Linguistics*, pp. 54-82.
- Bierwisch, M. & Lang, E. (1989). Somewhat Longer – Much Deeper – Further and Further: Epilogue to the Dimensional Adjective Project. In M. Bierwisch & E. Lang (eds.): *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*. Springer Series in Language and Communication; 26. Berlin: Springer, pp. 471-514.
- Bierwisch, M. & Schreuder R. (1992). From Concepts to Lexical Items. *Cognition*, 42, pp. 23-46.
- DeReKo - Das Deutsche Referenzkorpus.
<http://www1.ids-mannheim.de/kl/projekte/korpora/>
- FrameNet. <https://framenet.icsi.berkeley.edu/fndrupal/>
- Harras, G. (2001). Performativität, Sprechakte und Sprechaktverben. In G. Harras (ed.): *Kommunikationsverben: Konzeptuelle Ordnung und Semantische Repräsentation*. Studien zur deutschen Sprache; 24. Tübingen: Narr, pp. 11-32.
- Harras, G. (2007). Einleitung: Paradigmen und lexikalische Strukturen von Sprechaktverben. In G. Harras, K. Proost & E. Winkler: *Handbuch deutscher Kommunikationsverben, Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: Walter de Gruyter, pp.

11-24.

- Harras, G., Winkler, E., Erb, S. & Proost, K. (2004). *Handbuch deutscher Kommunikationsverben. Teil I: Wörterbuch*. Schriften des Instituts für Deutsche Sprache ; 10.1. Berlin/New York: Walter de Gruyter.
- Harras, G., Proost, K. & Winkler, E. (2007). *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: Walter de Gruyter.
- Kommunikationsverben. In: OWID (Online Wortschatz- Informationssystem Deutsch). <http://www.owid.de/docs/komvb/start.jsp>
- Lang, E. (1994). Semantische vs. konzeptuelle Struktur: Unterschneidung und Überschneidung. In M. Schwarz (ed.): *Kognitive Semantik: Ergebnisse, Probleme, Perspektiven*. Tübinger Beiträge zur Linguistik; 395. Tübingen: Narr, pp. 9-40.
- OWID – Online Wortschatz-Informationssystem Deutsch. <http://www.owid.de/>
- Proost, K. (2006). Speech Act Verbs. In K. Brown (Ed.-in-Chief): *Encyclopedia of Language & Linguistics*. 2nd. ed. Vol. XI. Oxford: Elsevier, pp. 651-656.
- Proost, K. (2007). *Conceptual Structure in Lexical Items: The Lexicalisation of Communication Concepts in English, German and Dutch*. Pragmatics & Beyond New Series ; 168. Amsterdam/Philadelphia: Benjamins.
- Searle, J. R. & Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge, UK: Cambridge University Press.
- Vanderveken, D. (1990). *Meaning and Speech Acts I: Principles of Language Use*. Cambridge, UK: Cambridge University Press.
- Verschueren, J. (1980). *On Speech Act Verbs*. Amsterdam: John Benjamins.
- Verschueren, J. (1985). *What people say they do with words: Prolegomena to an empirical-conceptual approach to linguistic action*. Norwood, NJ: ALEX.
- Wierzbicka, A. (1987). *English speech act verbs: A semantic dictionary*. Sydney: Academic Press.

Between Grammars and Dictionaries: a Swedish Constructicon

**Emma Sköldberg, Linnéa Bäckström, Lars Borin,
Markus Forsberg, Benjamin Lyngfelt,
Leif-Jöran Olsson, Julia Prentice, Rudolf Rydstedt,
Sofia Tingsell, Jonatan Uppström**

Department of Swedish, University of Gothenburg,
PO Box 200, SE-405 30 Gothenburg, Sweden

E-mail: {emma.skoeldberg; linnea.backstrom; lars.borin; markus.forsberg;
benjamin.lyngfelt; leif-joran.olsson; julia.prentice; rudolf.rydstedt;
sofia.tingsell; jonatan.uppstrom}@svenska.gu.se

Abstract

This paper introduces the Swedish Constructicon (SweCxn), a database of Swedish constructions currently under development. We also present a small study of the treatment of constructions in Swedish (paper) dictionaries, thus illustrating the need for a constructionist approach, and discuss three different methods used to identify potential constructions for inclusion in the Constructicon. SweCxn is a freely available electronic resource, with a particular focus on semi-general linguistic patterns of the type that are difficult to account for from a purely lexicographic or grammatical perspective, and which therefore have tended to be neglected in both dictionaries and grammars. Far from being a small set of borderline cases, such constructions are both numerous and common. They are also quite problematic for second language acquisition as well as LT applications. Accordingly, various kinds of multi-word units have received more attention in recent years, not least from a lexicographic perspective. The coverage, however, is only partial, and the productivity of many constructions is hard to capture from a lexical viewpoint. To identify constructions for SweCxn, we use a combination of methods, such as working from existing construction descriptions for Swedish and other languages, applying LT tools to discover recurring patterns in texts, and extrapolating constructional information from dictionaries.

Keywords: lexicography, construction, constructicon, Swedish, FrameNet, language technology

1. Introduction

Linguistic patterns that are too specific to be treated as general rules and too general to be tied to specific words are peripheral from both a grammatical and a lexicographic point of view. Hence, such patterns, which may be called constructions (cx), have (traditionally) tended to be neglected in grammars as well as dictionaries. Some typical Swedish examples are conventionalized time expressions like “[minuttal] i/över [timal]” ‘[minutes] to/past [hour]’ and semi-prefab phrases such as “i ADJEKTIV-aste laget” ‘in ADJECTIVE-superlative the-measure’. The latter cx basically means ‘too much’ of the quality expressed by the adjective: *i hetaste laget* ‘too hot for comfort’, *i minsta laget* ‘a bit on the small side’ and *i senaste laget* ‘at the last moment’.

These examples are partially schematic multi-word units, i.e. structures where at least one component is lexically fixed and at least one represents a morpho-syntactic category. Accounting for such constructions is a main priority for the Swedish Constructicon (SweCxn) currently under development. The resource is based on principles of Construction Grammar and developed as an addition to the Swedish FrameNet (SweFN). It is integrated with other freely available resources in Språkbanken (the Swedish Language Bank) by linked lexical entries (Lyngfelt et al., 2012). In most respects, the Swedish Constructicon is modelled on its English counterpart in Berkeley, and, thus, mostly adhering to the FrameNet format (see Fillmore, 2008; Fillmore et al., 2012). The SweCxn project is highly cross-disciplinary, involving experts on (construction) grammar, language technology, lexicography, phraseology, second language research, and semantics at the Department of Swedish, University of Gothenburg.

In the next section, the notion of constructions will be discussed. In section 3 we present a minor study of the treatment of cx in Swedish paper-dictionaries. The Swedish Cxn is presented (briefly) in section 4, followed by a presentation of possible methods to find new cx in section 5. Finally, in section 6, there is an outlook, addressing matters of cross-linguistic applicability.

2. Constructions

The type of cx mentioned above is far from being a small set of borderline cases that can simply be disregarded. On the contrary, semi-productive, partially schematic multi-word units are both numerous and common, arguably “of a comparable order of magnitude to the number of words” (Jackendoff, 2007: 57). Furthermore, these kinds of patterns have been shown to be highly problematic, e.g. in relation to L2 acquisition (cf. Ekberg, 2004; Wray, 2008; Prentice & Sköldböck, 2011) and language technology (LT; see Sag et al., 2002).

For the last few decades, however, the study of constructions is on the rise, due to the development of Construction Grammar (CxG; Fillmore et al., 1988; Goldberg, 1995; Hoffmann & Trousdale, 2013, and others) and other cx-oriented models. Furthermore, cx have also been gaining increased attention from some lexicalist perspectives, e.g., Head-Driven Phrase Structure Grammar (HPSG; Pollard & Sag, 1994), especially through the CxG-HPSG hybrid Sign-Based Construction Grammar (SBCG; Boas & Sag, 2012). Still, these approaches have mostly been applied to specific cx, or groups of such. To date, there are few, if any, large-scale constructional accounts.

Within CxG, cx are typically defined as conventionalized pairings of form and meaning/function. This definition can be compared to what in other linguistic contexts are called *signs* (cf. Saussure) or *symbolic units* (cf. Langacker). Linguistic patterns of any level, or combination of levels, from the most general to the most

specific, may be considered cx. Hence, instead of a sharp distinction between lexicon and grammar with a problematic grey area, one can see language as a network of cx along a continuum from extremely specific, lexically filled and fixed items to very general syntactic patterns (Fillmore et al., 1988; Lyngfelt & Sköldbberg, forthcoming).

Goldberg (2006) points out that one can often identify someone as a non-native speaker of a given language:

[...] because much of the phrasing used and combination of lexical choices are non-conventional, even if fully grammatical. It is in fact often the case that one particular formulation is much more conventional than another, even though both conform to the general grammatical patterns in a language

(Goldberg, 2006: 54; cf. Pawley & Syder, 1983).

Goldberg exemplifies the above with conventionalized time expressions that are language specific and have to be acquired through input like other lexical items, since a learner who has never met them before has no means to build them from scratch based on his knowledge of the L2-system (Goldberg, 2006: 54f.). A Swedish example of such language-specific properties concerns what Fillmore (2008) calls *day-level temporal units*. Although time adverbials are usually expressed as PPs, in Swedish as in many other languages, this is not the case if the time is a date: *Hon åker (*på) 7 maj* 'She will leave on May 7th', as opposed to *Hon åker på måndag* 'She will leave on Monday'. In L2 Swedish, incorrect inclusion of the preposition is not uncommon: **Jag är född på 2 mars* 'I was born on March 2nd' (cf. Fillmore, 2008).

So far, the project has to a large extent focused on partially schematic cx, where at least one of the component parts is lexically fixed. Such cx are somewhat similar to fixed multi-word expressions and are fairly close to the lexical end of the cx continuum (cf. Lyngfelt & Forsberg, 2012).

Of general theoretical interest are argument structure cx, which concern matters of transitivity, voice, and event structure, and are at the heart of discussions on the relationship between grammar and lexicon. Argument structure is usually assumed to be determined by lexical valence, but there are good reasons to assume that syntactic constructions also play a role (Goldberg, 1995).

Consider, for instance, the (Swedish) *reflexive resultative* cx (Jansson, 2006; Lyngfelt, 2007), as in *äta sig mätt* 'eat oneself full', *springa sig varm* 'run oneself warm', and *byta sig ledig* 'swap oneself free' (cf. Hanks, 2008). Its basic structure is Verb Reflexive Result, where the result is typically expressed by an AP, and its meaning can be described as 'achieve result by V-ing'. (Hence, an expression like *känna sig trött* 'feel tired' is not an instance of this cx, since it does not mean 'get tired by feeling'.) This pattern is applicable to both transitive and intransitive verbs, even when it conflicts with the verb's lexical valence patterns. Notably, the reflexive object does not correspond to a typical object role; for example, the *sig* in *äta sig*

mätt does not denote what is eaten. Such cx raise theoretically interesting questions regarding to what extent argument structure is lexically or constructionally determined.

3. Constructions in Swedish Dictionaries

To what extent are these and other constructions accounted for in dictionaries? Studies of the treatment of constructions in Swedish, or Nordic, dictionaries are few in number. However, Farø & Lorentzen (2009) have shown that the coverage of partially schematic cx is not satisfactory in Danish dictionaries. Many dictionaries tend to favor colorful fixed phrases, e.g. idioms, at the expense of more anonymous cx with variable component slots. This is a problem, as many such cx are arguably more relevant for language learners than, for example, the idioms which by comparison are used quite rarely (Farø & Lorentzen, 2009). The authors also observe that the dictionaries have problems in reproducing the productivity of these structures.

We studied to what extent, and how, about ten partially schematic cx already included in the SweCxn are treated in dictionaries of contemporary Swedish. More precisely, we examined the following four comprehensive monolingual paper dictionaries:

- *Natur och Kulturs Stora Svenska ordbok* (2006)
- *Svensk ordbok utgiven av Svenska Akademien* (2009)
- *Bonniers svenska ordbok* (2010)
- *Svenskt språkbruk. Ordbok över konstruktioner och fraser* (2003).

The three books mentioned first are general dictionaries and the fourth is a phraseological dictionary. Our study supports the results by Farø & Lorentzen (2009; cf. Lyngfelt & Sköldberg, forthcoming). A typical example is the treatment of the already mentioned time expression [minuttal] i/över [timal] '[minutes] to/past [hour]'. In one of the general dictionaries, *Natur och Kulturs Stora svenska ordbok*, you find the cx in two places: in the articles *i* 'to' and *över* 'past'. However, the other general dictionaries only account for one of the corresponding time expressions, the one with *i* 'to'. Surprisingly, in *Svenskt språkbruk*, the phraseological dictionary, this frequently used conventionalized expression is not mentioned at all. There might, of course, be many underlying causes behind this scanty and inconsistent treatment of this particular cx, but one plausible explanation is that the only lexically fixed components (*i* and *över*) are highly frequent prepositions. Hence, the cx can be hard to discern in corpora. In addition, the cx typically occurs in speech and not in newspaper texts (on which Swedish dictionaries are primarily based). Moreover, in many texts, time information is usually given in another way: you write, e.g. *06.15* or *18.15* instead of *kvart över sex* 'a quarter past six'. But even if this cx were accounted for in a more adequate way in the dictionaries, from a user's point of view the cx would still be difficult to find in a paper dictionary, as preposition articles like these

are extensive and hard to grasp. In that sense, an e-dictionary with more search options evidently has many advantages.

The other example mentioned in the introduction, “i ADJEKTIV-aste laget” ‘in ADJECTIVE-superlative the-measure’, has two lexical parts, the preposition *i* and the noun *lag*. In all the dictionaries the cx is treated in the noun entry. Due to space limitations, in the following we present only one of these entries, the one from *Bonniers svenska ordbok* (2010):

- (1) ¹**lag**³ (i många uttryck) *i längsta (största, minsta, kortaste, etc.) laget* nästan för lång osv. [...]
¹**lag**³ (in many expressions) *rather a bit long (big, little, short, etc.)* almost too long and so on [...]

In the case of “i ADJEKTIV-aste laget”, all lexicographers have tried to account for the productivity of the cx, but in different ways. In (1), the fact that this sense of *lag* appears in many expressions is commented. Similar comments, or other markers indicating the same thing, are also found in the other dictionaries. Furthermore, four different adjectives are given, i.e., *längsta* ‘longest’, *största* ‘biggest’, *minsta* ‘smallest’ and *kortaste* ‘shortest’ followed by an “etc.” indicating that these adjectives can be replaced by others.

The word combinations in the dictionary examples are without doubt recurrent in the corpora at Språkbanken (of more than 1 billion tokens). Many also appear in the other dictionaries. Still, they are not totally representative of authentic language as all the adjectives refer to size. Other recurrent adjectives in the corpora are, e.g. *dyr* ‘expensive’, *tidig* ‘early’, *sen* ‘late’ and *tuff* ‘tough’ which are all missing in the dictionaries. However, in a traditional dictionary it is very difficult to give exhaustive information in this respect, as the productivity cannot be captured on a lexical basis.

Finally, only the first example in the dictionary article above, i.e. *i längsta laget* ‘rather a bit long’, is explicitly explained. The idea is that the user can figure out the meaning of the other variants by analogy. One of the intended user groups of this particular dictionary (L1-speakers) might be able to understand this information. However, for L2-learners on all levels, it can be a hard nut to crack.

To conclude, our study of the treatment of partially schematic cx in dictionaries of Swedish is limited, but it supports the results presented by Farø & Lorentzen (2009). Even if cx with specific lexical parts, such as “i ADJEKTIV-aste laget”, to some extent are described in the dictionaries, many of them are missing. This also applies to dictionaries which normally account for many phraseological units, at least idioms. And even if all the dictionaries try to bring out the productivity of the cx, they cannot totally catch this characteristic feature due to the fact that they have lexical items as a starting point.

As shown by these examples, cx often combine features from several linguistic levels. They may be characterized by prosodic, morphological, lexical, syntactic, semantic, pragmatic features, in different combinations. How can such patterns be accounted for? Should cx of this type be described in dictionaries at all? Or do they “belong” to the grammars/grammarians? The questions bring to the fore an observation made by e.g. Hannesdóttir & Ralph (2010), who discuss the fact that lexicography and grammar description to a great extent are different activities. Patterns with both lexical and grammatical properties cannot be described in an adequate way as long as lexicography and grammar are kept strictly apart. Consequently, according to the authors, lexicographers and dictionary writers should cooperate more and jointly ensure that what is lacking in the one resource is covered in the other. Naturally an increased cooperation would be beneficial in many respects. However, such a solution is still based on a binary distinction between lexicon and grammar. Each linguistic phenomenon must be attributed to the one or the other – or perhaps both. In this paper we present a different approach, where the grammatical and lexical features are combined in the same description.

4. The Swedish Constructicon

The Swedish Constructicon (SweCxn) is a usage based database, where all cx descriptions are grounded in annotated corpus examples. At present, it consists of about 100 cx, still basically a pilot constructicon, but it is growing continually. Eventually, SweCxn is meant to be primarily a large-scale resource for linguistic research and language technology applications. In a longer perspective, the SweCxn should also be applicable in educational settings, not least for learners of Swedish as a second language. Today, the focus is on collecting the most essential linguistic information about a large number of cx often ignored by traditional lexicography and grammar, but the system is designed to be able to handle any kind of cx as the term is understood in Construction Grammar, including ordinary words, parts of speech, etc.

A typical example of cx currently in SweCxn is the so called *reflexiv resultativ* ‘reflexive resultative’ (cf. section 2 above), where the use of a reflexive pronoun adds a valency bound adjective phrase expressing the result of the action, as in *Han sprang sig varm* ‘He ran himself warm’ and *Kornet och havren får frysa sig mogen* ‘The barley and the oats may freeze themselves ripe’. From a structural point of view, the cx consists of a verb, a reflexive pronoun and an adjective phrase. Seen as a whole, it is a multi-word verb with the reflexive pronoun as a fixed, construction-evoking element. The verb, the reflexive pronoun and the adjective phrase are parts of the cx itself, the subject is also important but it is not a part of the construction proper. The adjective phrase expresses the Result while the subject and the reflexive pronoun may be an Agent or a Theme (according to the system of semantic roles employed in SweCxn). This information is captured in the following way in the entry for *reflexiv resultativ* in SweCxn:

Name: *reflexiv_resultativ*
Category: vbm ('multi-word verb')
Structure: vb refl AP
Construction evoking element: refl
Internal construction elements:
 role: name=Activity cat=vb
 role: cx=refl name=Actor
 role: cx=refl name=Theme
 role: name=Result cat=NP
External construction elements:
 role: name=Actor cat=NP
 role: name=Theme cat=NP

The set of labels used for the category and the structure is quite large, since different cx require different granularity. An element may belong to a very general phrase type like XP or NP but also specific lexical items (possibly in a certain inflectional form), with NPdef etc., in between. Truly fixed elements are noted as construction-evoking elements, but it is also useful to list **common words** and word combinations merely typical for a cx (cf. *collostructural elements*, Stefanowitsch & Gries, 2003). The list for *reflexiv_resultativ* is {*äta* 'eat': *mätt* 'full'}, {*supa* 'drink': *full* 'drunk'}, {*skrika* 'scream': *hes* 'hoarse'} and *springa* 'run'.

Construction elements are defined by a list of feature value pairs. There is no set of features fitting all construction elements, so it is not meaningful to require all of them to have the same features defined. The format does not imply that all possible construction elements are instantiated, which is why the external element and the reflexive have two alternatives with different definitions.

Semantic roles are described in two ways resembling Goldberg's (1995) argument roles and participant roles. Argument roles are typically small sets of general roles useful for describing general semantic features whereas participant roles give a local description of the frames of specific (lexical) items. Agent is an argument role and Eater is a participant role. But the neat distinction between argument and participant roles becomes less clear when dealing with cx in the continuum between the purely grammatical and lexical. In practice, this means that the set of semantic roles needed for describing general features with sufficient precision becomes larger than what is needed for arguments in traditional syntax. The set of general roles employed in SweCxn consists of 33 primitive roles augmented by some modifications and a mechanism for combining roles, e.g. Agent-Source.

General roles are noted explicitly whenever appropriate, as shown above. They are also used as the default name for the construction element. Local, frame specific roles are assigned indirectly when a construction **evokes** a FrameNet frame, e.g. the entry for *reflexiv_resultativ* is declared to evoke the frame Causation_scenario.

The meaning of a construction and how the construction elements contribute to it are described in the **definition**, in the case of *reflexiv_resultativ*:

Definition: [Någon]_{Actor} eller [något]_{Theme} utför eller undergår [en aktion]_{Activity} som leder (eller antas leda) till att [aktören]_{Actor} / [temat]_{Theme}, uttryckt med reflexiv, uppnår ett [tillstånd]_{Resultat}.

Definition: [Someone]_{Actor} or [something]_{Theme} performs or undergoes [an action]_{Activity} which leads to (or is assumed to lead to) the [actor]_{Actor} / the [theme]_{Theme}, expressed by a reflexive pronoun, reaching a [state]_{Result}'

The format for the definitions is inspired by ordinary dictionary type definitions but there are striking differences. One is that the definitions are annotated in almost the same way as the corpus examples included in the entry. The only difference is that the cx itself is delimited in the examples, as in:

[Kornet och havren]_{Theme} får [[frysa]_{Activity} [sig]_{Theme} [mogen]_{Result}]_{resultativ_reflexiv}
 '[The barley and the oats]_{Theme} may [[freeze]_{Activity} [themselves]_{Theme}
 [ripe]_{Result}]_{resultativ_reflexiv}'

Another difference between definitions in dictionaries and in SweCxn is that one does not expect explicit information in a dictionary definition about how parts of the meaning are expressed, e.g. that the theme is expressed by a reflexive pronoun. But there are also deep similarities. One is that readability for humans gets a higher priority than tractability for computers. Another, not apparent from the definition of *reflexiv_resultativ*, is the use of dictionary type modifications as *typically*, *also* and *etc.* This makes it relatively easy to write definitions which are reasonably nuanced and easy to understand. The price is that further formalization will be required to make some information in the definitions useful for technical systems, but that is probably a price worth paying to facilitate the collection of the information in the first place.

But it is worth noting that SweCxn is a formally well defined system in most respects. All names of semantic roles, lexical units etc., are declared or defined either within SweCxn proper or imported in an orderly way from external resources, such as FrameNet or the lexical resource SALDO at Språkbanken. The cx are also ordered in an inheritance hierarchy so that more specialized cx, e.g. *jämförelse.likhet* [*comparison.equality*] and *jämförelse.olikhet* [*comparison.inequality*] inherit from the more general *jämförelse* [*comparison*] in order to increase consistency and maintainability.

5. Data and Methods

Since no comprehensive collection of cx descriptions has ever existed for Swedish, an important methodological question for the project is to discover those cx that have not been described before. To identify cx for SweCxn, we use a combination of methods, such as working from existing cx descriptions for Swedish and other

languages (section 5.1), applying LT tools to discover recurring patterns in texts (5.2), and extrapolating constructional information from dictionaries (5.3).

5.1 Digging where we stand

The natural starting point for SweCxn has been existing cx analyses, for Swedish and for other languages. These analyses include quite a few term papers by our own students, produced over the years in relation to CxG courses and earlier CxG projects. The typical CxG paper presents an in-depth analysis of a particular type of cx. From there, we can a) make simplified analyses to include in SweCxn, and b) trace other cx with related properties. On the basis of the initial, familiar cx, we have developed preliminary standards for SweCxn descriptions. It should be noted in this context that we always consult corpora before arriving at a SweCxn account, even when the cx in question has been described by others.

Cx descriptions for other languages provide a more indirect source of inspiration. Each cx in another language raises the question what more or less corresponding patterns exist in Swedish. However, cx are essentially language specific, and even when similar cx occur in different languages, they cannot be presumed to be identical. The SweCxn entries must always be based on Swedish data, but the foreign cx provide hypotheses to explore.

Of particular interest are cxn resources for other languages. There is a small cxn for English (Fillmore et al., 2012), and cxn projects are under way for Japanese (Ohara, 2012) and Brazilian Portuguese (Torrent et al., 2013). In this case, the cxn entries are not only a source of inspiration; we also wish to establish correspondences for future cross-linguistic cxn applications. Such applications, however, require compatible description formats for the cxn resources involved. We will return to this issue in the final section.

As a first step in this direction, we conducted an inventory of the entries in the English cxn at Berkeley (BCxn), investigating to what extent there are corresponding Swedish cx for each of them (Bäckström et al., 2013b). BCxn consists of 50 complete and 23 incomplete cx entries. Out of the 50 full cxn entries, we established 37 one-to-one correspondents. In five cases, one BCxn entry corresponds to two Swedish cx, whereas the remaining eight entries lack satisfactory matches in Swedish.

As might be expected, more general and abstract cx are typically among the closest cx equivalents, whereas more specific idioms tend to differ to a greater extent. Formal differences between corresponding cx typically concern grammatical markers for number, agreement, definiteness, etc., and relational expressions within the cx. For instance, consider the following pair of examples of corresponding Rate cx in English and Swedish:

- (2) a. twice an hour
- b. *två gånger i timmen*
‘two times in hour-DEF’

As shown in (2), the denominator is headed by an article in English but by a preposition (*i* / ‘in’) in Swedish, and its complement is indefinite in English but definite in Swedish. (In addition, the word *twice* corresponds to a phrase, *två gånger* / ‘two times’, but this last difference is not a property of the respective rate *cx per se*.) Functionally, however, the two Rate *cx* are basically equivalent, although their distribution may differ somewhat. In summary, the comparison with BCxn both provided SweCxn with a set of *cx* entries and may serve as a first step towards multilingual constructicography.

5.2 Cx-candidates via corpora

One of the goals of SweCxn is to develop tools for automatic identification of constructions in authentic texts. This is a highly desirable research objective in itself, with potential uses in a number of LT applications. In addition, the same methods provide the project with a heuristic tool. By automatically extracting various kinds of regularities in texts, we may discover patterns that might otherwise have been overlooked. This especially concerns seemingly insignificant constructions that do not stand out against the context the way spectacular idioms do. The resulting findings are treated as *cx* candidates, a subset of which may be considered actual *cx* after manual evaluation (see Bäckström et al., 2013a).

The general setting for our experiment is the resource infrastructure of Språkbanken, a modular set of resources and tools in the form of web services for accessing, browsing, editing and automatically annotating resources. The two facets of the infrastructure most relevant for the present purposes are the corpus infrastructure Korp (Borin et al., 2012b) and the lexicon infrastructure Karp (Borin et al., 2012a).

The data source for the experiment is SUC 2.0, a balanced text corpus for Swedish consisting of 1.17M tokens that have been manually annotated with lemmas and MSDs (morpho-syntactic description). SUC was selected in order to avoid annotation errors confounding the experiment results, but the experiment can be (and has been) run on any of the more than hundred corpora of Språkbanken that have been automatically annotated with the same information.

The experiment is based on the work on StringNet (Tsao & Wible, 2009; Wible & Tsao, 2010, 2011), where the notion of *hybrid n-gram* plays a central role. A hybrid *n-gram* is a generalization of an *n-gram* where not only the word forms are included in the process, but also the information from the annotation layers. If we limit ourselves to lemmas and part-of-speech, which is the case for this experiment, then the 2-gram *Hur är* ‘How is’ would generate four *cx* candidates: *hur vara* ‘how be’, *hur VB* ‘how VB’, *HA vara* ‘HA be’, and *HA VB*.

Focusing on the discovery of partially schematic constructions, we discarded all candidates that are fully schematic or fully lexical, i.e., consisting of only PoS tags (e.g., *HA VB*) or lemmas (e.g., *hur vara* ‘how be’). Moreover, we removed all hybrid n-grams containing punctuation marks and/or words marked as foreign. They are not necessarily uninteresting, but since they did introduce a lot of noise in the candidate list, we decided to remove them. For SUC 2.0 with 2-, 3- and 4-grams we ended up with 16M hybrid n-grams of which 8.8M were unique.

The next step was to rank all hybrid n-grams, which can be done with a wide range of association measures. We have followed StringNet in using point-wise mutual information (PMI). PMI has a known shortcoming in these kinds of experiments – it has a preference for the low-frequency items – which can be remedied by multiplying PMI with the absolute frequency. This does not solve another problem, however, which is boilerplate text, e.g., “For subscription enquiries e-mail:...”. But with a small modification, instead of counting hybrid n-grams, we count UIF (unique instance frequency), which is the number of unique n-grams underlying the target hybrid n-gram, we can counteract that problem too.

There was still one more problem that needed to be solved: since the bulk of the hybrid n-grams are subsets of other hybrid n-grams, we first arrived at a ranking list with massive redundancy. This was solved, in the same spirit as StringNet’s vertical/horizontal pruning (Tsao & Wible, 2009; Wible & Tsao, 2010), by removing all hybrid n-grams that were subsets of other hybrid n-grams with a higher PMI-UIF. A hybrid n-gram is considered a subset of another if it occurs as a subsequence that is either equal or consisting of non-conflicting items sharing the same part-of-speech; e.g., *vara_{VB}* is considered equal to *VB*.

Some sample candidates are given in Figure 1. The hybrid n-grams are linked to the Korp interface to enable inspection of their instances in the corpus. We also see the most frequent instance, followed by the absolute frequency, relative frequency, and the PMI-UIF.

vara_{VB} ute_{AB} och_{KN} VB	<i>är ute och letar (3)</i>	15	0.93	52.24
vara_{VB} JJ för_{PP} att_E	<i>är viktiga för att (2)</i>	26	1.61	52.83
stänga_{VB} av_{PL} NN	<i>stängt av motorn (1)</i>	11	0.68	52.25

Figure 1. Some example hybrid n-grams from SUC 2.0 ranked by PMI-UIF

The candidate lists are accessible from here: <<http://spraakbanken.gu.se/eng/resource/konstruktikon/candidates>>. Here you will find other materials as well that have been annotated automatically using the Korp pipeline (see 5.3 below).

The construction candidate list makes it possible to go through a large amount of examples quickly, since every hybrid n-gram is directly linked to the instances in the

corpus. However, it was a difficult task to draw the line between relevant and non-relevant constructions and this is still an ongoing matter of discussion in the project group. Of the 2500 items included in the list, 50 constructions were decided to be relevant construction candidates according to our criteria, i.e., that they are partially schematic and productive multiword units that are “too general to be attributed to individual words but too specific to be considered general rules” (Lyngfelt et al., 2012).

The final list of 50 relevant constructions was extracted in several steps. First, one project member went through the whole list extracting a list of 143 interesting candidates (approximately a day’s work). This list was then, in consultation with the other members of the project group, gradually reduced and the final result of this process was, as mentioned above, 50 cx that were found relevant for entries in the SweCxn. As the main goal was to discover cx that are difficult to find with other methods, the result of 50 is not the whole story: a cx candidate can also inspire descriptions of other similar cx, which is a question of the researchers’ capacity for creative thinking at a given moment in time.

5.3 Cx-candidates from general dictionaries

Currently we are also exploring the possibility of finding relevant cx-candidates within the articles in Swedish definition dictionaries. First of all we are interested in partly schematic patterns not so emphasized but rather indicated by comments like “in many expressions” (cf. section 3 above). Of course, this kind of usage marker is more easily found in e-dictionaries. Unfortunately, there are very few modern electronic definition dictionaries of high quality for Swedish. As a matter of fact, the existing ones are just e-versions of older paper dictionaries, which now have been subject to extensive revisions. Unfortunately, these revised versions are not published electronically (cf. i.e. NEO from 1995–1996 online with the printed SO from 2009; see below).

However, in the SweCxn project we have access to the whole database of the two-volume paper dictionary of Swedish published by the Swedish Academy (2009; henceforward SO). The dictionary, comprising about 65,000 lemmas, is the most comprehensive monolingual dictionary of contemporary Swedish that there is. By advanced search options in the database, we can extract information on different kinds of relatively anonymous word combinations indicated in the microstructure.

For example, the marker “i uttryck” ‘in expression(s)’ is used about 700 times within the SO articles. One cx observed by this method is “[X efter X]” ‘[X after X]’, i.e. a certain lexical item appears just before and after the preposition *efter* ‘after’. SO have tried to capture the cx as a subordinate sense of the word *efter* (‘after’) in the following way:

- (3) **efter** prep. (...) [äv. i uttr. för upprepning] *dag efter dag; mil efter mil; (...)*
 ‘**after** prep. (...) [also in expressions of repetition] *day after day; mile after mile;*
 (...)’

In the dictionary only two examples are given, including the nouns *dag* (‘day’) and *mil* (‘mile’). Furthermore, the information on the semantic and pragmatic characteristics of the cx is very scanty. However, by searching in the corpora of Språkbanken, you get more data on this structure. In the texts the cx is used in a frequent and productive way. The repeated word may be a noun (as in the dictionary examples), but it can also be a numeral (*en, ett*):

- (4) ... *hon dricker glas efter glas*
 ‘... she drinks glass after glass’
- (5) *I brev efter brev utbytte de tankar om kriget*
 ‘In letter after letter they were exchanging their thoughts about the war’
- (6) *De kom allesammans, en efter en*
 ‘They all came, one by one’
- (7) *Också träden försvann, ett efter ett*
 ‘All the trees disappeared, one by one’

Many of the hits (here from a corpus of modern novels) can be paraphrased by ‘many X in succession’, emphasizing the repetition. As indicated by the examples, the cx also infers some kind of process. If the repeated word is a noun referring to time, the cx also expresses extension in time and some kind of continuity. This is the case with *dag efter dag* ‘day after day’ in SO. Other typical examples from the corpora are *kväll efter kväll* ‘evening after evening’, *natt efter natt* ‘night after night’ and *år efter år* ‘year after year’.

In other words, well hidden in the SO-articles you find several partially schematic patterns – like “[X efter X]” “[X after X]” – that could be emphasized and accounted for in a more exhaustive way. In SweCxn this problem can be solved. In that sense, the SweCxn can serve an important purpose towards a more detailed description of different kinds of Swedish word combinations.

In the project we also have access to the about 90,000 editorial examples found in the SO articles. One important function of the examples is, of course, to clarify the meaning(s) of the lemmas in the dictionary. But they also reveal typical usage of the lemmas by specifying constructions and collocations (Svensén, 2009: 285). The examples have been tokenized, lemmatized and PoS-tagged and constitute a corpus of its own in Språkbanken. Using the method described in section 5.2 above, we have also extracted SweCxn candidates from that corpus. On the list one can find, for example, the structure [*var*_{DT} RO NN] which is typically realized in the following ways in the corpus:

- (8) *var tjugonde minut* ‘every twenty minutes’

(9) *var tredje timme* ‘every three hours’

(10) *vart fjärde år* ‘every four years’.

In other words, the method reveals another highly productive cx, which also is a challenge to language learners. First of all, as hinted by the examples, the noun can be composed by any time expression. Secondly, the cx includes a variable ordinal number. Thirdly, the pronoun *var* ‘every’, constituting the only lexically-filled component of the cx, has to agree in gender with the noun. And, once again, the cx is an ordeal to lexicographers; it is hard to place and render adequately in the dictionary as the only lexically-filled component is the unstressed pronoun.

6. Outlook

SweCxn is a resource under development, initially designed to suit the needs of linguistic research and LT application. In a longer perspective, it is meant to also support (second) language pedagogy and eventually be presented in a format adapted to a wider audience. Furthermore, in collaboration with the cxn projects of other languages, we are working towards cross-linguistic applicability.

The latter endeavor is probably best characterized as multilingual constructigraphy. It differs from lexicography in that a cxn must also account for the formal structure of a cx and its constituents. What is expressed by syntax or morphology is highly relevant, whether a certain construction element is an NP or a PP, whether NPs are definite or indefinite, if any particular agreement patterns apply, etc. Such features are language-specific, but must be represented in a way in which the relevant information may be linked across languages.

Since all existing cxn resources are developed in relation to a FrameNet of that language, it is desirable to make the two types of resource compatible from a cross-linguistic perspective as well. In FrameNet, which is essentially a lexicographic resource, all cross-linguistic relations are established through the frames. These are semantic units, which have been fairly successfully applied to different languages, since language-specific idiosyncrasies are instead attributed to the lexical units instantiating the frames in each language (cf., however, Pado, 2007; Friberg Heppin & Toporowska Gronostaj, 2012).

For cx with a meaning roughly equivalent to a frame, the same strategy is a viable option, provided that information about cx internal structure is added; but not all cx correspond to frames. Alternatively, as mentioned in section 5.1 above, some cx might be treated as direct equivalents in different languages, but clearly not all of them: especially not when languages less similar than English and Swedish are taken into account. Hence, a cross-linguistically applicable format for cx descriptions is required. Devising such a format will be a challenge for future constructicon development.

Awaiting that, each cxn should be nonetheless useful as a monolingual resource. SweCxn is still small, compared to a comprehensive dictionary, but it already contains a substantial number of linguistic patterns that would be hard to account for from a lexical viewpoint. Some of these cx are of course relevant for lexicography as well – to the extent that they are lexically entrenched. Their productivity, however, is beyond any resource restricted to lexical entries. An appealing future development would be to integrate the constructicon with an e-dictionary, where the possible entries are no longer limited to lexical units. In such a resource, one could navigate from grammatical constructions to the lexical entries that instantiate them and vice versa. Ideally, a user should only have to enter an expression, and the e-resource would be able to identify the constructional pattern to which it corresponds.

7. Acknowledgements

The research presented here was supported by the Swedish Research Council (grant agreement 2010-6013), by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldbberg, sponsored by the Knut and Alice Wallenberg Foundation.

8. References

- Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Prentice, J. & Sköldbberg, E. (2013a). Automatic identification of construction candidates for a Swedish constructicon. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013*. Linköping Electronic Conference Proceedings 88, pp. 2-11.
- Bäckström, L., Lyngfelt, B. & Sköldbberg, E. (2013b). Constructions in contrast. Approaching Swedish correspondents to the entries in the Berkeley FrameNet Constructicon. *International FrameNet Workshop 2013 (IFNW-13)*, Berkeley, CA.
- Boas, H. C. & Sag, I. A. (eds.) (2012). *Sign-Based Construction Grammar*. Stanford: CSLI Publications.
- Bonniers svenska ordbok* (2010). (10 ed.) Stockholm: Bonniers.
- Borin, L., Forsberg, M., Olsson, L.-J., & Uppström, J. (2012a). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*. Istanbul. ELRA, pp. 3598-3602.
- Borin, L., Forsberg, M., & Roxendal, J. (2012b). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*. Istanbul. ELRA, pp. 474-478.
- Ekberg, L. (2004). Grammatik och lexikon i svenska som andraspråk på nästan infödd nivå. I: Hyltenstam, K. & Lindberg, I. (eds.), *Svenska som andraspråk –*

- i forskning, undervisning och samhälle*. Lund: Studentlitteratur, pp. 259-276.
- Farø, K. & Lorentzen, H. (2009). De oversete og mishandlede ordforbindelser – hvilke, hvor og hvorfor? *LexicoNordica* 16, pp. 75-101.
- Fillmore, C. (2008). Border Conflicts: FrameNet Meets Construction Grammar. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 49-68.
- Fillmore, C., Kay, P. & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64, pp. 501-538.
- Fillmore, C., Lee-Goldman, R. & Rhomieux, R. (2012). The FrameNet Constructicon I: Boas, H. & Sag, I. (eds.) *Sign-Based Construction Grammar*. Stanford: CSLI, pp. 309-372.
- Friberg Heppin, K. & Toporowska Gronostaj, M. (2012). The Rocky Road towards a Swedish FrameNet – Creating SweFN. In *Proceedings of LREC 2012*, Istanbul.
- Goldberg, A. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago/ London: University of Chicago Press.
- Goldberg, A. (2006). *Constructions at work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Hanks, P. (2008). The lexicographical legacy of John Sinclair. In *International Journal of Lexicography* 21(3), pp. 219-229.
- Hannedóttir, A. H. & Ralph, B. (2010). Explicit och implicit information i tvåspråkig lexikografi. I: Lönnroth, H. & Nikula, K. (eds.), *Nordiska studier i lexikografi* 10. Tammerfors, pp.150-163.
- Hoffmann, T. & Trousdale, G. (eds.) (2013): *The Oxford Handbook of Construction Grammar*. Oxford: OUP.
- Jackendoff, R. (2007). *Language, Consciousness, Culture: Essays on Mental Structure*. Cambridge: MA: MIT Press.
- Jansson, H. (2006). *Har du ölat dig odödlig? En undersökning av resultativkonstruktioner i svenskan* (MISS 57). Inst. f. svenska språket, Göteborgs universitet.
- Lyngfelt, B. (2007). Mellan polerna. Reflexiv- och deponens-konstruktioner i svenskan. *Språk och stil* NF 17, pp. 86-134.
- Lyngfelt, B., Borin, L. Forsberg, M., Prentice, J., Rydstedt, R., Sköldberg, E. & Tingsell, S. (2012). Adding a Constructicon to the Swedish resource network of Språkbanken. *Proceedings of KONVENS 2012 (LexSem 2012 workshop)*, Wien, pp. 452-461.
- <http://www.oegai.at/konvens2012/proceedings/66_lyngfelt12w/>.
- Lyngfelt, B. & Forsberg, M. (2012). *Ett svenskt konstruktikon. Utgångspunkter och preliminära ramar*. (GU-ISS 2012-02) Inst. f. svenska språket, Göteborgs

- universitet. <<http://hdl.handle.net/2077/29198>>.
- Lyngfelt, B. & Sköldberg, E. (forthcoming). Lexikon och konstruktikon – ett konstruktionsgrammatiskt perspektiv på lexikografi. *LexicoNordica* 20.
- Natur och Kulturs Stora Svenska Ordbok* (2006). Stockholm: Natur och Kultur.
- Ohara K. (2012). Japanese FrameNet: Toward construction building for Japanese. *Seventh International Conference on Construction Grammar (ICCG-7)*, Seoul, Korea
- NEO = *Nationalencyklopedins ordbok* (1995-96). Höganäs: Bra böcker.
- Pado, S. (2007). Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames. *Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*. Tartu, Estonia.
- Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: nativelylike selection & nativelylike fluency. In Richards, J. & Smith, R. (eds.) *Language and communication*. London: Longman, pp. 191-221.
- Pollard, C. & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar* Chicago: University of Chicago Press and Stanford: CSLI Publications.
- Prentice, J. & Sköldberg, E. (2011). Figurative word combinations in texts written by adolescents in multilingual school environments. In Källström, R. & Lindberg, I. (eds.) *Young urban Swedish. Variation and change in multilingual settings*. University of Gothenburg: Dept. of Swedish, pp. 195-217.
- Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multi-word expressions: A pain in the neck for NLP. I: Gelbukh, A. (ed.) *Computational Linguistics and Intelligent Text Processing* (Proceedings of CICLING-2002). Berlin: Springer, pp. 1-15.
- SO = *Svensk ordbok utgiven av Svenska Akademien* (2009). Stockholm: Norstedts. Språkbanken <<http://spraakbanken.gu.se/>>.
- Stefanowitsch, A. & Gries S. (2003). Collocations: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2, pp. 209-43.
- Svensén, B. (2009). *A handbook of lexicography. The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.
- Svenskt språkbruk. Ordbok över konstruktioner och fraser* (2003). Utarbetad av Svenska språknämnden. Stockholm: Norstedts Ordbok.
- SweCxn = *Svenskt konstruktikon*.
<<http://spraakbanken.gu.se/swe/resurs/konstruktikon>>.
- Torrent, T., Lage, L. Sampaio, T., Tavares, T. & Matos, E. (2013). Revisiting Border

Conflicts between FrameNet and Construction Grammar: annotation policies for the Brazilian Portuguese Constructicon. *International FrameNet Workshop 2013* (IFNW-13), Berkeley, CA.

Tsao, N.-L. & Wible, D. (2009). A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Boulder. ACL, pp. 51-54.

Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford: OUP.

Wible, D. & Tsao, N.-L. (2010). StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*. Los Angeles. ACL, pp. 25-31.

Wible, D. & Tsao, N.-L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland. ACL, pp. 128-130.

Testing an electronic collocation dictionary interface: *Diccionario de Colocaciones del Español*

Orsolya Vincze, Margarita Alonso Ramos

Universidade da Coruña, Campus da Zapateira s/n, A Coruña 15071, Spain
E-mail: ovincze@udc.es, lxalonso@udc.es

Abstract

This paper describes the results of a usability study that tests the online interface of the *Diccionario de Colocaciones del Español* (DiCE). This dictionary was conceived with the purpose of providing a detailed description of Spanish collocations in accordance with the theoretical guidelines of the Explanatory and Combinatorial Lexicology. Although from the outset of dictionary compilation, accessibility of the DiCE interface has always been taken into account, no usability test has been carried out to see how different target user groups are able to perform with the dictionary. Our aim was to assess the functionality of the different search options offered by the interface, both in terms of their efficiency and the adequacy of presentation from the point of view of the user. As the results of the test show, the overall quality of interaction between users and the dictionary was good, although we have also identified some areas for improvement, which are provided as design recommendations in the concluding part of the paper.

Keywords: usability testing; collocation dictionary; explanatory and combinatorial lexicology, search strategies, log file analysis

1. Introduction

The present paper describes the results of a usability study that tests the online interface of the *Diccionario de Colocaciones del Español* (DiCE, Alonso Ramos, 2004, Alonso Ramos et al., 2010 and Vincze et al., 2011). This dictionary was conceived with the purpose of providing a theoretically well-founded and detailed description of Spanish collocations. However, it was always intended as a useful tool for its users. It is for this reason that, after different modifications of the interface, we decided to carry out a usability test to see how different target user groups are able to perform with the dictionary. Our aim was to assess the different search options offered by the interface both in terms of their efficiency and the adequacy of their presentation from the point of view of the user. In the next section we briefly overview similar dictionary usability studies. Subsequently, we present our own study and the conclusions drawn from the results obtained.

2. Dictionary usability studies

Various aspects of dictionary use are studied. Most studies aim to decipher for which purposes dictionaries are used, what knowledge or abilities dictionary users have or require, or how dictionaries contribute to language learning. Heid (2011) proposes a different approach: the application to electronic dictionaries of *usability testing*, as defined by information science. This line of research implies testing dictionaries at

the level of functionality, much like in the case of other kinds of software tools. Studies that have applied usability testing methodology include Heid and Zimmerman (2012), which compares different types of access to collocations in mock-up dictionary interfaces, and Hamel (2012), which provides a detailed description of a usability experiment with a dictionary prototype concentrating on lexical selection, combination and paraphrase. Jousse et al. (2011) reports briefly on a test performed on a prototype collocation dictionary developed following the same theoretical framework as DiCE (see below), without providing quantitative results.

3. The study

3.1 The interface tested

The *Diccionario de Colocaciones del Español* (DiCE) is an online collocation dictionary of Spanish, which has been designed in accordance with the postulates of the Explanatory Combinatorial Lexicography (Mel'čuk et al., 1995), and is mainly oriented to language production. The DiCE represents collocations as restricted combinations of two lexical elements: the *base*, the element with more semantic weight which is freely selected in language production, and the *collocate*, an element whose selection is conditioned by lexical restrictions imposed by the base. For instance, in the combination *reanudar una amistad* 'renew a friendship', the noun is the base, and it conditions the selection of the collocate verb.

In order to offer dynamic access to the information stored in the DiCE database, in addition to the *dictionary module*, the current user interface incorporates various advanced search options. Each of these was conceived to provide the user with a more direct path of access to a specific type of information. Since the main objective of the usability test was to measure the functionality of the different search options, we provide a brief description of these.

1. *Dictionary module*: This option provides a traditional collocation dictionary type access to combinatorial information. The entry of each lemma contains the subentries of its corresponding lexical units, where collocations are grouped according to their syntactic pattern and semantic content.
2. *Advanced search module*:
 - a. *What does it mean?*: This reception-oriented module provides direct access to the entry of a specific collocation. The user is prompted to introduce a base (e.g. *amistad*) and a collocate (e.g. *reanudar*).
 - b. *Writing aid*: This is a production-oriented module, which allows the user to find collocates of a given base (e.g. *amor* 'love'), corresponding to a specific part of speech, and a meaning (e.g. 'felt for one another'), such as *amor mutuo* 'mutual love'.
 - c. *Direct search*: This option allows finding collocations in DiCE encoded

by a specific Lexical Function (Mel'čuk et al., 1995) (e.g. Sing(remordimiento) = *acceso de* ~ 'fit of remorse').

- d. *Inverse search*: This last module prompts the user to introduce a collocate (e.g. *cumplir* 'fulfill') in order to find the bases with which it can be combined (e.g. *deseo* 'wish', *esperanza* 'expectation').

3.2 The questionnaire

The questionnaire used in the usability test consisted of 13 questions. Participants were instructed to conduct searches on the dictionary interface in order to retrieve the answer for each item, even if they did feel able to provide a solution relying only on their own knowledge. Questionnaire items were designed in such a way that, although in most cases they could be resolved via navigating the dictionary module, the most direct path to obtain an answer was through using the advanced search options. In Figure 1 we show a few sample questions together with the optimal query type to be used. The numbers in brackets indicate the number of questionnaire items corresponding the given query type; note that items indicated as optimally searched by the same query type are not formulated in exactly the same way. Following the usability test itself, a brief post-test questionnaire was administered in order to measure user satisfaction.

<p>What verbs can be used with the lexical unit <i>cariño</i> 2 'affection'?</p> <ul style="list-style-type: none"> ○ optimal query type: Dictionary module/Writing aid (2) <p>What does <i>reanudar la amistad</i> 'renew a friendship' mean?</p> <ul style="list-style-type: none"> ○ optimal query type: What does it mean? (4) <p>Find the adjectives you can use to speak about <i>amor</i> 'love' 'that is felt for one another'</p> <ul style="list-style-type: none"> ○ optimal query type: Writing aid (3) <p>Find the collocates of <i>remordimiento</i> 'remorse' codified by the Lexical Function <i>Sing</i>.</p> <ul style="list-style-type: none"> ○ optimal query type: Direct search (2) <p>Find all collocations with the verbal collocate <i>cumplir</i> 'fulfill'.</p> <ul style="list-style-type: none"> ○ optimal query type: Inverse search (2)

Figure 1: Sample questionnaire items

3.3 Participants

The 26 informants who participated in the study represent four groups of different target user-profiles of DiCE: 1) Eight informants are Spanish university students. They represent a group of native Spanish users with certain language awareness. 2) Nine participants are foreign university students majoring in Spanish. These informants are upper-intermediate or advanced learners of Spanish as L2. 3) Five

informants are teachers of Spanish or English as a foreign language, all of them native speakers of Spanish. 4) Finally, the last five informants are Spanish PhD students of translation studies, all native speakers of Spanish. As a group, they can be considered as language professionals, characterized by an elevated language awareness and considerable expertise in the use of lexicographic tools.

3.4 Procedure

The experiment can be divided into three main phases: an informative session, the usability test proper, and a post-test questionnaire. Previous to completion of the usability questionnaire, the participants received a brief introduction to the concept of collocations, and were given some instructions on the completion of the usability test; however they were not instructed in the use of DiCE. After having received all necessary information, participants completed the usability questionnaire on their home computers. They were asked to provide the IP address of their computer, and the time and date of connection, so that their actions could be tracked in the DiCE website log files.

3.5 Data analysis

For quantitative analysis of the results of the usability test, we adopted the criteria described in e.g. Nielsen (1993). The usability of an interface can be measured along three main aspects: *effectiveness*, *efficiency* and *user satisfaction*. *Effectiveness* of the interaction can be measured through the task outcome, in our case, the participants' performance on the usability questionnaire represented by the number of correct answers provided.

Efficiency of the interaction is measured through task duration and the efforts of the user to accomplish the task, i.e. the degree of interaction with the dictionary interface. In our case, we established three parameters for measuring efficiency: 1) the *net time* required to complete the query in the case of each individual test item; 2) the *effort measure* calculated as the sum of the number of times a specific search option is chosen by the participant, the number of times a search filter is set, and the number of times the participant hits the *Search* button before obtaining the definitive answer for the test item; and 3) *query-type adequacy* based on the search option used to retrieve a correct answer. Here, 3 points were assigned when the participant used the most optimal search option for the question with all filters correctly set; 2 points when they used one of the advanced search options – though not the most adequate one – or when they failed to optimally set some of the search filters; and 1 point when they used the dictionary module in place of another search option which would have provided a more direct access to the information.

While effectiveness and efficiency constitute objective measures, and can be assessed on the basis of participants' answers to the items of the usability questionnaire together with the data obtained from the log files, the third aspect, *user satisfaction*,

being a subjective indicator, is evaluated on the basis of the results of the post-test questionnaire.

4. Results

The mean number of correct answers provided per participant was 9.62, with a standard deviation (SD) of 3.35, out of the total number of 13 questions. Four participants out of the 26 succeeded in finding the correct answer for all questions and 10 participants answered 11 or 12 questions correctly. Two participants only provided one correct answer, before deciding not to continue with the test.

From the efficiency scores (see Table 1), we can conclude that participants who obtained 12 or more correct answers tended to need less time, made less effort per query and simultaneously used the more adequate access path more often than others; a fact that suggests that they can be considered more skillful users. Note that both mean time and effort indicate the difficulty faced as a result of the participants' unfamiliarity with the user interface. Another tendency that can be observed in the data is that often participants who obtained a higher query adequacy score made more effort. This may be a result of these participants tending to experiment more with the different search options available on the DiCE interface, and managing to find the more straightforward ways to access information. Indeed, these participants provided more correct answers than users who tended to employ almost exclusively the more basic traditional dictionary-type access.

	Net time	Net time per test item	Total efforts	Efforts per test item	Query-type adequacy
1-4 corr. ans. (n=3)	28:39	03:39	196.67	15.13	2.33
SD	18:32	00:54	167.41	12.88	1.15
7-9 corr. ans. (n=6)	44:58	03:28	264.83	20.37	1.87
SD	20:14	01:33	107.68	8.28	0.71
10-11 corr. ans. (n=9)	50:18	03:53	355.67	27.36	2.36
SD	28:06	02:10	142.31	10.95	0.52
12-13 corr. ans. (n=8)	25:49	02:06	202.63	15.59	2.60
SD	11:57	00:59	56.61	4.35	0.48
MEAN	39:02	03:12	269.27	20.71	2.32
SD	22:54	01:42	129.17	9.74	0.66

Table 1: Summary of overall task efficiency

The number of participants who managed to find the correct answer, together with efficiency measures for each group of questions representing a specific anticipated optimal query type, provides information on which items of the usability questionnaire were especially problematic (see Table 2).

	Correct answers	Net time	Efforts	Query-type adequacy
Dictionary/ Writing aid (Qs 1, 10)	15.50 SD=0.71	03:26 SD=03:06	17.58 SD=16.84	2.94 SD=0.25
What does it mean? (Qs 2, 4, 11, 13)	23.33 SD=1.53	02:23 SD=03:05	18.15 SD=22.68	2.24 SD=0.96
Writing aid Qs 3, 6, 12)	17.00 SD=5.29	03:10 SD=3:00	20.07 SD=18.96	1.82 SD=0.99
Direct search (Qs 7, 9)	19.50 SD=2.12	03:35 SD=04:14	26.29 SD=22.72	2.23 SD=0.84
Inverse search (Qs 5, 8)	18.50 SD=0.71	04:27 SD=04:22	31.27 SD=25.90	2.78 SD=0.48

Table 2: Summary of effectiveness and efficiency for question groups according to optimal query type

Participants were most successful in answering questionnaire items which were categorized as most suitable for the *What does it mean?* search option. They also needed the least time, and made on average less effort than in the case of most other questionnaire items. The second highest mean of correct answers was achieved in the case of items which were classified as optimally queried using *Direct search*, despite the fact that these involved the use of Lexical Functions, with which participants were not familiar. A slightly lower number of participants answered correctly in the case of the questions which prompted finding collocations in the dictionary starting from the collocate, and could be resolved using the *Inverse search* option. Note that these were the only questionnaire items where participants necessarily had to make use of a specific advanced search option; whereas the answers to all other items could be queried using the *Dictionary module*. Accordingly, in the case of these items, participants spent the highest mean time and made the most effort, while the mean query-type adequacy score is the highest. In the case of items where subjects were expected to use the *Writing aid* option, there was considerable difference between individual questions in terms of the number of correct answers provided. Finally, the two questionnaire items where we considered as optimal access paths both the *Dictionary module* and the *Writing aid* option, are among the questions with the lowest number of correct answers.

Table 3 provides a summary of search options used to obtain correct answers in the case of each questionnaire item. The highlighted squares represent the optimal query type in each case, which was in fact the most frequently used search option for the majority of questionnaire items. However, it can be seen that the advanced search options were generally under-used, especially the *Writing aid*.

As for effectiveness and efficiency according to user profiles, the group of translation students performed best since they obtained the highest number of correct answers (mean = 10.6, SD = 1.67), and took the least time (mean = 28:55, SD = 11:32) to complete the queries. The native university students performed slightly better

concerning the number of correct answers (mean = 10.38, SD = 2.8) than the foreign university students (mean = 10.0, SD = 4.14), whereas the group of native Spanish language teachers seemed to have the greatest difficulty in using the interface (mean = 6.8, SD = 4.21).

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Dictionary module	10	8	9	7		12	5		5	12	6	9	12
What does it mean?		17		16	1						14		10
Writing aid	6		12			7				3	1	2	
Direct search							16		13				
Inverse search					17			19				1	
TOTAL	16	25	21	23	18	19	21	19	18	15	22	11	22

Table 3: Summary of the number of correct answers generated using each query type per question

Information on user satisfaction was collected in a post-test questionnaire. In addition, following each query during the usability test, participants were asked to assess its difficulty on a 1–5 Likert-type scale. In the post-test questionnaire, participants were asked whether they had used DiCE before and whether they used it frequently. The answers to these questions reveal that none of the participants had substantial experience with the dictionary. In contrast, 20 participants answered “yes” when asked whether they would use the dictionary in the future, while the remaining six said “maybe”. Finally, when asked whether they would recommend the dictionary to others, 20 informants said “yes”, three said “maybe”, and the remaining three participants said that they would recommend it but it is not easy to use, or they would recommend only the simpler features. In conclusion, participants’ answers reveal a clear positive attitude towards DiCE, although some have reservations about its ease of use. This last point is also apparent if we observe the difficulty score assigned to questionnaire items. The mean difficulty score assigned by participants is 2.65 (SD = 0.77).

5. Discussion

As we have seen, the items of the usability questionnaire were designed in a way that they encourage users to experiment with the different advanced search options available in the DiCE web interface. However, as the results presented above suggest, subjects most frequently used the *Dictionary module*. The reasons for this are twofold. On the one hand, this access path is offered by default in the web interface, and, in addition, it assists in retrieval of the correct answer in the case of most questionnaire items; consequently when participants managed to find the required information using this feature, they did not subsequently employ any advanced

search options for the task. On the other hand, this module provides a type of access similar to paper dictionaries, which may therefore be more familiar to users. Among the advanced search options, the most frequently and successfully used query type was *What does it mean?*. We believe that this can be accounted for by the way dictionaries are most commonly used: users tend to check a given lexical item (either for its meaning or spelling), but they generally do not search for how to express a specific meaning. Also, note that two of the four questions where this option was indicated as an optimal query type explicitly asked about the *meaning* of collocations, which might have served as a clue for users as to which search option to choose.

A qualitative assessment of individual search options has allowed us to explore what details in particular were problematic from the user's point of view. Most identified problem areas can be referred to as *content related problems*, given that they reflect the informants' difficulties in interpreting the dictionary content and the presentation of lexicographic data. The most prominent of these was a lack of familiarity with the notion of collocations and the specific terminology applied in DiCE. Subjects tended to confuse the elements of a collocation (base and collocate) leading to difficulties in using a number of search options. For instance, in the *What does it mean?* search option the search form requires introducing the base and the collocate in individual search boxes; nevertheless, one participant typed a whole collocation string in the box corresponding to the collocate, while others interchanged the two elements of the collocation instead of writing them in the corresponding search boxes. We also noticed that participants tended to confuse the *Direct search* and the *Inverse search* options, which might be a consequence of the fact that both search forms require the introduction of an element of a collocation (the base or the collocate, respectively).

In DiCE, the approximate meaning of collocations is described via a semantic gloss, which some participants tended to confuse with the collocates themselves. For example, when accessing the lexical entry of a base through the *Dictionary module*, collocates are grouped in such a way that the user is provided with a list of gloss tabs, which must be opened to access the collocates. After using the *Dictionary module*, some participants included glosses in their answers, listing them together with collocates, while a few subjects only listed the glosses themselves, which suggests that they were unaware of the need to open the gloss tab to visualize the collocates. We also noticed a few cases when participants typed a semantic gloss in a search box corresponding to the collocate, for instance, in the *What does it mean?* search option.

Some participants proved to be unfamiliar with the more general concepts of word form and lemma. In the case of *Inverse search*, when introducing a collocate in the search box, users can choose between searching for the exact word form (e.g. the feminine or the masculine form of an adjective) or the lemma, the former being the default search option. A number of queries reveal that the distinction between lemma and word form was not familiar to a few participants. In addition, we also noticed that participants experienced some difficulty in identifying and distinguishing lexical

units. In fact the two questionnaire items for which we obtained the lowest number of correct answers involved identifying a particular lexical unit, and providing its collocates, on the basis of example sentences.

It follows from the above considerations that although the user-friendliness of the DiCE interface can clearly be improved, our results also imply the importance of users' reference skills. We have seen that the participants of our experiment lacked some of the knowledge necessary to successfully use the more advanced functions of the dictionary. This claim is supported by the comparison of the performance of the participant groups of varying user profiles. We have seen that the group of translators performed best, which may be a result of the fact that they may be more used to dealing with different lexical tools. The group of language teachers displayed the poorest results, though it should be noted that they incidentally also belong to an older age group than the rest of the participants, and probably have less experience in using web interfaces in general. In any case, we believe that a demonstration of the DiCE website or the use of familiarization activities prior to the experiment, as in the case of Hamel (2012), would have resulted in a considerably better test performance of most participants. In fact, it should be noted that the only informant who claimed to have completed the web tutorial prior to the experiment itself, performed substantially better on the test than the rest of the participants.

6. Conclusion and future work

This paper has described a usability study of the DiCE web interface. The results of the test above all point to the importance of user familiarization with the concepts used by the dictionary. On the one hand, we believe that a number of changes to the current design can considerably improve dictionary usability. These include a more consistent exemplification of the content to be introduced in each search box, a clear indication of obligatory search boxes and filters, as well as the enhancement of the visibility and distinguishability of navigation aids, e.g. semantic glosses and buttons that allow expansion and contraction information to be shown on the screen. On the other hand, we think that, in order to obtain a clearer picture of the usability of DiCE, future research should better control for reference skills of participants, and include familiarization tasks. Finally, we would like to emphasize that the methodology applied in this experiment implies that the test can be completed on participants' home computers, which considerably facilitates data collection and, therefore, may be of interest for future user experiments.

7. Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the FEDER Funds of the European Commission under the contract number FFI2011-30219-Co2-01, as well as the Spanish Ministry of Education under the FPU grant AP2010-4334. We would also like to thank Miguel Ramos Naveira for his help with retrieving the log files.

8. References

- Alonso Ramos, M. (2004). *Diccionario de colocaciones del español*. Accessed at: <http://www.dicesp.com>.
- Alonso Ramos, M., Nishikawa, A., & Vincze, O. (2010). DiCE in the web: An online Spanish collocation dictionary. In S. Granger & M. Paquot (eds.) *eLexicography in the 21st century: New challenges, new applications. Proceedings of eLex 2009*. Presses Louvain-la-Neuve: Universitaires de Louvain, pp. 367–368.
- Hamel, M.-J. (2012). Testing aspects of the usability of an online learner dictionary prototype: a product- and process-oriented study. *Computer Assisted Language Learning*, 25(4), pp. 339–365.
- Heid, U. (2011). Electronic dictionaries as tools: Towards an assessment of usability. In P. A. Fuentes-Oliveira & H. Bergenholtz (eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. New York: Continuum, pp. 287–304.
- Heid, U., & Zimmermann, J. T. (2012). Usability testing as a tool for e-dictionary design: collocations as a case in point. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of Euralex 2012*. Oslo, pp. 661–671.
- 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.
- Mel'čuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.
- Nielsen, J. (1993). *Usability engineering*. San Francisco: Morgan Kaufman.
- Vincze, O., Mosqueira, E., & Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In I. Boguslavsky & L. Wanner (eds.) *Proceedings of the 5th International Conference on Meaning-Text Theory*. Barcelona, pp. 275–286.

Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes

Carla Parra Escartín,¹ Gyri Smørdal Losnegaard,¹
Gunn Inger Lyse Samdal,¹ Pedro Patiño García²

¹University of Bergen, Norway

²NHH Norwegian School of Economics, Norway

Carla.Parra@uib.no, Gyri.Losnegaard@uib.no, Gunn.Lyse@uib.no, Pedro.Patino@nhh.no

Abstract

In the context of standardisation and interoperability of Language Resources and Tools (LRT), this paper addresses the formal representation of multiword expressions (MWEs) for Natural Language Processing (NLP) purposes. By formal representation we mean the encoding of MWEs in lexical and terminological databases. The representation should render a language resource maximally reusable and ideally allow for seamless integration into any type of NLP application. In the case of MWEs, the situation is particularly complex due to their lexical properties on the one hand, and morphosyntactic variation on the other. Furthermore, their representation in multilingual resources poses even bigger challenges due to extensive translational asymmetry. In this paper we discuss the challenges posed by the formal representation of MWEs. We analyse the needs of four different projects, all NLP oriented, but with slightly different approaches to the collection and representation of MWEs. Based on the analysis, we identify a minimal set of features to be accounted for in any formal representation of MWEs, as well as a set of more specific task-dependent requirements hinging on the intended use of the lexical resource. Finally, we assess to what extent existing standards meet these requirements.

Keywords: Multiword Expressions, Harmonisation, Standardisation, Interoperability, Natural Language Processing Applications, Terminological Resources, Language Resources

1. Introduction

Lexical Language Resources and Tools (LRT), such as machine-readable dictionaries and lexical and terminological databases, constitute a key element of advanced Natural Language Processing (NLP) systems. For the last two decades, researchers in computational lexicography have promoted the importance of designing a set of standards for the creation of reusable and interoperable lexical resources (Moreno Ortiz, 2000; Copestake et al., 2002; Francopoulo et al., 2006b; Francopoulo et al., 2009).

However, the lexis of a language is more than just single words, and in this regard there are still challenges to be overcome. Expressions such as “*fit as a fiddle*”, “*give in*”, “*pose a problem*” and “*as a matter of fact*” are multiword units that need to be

appropriately represented in computational lexicons and yet are difficult to represent in a standardised manner. In their seminal “pain in the neck” article, Sag et al. (2001) point out that multiword expressions (MWEs) constitute a major bottleneck in NLP applications, and recent work and initiatives suggest that this is still the case¹. Moon (1998), Sag et al. (2001) and Baldwin and Kim (2010) note that MWEs exceed word boundaries and have unpredictable properties. Research in the MWE field has also shown that one of the most salient and defining features of MWEs is their semantic non-transparency or non-compositionality. However, there is no widely agreed upon definition or typology of MWEs (Moon, 1998; Cowie, 1998; Sag et al., 2001; Baldwin and Kim, 2010, among others). We adopt a broad definition of MWEs as word combinations that form a unit at some level of linguistic analysis (Ramisch, 2012), and which deviate from regular language lexically, syntactically, semantically, pragmatically and/or statistically (Moon, 1998; Baldwin and Kim, 2010). Thus, although collocations are not always considered MWEs, we also include statistically marked or institutionalised collocations as a type of MWE. The aim of this paper is to capture all kinds of constructions that may pose problems in automatic analysis, and to determine which information should be recorded if such expressions are to be represented in a lexical inventory for NLP purposes. Managing to successfully represent MWEs in lexical and terminological resources is essential to ensure their successful integration in NLP applications, workflows and infrastructures.

The remainder of this paper will focus on this issue from different perspectives, based on four different use case scenarios. Particularly, we will concentrate on defining what information shall be recorded when including MWEs in lexical and terminological resources. How to encode such information will be the subject of further research.

In section 2, four different research projects dealing with MWEs are used as case studies, and their requirements as regards the representation of MWEs are discussed. Section 3 discusses how different standards may be used to formally represent MWEs, and the prerequisites needed to ensure that the final resource is reusable in NLP applications. Section 4 consolidates the results of our analyses and discusses the prerequisites for improved representations of MWEs, and sections 5 and 6 discuss future work to be carried out and sum up the main findings of the study reported here.

2. Case studies: Projects representing MWEs

In the creation of a new lexical or terminological resource the intended usage of such resource may condition its layout and the information recorded in it. In the case of resources including MWEs, what properties to record and represent will depend both

¹ <http://multiword.sourceforge.net>; <http://typo.uni-konstanz.de/parseme>

on the specific purpose and the type(s) of MWE. For certain purposes, a purely lexical account will do: if the end users of a MWE resource are human translators or second language learners, a simple entry with the MWE, its correspondence in the second language, and maybe examples of use, will be sufficient. However, if we intend to reuse the same resource within an NLP application, in order to ensure that the MWE is correctly processed, the computer will probably need additional information for each MWE unit, such as its morphosyntactic properties and its particular behaviour. Different kinds of MWEs may also have different intrinsic features, and the information needed for each particular entry will thus vary with the type in question as well as with the intended final usage of the resource.

In the following subsections four different research projects dealing with MWEs are presented. These projects have been selected because they have been or are currently being carried out by the authors and are presented in chronological order. Two of the projects approach MWEs from a mainly monolingual perspective (subsections 2.1 and 2.4). The other two are multilingual and concern translational correspondences (subsections 2.2 and 2.3).

Each subsection starts with a brief summary of the research project and then proceeds to briefly discuss what information should be recorded for each MWE in the frame of that particular project. Project-specific requirements are then discussed and analysed further in section 4, focusing on the properties that should be mandatory in the representation of MWEs.

2.1 Collocations and statistical analysis of n-grams: Multiword expressions in newspaper text

Lyse and Andersen (2012) describe an empirical study carried out in 2009 which applied various statistical association measures (AMs) to two- and three-word sequences (bigrams and trigrams) from the Norwegian Newspaper Corpus (NNC)². The aim of this study was to determine which AMs are better at picking out relevant MWEs representing different lexical and terminological categories.

The NNC contains ca. 1.3 billion running words and it is the largest searchable corpus of contemporary Norwegian language. With such large amounts of data, efficient tools to identify different kinds of MWEs automatically are of great interest. In fact, recurring MWEs could be thus systematically identified, correctly segmented and added to lexical databases. This could in turn improve the syntactic tagging of the corpus since certain MWEs could be stopped from being further processed by the tagger. Moreover, technical terminology is often realised as MWEs, and the identification of recurrent collocational patterns is relevant for term extraction, even in non-technical texts such as newspaper language.

² <http://avis.uib.no/om-aviskorpuset/english>

Within the context of this study, nine common AMs were applied to bigrams in the NNC and four AMs to trigrams. To analyse the behaviour of each AM in more detail, the 500 top-ranked MWE candidates for each AM were classified manually. A relatively broad definition of MWEs was adopted, taking MWEs to be words that co-occur so often that they are perceived as a linguistic unit. The high-ranked terms were classified according to the following set of categories: *anglicism MWE*, *foreign MWE* (e.g. Latin expressions), *grammatical MWE* (e.g. multiword adverbs), *idiomatic phrase*, *term candidate* and *concept structure appositional phrase* (a term preceded by its superordinate concept).

Table 1 presents some examples of the kinds of MWEs that were highly ranked in the study.

In order to record the identified MWEs, the main requirement would be a standardised way of expressing statistical information about the rank of an item, preferably also including information about the raw frequencies on which the rank was based and the AM used. The extraction of n-grams and their statistical ranking in Lyse and Andersen (2012) did not rely on any linguistic annotation of the data, such as part of speech or lemma information.

The manually categorised MWE units could be interesting for reuse as a gold standard for new statistical experiments, which then imposes further formal representation requirements. To represent foreign MWEs, such as the anglicism “*consumer confidence*” and the Latin expression “*annus horribilis*”, additional attributes for encoding the meaning of the expression itself and the language in which they appear would be needed. Furthermore, foreign expressions raise the need to emphasise that some expressions maintain a foreign inflectional paradigm (e.g. the anglicism “*practical joke*” (*sg.*), “*practical jokes*” (*pl.*)) whereas others adopt the Norwegian one (“*walkie-talkie*” (*sg.*), “*walkie-talkier*” (*pl.*)) and some are only used as frozen expressions without a productive inflectional paradigm (“*freezing fog*”). For term candidates, such as “*alternative energikilder*” (alternative energy sources), morphological information about inflection and internal structure is also necessary.

Multiword unit	English translation	Suggested classification
consumer confidence	-	anglicism MWE
annus horribilis	(Lat.) horrible year	foreign MWE
etter hvert	gradually	grammatical MWE
grøss og gru	shiver and horror	idiomatic phrase
alternative energikilder	alternative energy sources	term candidate

Table 1: Examples of high-ranked collocations in our study

2.2 English and Spanish specialised collocations found in Free Trade Agreements

This project is aimed at approaching the study of the type of collocations that appear in specialised texts from the subject field of international trade, i.e. legal and economics texts. The project also concerns the formal representation of these lexical units, in such a way that the data is machine readable and thus, interchangeable across different language resources (Litkowski, 2006). The data were obtained from the FTA parallel corpus (Patiño García, 2013), with English and Spanish data drawn from 16 official Free Trade Agreements (FTA) including texts from the American and European varieties of the two languages.

Within the frame of this project, a specialised collocation is defined as a type of MWE composed of at least one term that serves as the node. The collocates of this term can be nouns, verbs, adjectives or adverbs in a direct syntactic relation with the node and which do not necessarily appear adjacent to it.

Collocations constitute a challenge for several reasons. First, they can be unpredictable lexical combinations, appearing either adjacent to each other or in a span of several words to the left or right of the node word. Second, in a specialised context, terminology alone is not enough since it is also necessary to master the collocations that are used with these terms. Third, non-experts may encounter problems producing the correct verb, noun or adjective that is typically combined with a specific term (Bartsch, 2004; L’Homme, 2009). However, the lexical combinations of terms do not receive enough attention in lexicography and terminography and are therefore underrepresented in language resources (Pavel, 1993).

English	Spanish
accord favorable treatment	otorgar trato favorable
labor or environmental law enforcement	cumplimiento de la legislación laboral o ambiental
prescribe a conformity assessment procedure	exigir un procedimiento de evaluación de conformidad
prepare adopt apply a technical specification	preparar adoptar aplicar una especificación técnica

Table 2: Specialised collocations in English and their Spanish equivalents [Source: FTA Corpus]

Table 2 presents some English and Spanish examples of specialised collocations that appear in the FTA parallel corpus. In order to produce a language resource which is reusable and interoperable, particular features of every specialised collocation should be properly represented. First of all, the node of the collocation shall be properly

detected and annotated as a term used in a specific subject field. Secondly, all collocates that this term may take should be appropriately tagged as well together with the subject field in which this collocation occurs. In addition to this, information on syntactic and morphological, as well as dialectal, aspects should be included to account for the multiple realisations of these collocations in different varieties of the same language.

2.3 Spanish MWEs as the translational correspondence of German compounds

This project deals with nominal compound words in German and their phraseological correspondences in Spanish. The project aims at improving 1:n word alignment within Germanic and Romance languages and the automatic extraction of compound dictionaries. Such dictionaries need to be appropriately encoded to ensure their reusability, and thus the question of how to represent the correspondence between one word in a language and an MWE in another arises.

Spanish translational correspondences of German compounds usually have the form of regular noun phrases. However, they need to be appropriately represented to yield satisfactory results in NLP applications such as Machine Translation (MT) systems and Terminology Extractors. As an illustration of the kind of units studied in this project, Table 3 shows some of the German compounds found in the TRIS corpus³ (Parra Escartín, 2012) and their translations into Spanish.

German compound	Compound constituents	Spanish correspondence
Wohnungsförderungsverordnung	Wohnung·s·förderung ·s·verordnung	Ley de promoción de viviendas
Warmwasserbereitung	Warm·wasser· bereitung	preparación de agua caliente
Wärmepumpeanlagenförderung	Wärme·pumpe· anlagen·förderung	promoción de instalaciones de bombas de calor

Table 3: German compounds and their correspondences into Spanish
 [English: *Housing Promotion Act / Water heating / Promotion of heat pumping systems*]
 [Source: TRIS Corpus]

As can be observed in Table 3, German compounds constitute a single unit and thus their formal representation does not seem particularly problematic. However, their Spanish translational equivalents may indeed pose a challenge for bilingual and/or multilingual projects, as their representation will need to be more detailed and complex.

³ The TRIS corpus has been compiled for the purposes of the project described here.

As far as German compounds are concerned, it would be desirable to have an indication as to which is the “head” of the compound as it selects inflection and gender. This is usually the most-right element of the compound. Moreover, additional morphological information as regards the rest of the elements forming part of the compound and their internal structure would also be desirable as this conditions the translation of a compound. For instance, the fact that the word “*Anlage*” appears in plural in the middle of the third compound (“*Wärmepumpeanlag****en****förderung*”) requires the Spanish translation to be plural as well (“*instalaciones*”) and translating it in singular would imply a semantic change.

It is also necessary to indicate which elements may be inflected in general language but are fixed or semi-fixed when part of the nominal phrase which translates into German as a compound. And finally, it would also be important to indicate whether other modifiers could be accepted (e.g. an adjective preceding the nominal compound in German) and their position within the nominal phrase in Spanish.

2.4 An NLP study of Norwegian MWEs

The last project is still at an initial stage. It aims to build the first extensive inventory of MWEs for Norwegian, which will serve as a basis for a typology of Norwegian MWEs and for the integration of different types of MWE into NorGram, a computational LFG grammar for Norwegian⁴. The representation requirements presented here are preliminary results based on a pilot analysis of MWE candidates identified during the annotation of the Norwegian treebank INESS⁵. The MWEs in Table 4 are taken from the first chapter of the novel *Sofies verden* (*Sophie's world*) by Jostein Gaarder. They exemplify, although not exhaustively, different kinds of MWEs found in this text.

Norwegian MWE	Literal translation	Idiomatic translation
snakke om	talk of, about	talk about
stå igjen	stand again	be left, remain
gjøre lekser	do homework	do (one's) homework
skille lag	divide team	split, part (ways)
komme rekende på en fjøl	come drifting on a board	come from nowhere (with origin unknown)
sikker på	sure on	sure that, sure of/about
et eller annet	one or other	something

Table 4: MWEs in *Sofies verden*

⁴ http://iness.uib.no/redmine/projects/inesspublic/wiki/NorGram_documentation

⁵ <http://iness.uib.no>

The verbal MWEs in Table 4 exemplify verb-preposition constructions (“*snakke om*”), verb-particle constructions (“*stå igjen*”), verb-object constructions (“*gjøre lekser*” and “*skille lag*”), and idioms (“*komme rekende på en fjøl*”). Each of these types of MWEs has different inherent features that need to be accounted for correspondingly. Verb-preposition and verb-particle constructions tend to be syntactically quite flexible, as opposed to idioms, for instance. On the other hand, we may have different degrees of semantic compositionality even within the same category. In the case of verb-object combinations, there may be expressions whose meaning is fairly transparent, such as the light-verb (or support verb) construction “*gjøre lekser*”, while in other cases the meaning is contributed by all the component words and is less transparent, such as “*skille lag*” (lit. “divide team”). Last but not least, it is also important to highlight that idioms also pose challenges as regards their formal representation because they are syntactically restricted. In the more idiomatic of the two verb-object examples, “*skille lag*”, the object “*lag*” cannot take a determiner and must be in singular and indefinite form. The idiom “*komme rekende på en fjøl*” cannot be passivised without losing its figurative meaning, the verb “*reke*” (“drift”) must be in present participle form, and the object noun “*fjøl*” (“board”) must be in singular indefinite form. It is semantically non-transparent, and like most idiomatic expressions, its lexical components and their morphological form are fairly invariable (Moon, 1998).

If we now focus on the non-verbal MWEs in Table 4, differences arise again with respect to syntactic flexibility and semantic transparency. “*Sikker på*” is an adjective-preposition construction which fills the same syntactic function in the sentence as a simple adjective. Like prepositional verbs, adjectives with selected prepositions require a clausal or nominal argument, and they are transparent in meaning. “*Et eller annet*” (literally “one or other”) functions as a pronoun at clause level. Its meaning is semi-transparent, and it is syntactically fixed in the sense that the word order is invariable and no other words may intervene. However, the disjuncts “*et*” and “*annet*” inflect, and must agree in gender with its anaphoric referent.

The MWE candidates compiled in this project will be stored as entries in a database. For the most general level of use, each entry will contain lexical information as typically found in dictionaries, such as lexical category (part of speech), definition, canonical form (dictionary entry form), surface form (the instance as it occurs in the source text) and, if relevant, context (the sentence from which they were extracted). For research documentation and organisational purposes, it will be necessary to supply each MWE instance with a unique identifier and an identifier for the MWE “lemma”. Information about the source (type, genre, publication date, author etc.), the method used to extract the MWE, the MWE frequency, and pointers to other occurrences of a given expression will also be recorded.

Further, to ensure an adequate level of description for an empirically based, formal classification of MWEs, it will be relevant to know on which linguistic level(s) the MWE exhibits anomalous behaviour, as well as its degree of semantic transparency and syntactic flexibility. As MWEs have varying degrees of semantic transparency and syntactic flexibility, they should be described with reference to a semantic scale ranging from totally transparent in meaning to completely opaque, and a syntactic scale ranging from syntactically flexible to completely restricted (or fixed). Finally, it will also be necessary to represent the internal structure and the morphosyntactic restrictions of each MWE, such as the argument structure of idioms. Whether the relevant properties for each MWE will be identified through manual analysis or by using automated methods is an open methodological question at this stage of the project. However, bearing in mind that the database will be integrated in a computational grammar, this information will have to be included in such a way that the resource can be easily integrated in the grammar and yet contain all relevant information for stand-alone usage.

3. Existing standards for representing MWEs

As we have shown in section 2, the formal representation of MWEs poses several challenges for resource developers, in particular if we aim at the interoperability and reusability of the lexical resource. From a monolingual perspective, a standard for formal representation will have to adequately account for the semantic and morphosyntactic properties of the overall expression and of the component words, internal structure and dependencies, syntactic variation, and potentially also regional language varieties for the given language. For instance, in Spanish, the English idiom “*it’s raining cats and dogs*” may be “*está lloviendo a cántaros*” (lit. “it’s raining pitchers”), “*caen chuzos de punta*” (lit. pointed “pikes are falling”), or “*llueven hasta maridos*” (lit. “it’s raining husbands”), among others, depending on the regional variety of the speaker. For multilingual resources, translational correspondences must be accounted for, and the properties above must also be described for each language and/or language variety. If resource developers aim to create a scalable resource which can also be used by NLP applications, the formal representation of such a resource must also be compliant with the input format accepted by the tools that will process the resource.

Several projects have been undertaken in the last decades with the aim of unifying the coding of computational lexicons and terminologies through the creation of norms. The proposed standards are implemented by organisations, research groups, companies and professionals in the field and foster the exchange of information without losses or obstacles in transmission. Among these projects we can mention

GENELEX⁶, MULTEXT⁷, EAGLES⁸, SIMPLE⁹ and ISLE¹⁰. However, no standard has been broadly accepted thus far.

A quick look at the deliverables written in projects promoting the standardisation, interoperability and reusability of language resources (Rirdance and Vasiljevs, 2006; Hinrichs and Vogel, 2010; Calzolari et al., 2011; Monachini et al., 2011; Borin and Lindh, 2011) reveals that in the case of lexical and terminological resources, there are two standards that are commonly being used and fostered: TBX and LMF. Here, we also look at the TEI initiative, a well-known standard for general text encoding. Table 5 summarises the main features of the three standards.

Standard	Monolingual	Bilingual	Encoding of morphosyntactic features	
			MWE level	Token level
TBX	No	Yes	Yes	No
LMF	Yes	Yes	Yes	Yes
TEI	Yes	Yes	Yes	Yes

Table 5: Summary of standards and encoding

3.1 The TermBase eXchange format

If we first consider the TermBase eXchange format (TBX)¹¹, one of its main advantages is also one of its main drawbacks: its DTD is extremely flexible. This flexibility makes it possible for the user to customise the database and use attribute names suiting the project in which the termbase is created, but comes at the cost of interoperability since the resource will be incompatible with the representation requirements of NLP tools and applications. Furthermore, in TBX MWEs can only be registered as strings. Since they cannot be tagged in a fine-grained manner at token level, TBX prevents the possibility of processing non-fixed MWEs successfully with automatic methods. For instance, it would be impossible to account for the fact that in English the idiom “*it’s raining cats and dogs*” may take internal modification as in “*it’s **certainly** raining cats and dogs today*”. Furthermore, although it would be possible to represent the MWE in all tenses (e.g. “*it **is/was/will be/has been** raining cats and dogs*”) as separate entries, this is clearly not a very efficient way of dealing with its completely regular inflection. The TBX standard was created within the localisation industry and with translators and terminologists as its main target

⁶ <http://llc.oxfordjournals.org/cgi/content/abstract/9/1/47>

⁷ <http://acl.ldc.upenn.edu/C/C94/C94-1097.pdf>

⁸ <http://www.ilc.cnr.it/EAGLES/browse.html>

⁹ <http://www.ub.edu/gilcub/SIMPLE/simple.html>

¹⁰ <http://portal.acm.org/citation.cfm?doid=1118062>

¹¹ http://www.gala-global.org/oscarStandards/tbx/tbx_oscar.pdf

users and it was primarily envisaged for the creation of bilingual and/or multilingual resources, not monolingual ones. Although it is not adequate for monolingual description, other important features such as the regional language variety and multilingual translational correspondences are easily encoded.

In short, in order for TBX to be appropriate for the encoding of MWEs, the names of attributes and values would need to be restricted and agreed upon. Granularity up to token level should be integrated as well as the possibility of assigning inflectional paradigms and other features to allow for language processing and generation in NLP applications. Finally, it should also allow for the proper representation of monolingual lexicons without requiring at least a second language. Until these requirements are met, TBX does not serve as an appropriate standard for encoding MWEs.

3.2 The Lexical Markup Framework

The Lexical Markup Framework is another of the standards encouraged by major standardisation initiatives. It was developed by the Technical Committee 37 of the International Organisation for Standardisation, Subcommittee 4 (ISO TC37/SC4¹²) and, as stated on their website¹³, LMF was developed combining the best designs and methods from many NLP lexicons. However, it was developed for NLP use and not for human users, which is unfortunate since lexical resources are extremely useful in related fields such as second language acquisition. Among its features, there is an extension for bilingual or multilingual dictionaries, designed to express equivalence relations applicable in automatic translation (ISO, 2008). It also includes a module for the representation of MWEs, known as NLP Multiword Expression Pattern, which allows the representation of the internal structure of fixed, semi-fixed and flexible lexical units in a computational lexicon (Francopoulo et al., 2006a; Francopoulo et al., 2006b; Francopoulo et al., 2009).

More recently, UBY-LMF has been published. UBY is a large-scale lexical-semantic resource based on LMF and has been developed with the aim of interoperability and the smooth integration of resources (Gurevych et al., 2012). Despite capturing lexical information at a fine-grained level, using ISOcat data categories and being directly extensible by new languages and resources, this LMF-compliant model currently fails to offer an appropriate representation of MWEs. In fact, MWEs seem to have been overlooked by the developers of this model since they have rather focused on the standardisation of the semantic encoding of the entries of lexical semantic resources.

However, a priori, LMF seems a promising candidate for the encoding of MWEs.

¹² http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee_participation.htm?commid=297592

¹³ <http://www.lexicalmarkupframework.org/>

Spohr (2012, p. 25 ff.) explores this possibility and acknowledges that although it is feasible to represent MWEs in LMF, this has several drawbacks which he further discusses after demonstrating the representation of “*throw to the lion*”. In future work we will try encoding samples from our case studies in this format to test to which degree it actually meets all the encoding requirements we have detected for the different projects accounted for in section 2. The findings of Spohr (2012), however, seem to suggest that although it may be possible to represent MWEs successfully, such a representation might not be the optimal one.

3.3 The Text Encoding Initiative

The Text Encoding Initiative (TEI) also has a specific module for encoding dictionaries. The TEI guidelines (Sperberg-McQueen and Burnard, 2009) explain how to appropriately encode all relevant information for each entry. Concretely, page 262 offers an example in which a compound is encoded as part of a larger lexical entry. TEI dictionaries allow for the encoding of multiple properties of relevance for NLP applications, such as part of speech, geographical area and etymological information, and also include the possibility of adding links and cross-references to other entries in the same resource. This makes TEI particularly interesting for the encoding of lexical and terminological resources, even though it seems to have been disregarded by major standardisation and infrastructure initiatives. The main drawback of TEI – besides the fact that it is not encouraged by the major standardisation initiatives – is that it is very flexible, which again introduces the possibility that different resource developers use different approaches for the encoding of their resources.

4. Prerequisites for improved representations of MWEs

In the following we merge the requirements we have identified in the four projects described in section 2, offering an overview of properties that we believe should be mandatory in the formal representation of MWEs, regardless of the standard used. The differences in the nature of our research projects make us think that we have covered most of the main possible usages a lexical resource could have in NLP applications. As has also been discussed in section 3, existing standards do not currently seem to be fully appropriate for the encoding of MWEs. Although further analysis is required, it seems reasonable to conclude that a set of required features for the representation of MWEs needs to be agreed upon and that standards should comply with successfully encoding all those features. Spohr (2008) divides his requirements for the model of a multifunctional electronic dictionary into the categories *detail of description*, *access and retrieval*, *consistency and integrity*, *specific users' needs* and *specific needs of NLP applications*. He observes that “[o]ne of the most striking requirements, which can be directly derived from the above analysis, is the fact that the underlying formalism cannot be entirely unconstrained, but rather has to be strongly typed”. This leads Spohr to propose the OWL

formalism¹⁴ for representation, a formalism based on the Resource Description Framework (RDF). Although we have not gotten as far as Spohr and we do not attempt here to define which formalism is best for representing MWEs, we have devised a modular representation schema which we believe would meet the requirements we identified. This schema, which has been designed after the representation model envisaged by META-SHARE, consists of three levels of detailed representation, one mandatory and two optional but recommended. We further suggest a need for optional type and purpose dependent representation schemas, or *encoding modules*. In the general, main schema (or module) described below, levels 1 and 2 both describe properties relevant for the description of the overall expression (type level). The second level is an extension of the first and targets more advanced users and usages, while the third level provides information about the MWE at token level. Ideally, levels 1 and 3 should be mandatory, but it is not feasible that every resource creator will be able to encode a potentially large number of expressions in such detail. We therefore propose level 1 as the minimum representation schema for every MWE, and thus the only mandatory level.

1. Type level (mandatory)
 - a. Part of Speech (PoS)
 - b. PoS standard
 - c. Meaning
 - d. The number of component words
2. Type level, extended description (optional)
 - a. Canonical (base) form
 - b. Level(s) of idiosyncrasy
 - c. Translational correspondences
 - d. Language variety
3. Token level (optional)
 - a. PoS
 - b. Lemma
 - c. Grammatical features

4.1 Level 1: Type level

Many MWEs correspond syntactically to simple words or constituents in a sentence, such as the complex adverb “*etter hver*” (lit. after each, “gradually”) and the noun phrase “*preparación de agua caliente*” (lit. preparation of water hot, “water heating”). For such MWEs, the lexical category (part of speech, PoS) should be assigned (1a in the proposed schema). Not all MWEs correspond to one word or constituent, as is the

¹⁴ Web Ontology Language, <http://www.w3.org/TR/owl-ref/>

case with most verbal expressions. The specialised collocation “*accord favorable treatment*” in Table 2 and the verb-object construction “*skille lag*” (lit. divide team, “part”) in Table 4 both exceed constituent level. It should thus be possible to express that the PoS category is “non-applicable”. In such cases, the additional classification module could be used to assign the MWE a type label instead, such as *sentence* (like “*it’s raining cats and dogs*”), *verb-particle construction* (VPC), *light verb construction* (LVC), etc.

Which PoS standard is used should also be accounted for (1b). Even though there is no specific standard that is commonly used in NLP, the PoS inventory (for European languages) normally includes the traditional categories noun, verb, adjective, adverb, pronoun, conjunction, preposition and interjection. Most linguists would probably not settle for such a crude classification, and for encoding purposes we recommend that the representation schema is equipped with the most widely used PoS standards. In case these are not applicable, the representation schema should also allow users to define their own custom-made inventories of lexical categories that are suitable for their individual projects or needs.

Meaning can be represented with a synonym, a definition, a translation or a transliteration. All of these possibilities should be available in the encoding schema (1c). 1d accounts for the number of constituent words.

All features at this level are mandatory, and features that are not relevant for a given MWE should be marked as non-applicable.

4.2 Level 2: Type level, extended description

Level 2 of our proposed schema targets more advanced usages and is recommended, but optional. After all, a particular resource may not be bilingual or account for dialectal varieties; or the MWEs may not have been analysed and thus may not be classified or described in terms of idiosyncrasy (at which linguistic levels they deviate from “regular” language; syntactic, semantic etc.). However, having a pre-defined module that envisages the addition of such information would ease the scalability and reusability of the resource in the long run. As for the canonical form, it would be desirable to have a standardised way of representing this, e.g. the base form of each component word.

4.3 Level 3: Token level

The final level describes the properties of the component words and again is recommended, but optional. This level allows for the annotation of component words with grammatical information.

4.4 Additional encoding modules

The provision of additional modules to the main schema will allow for optional

representation of different types of MWEs, of information particular to a given field, topic or discipline, and of purpose-dependent properties. A modular representation schema thus makes it possible to describe MWEs from different perspectives according to the needs of the individual user or resource developer. Furthermore, optional modules for specialised information may simply be ignored by processing tools which do not make use of that particular type of information. Additional modules depending on the particular research project and the final usage of the resource could be:

- Classification
- Morphosyntactic profile
- Metadata
- Organisational data
- Semantic profile
- Terminology
- Multilinguality
- Named Entity

Due to the lack of agreement with respect to the definition and classification of MWEs, information about the type of MWE could be represented in a dedicated *classification* encoding module. This module should offer predefined MWE categories from existing typologies. It should also allow for customisation of classification schemas, so that users may classify the MWEs according to his/her own schema, and if desired, according to several schemas. Categories that reflect syntactic structure, such as *light verb construction* and *particle verb*, could be represented here, as well as more general types such as *collocation*, *idiom* or *metaphor*.

The description of the more complex morphological syntactic properties of an MWE would be difficult to account for at token level, since such properties often involve dependencies between words. We thus propose to have a dedicated *morphosyntactic* module. This would be the most important component for ensuring interoperability with and integration in NLP applications. The module should account for aspects that cannot easily be represented at word level, such as the internal structure of the MWE, morphosyntactic restrictions (e.g. the indication of morphosyntactically “frozen” words), subcategorisation information, description of internal modifiers, their type and position within the expression, etc. Dependency descriptions involve marking phrasal heads, node words and collocates, indicating which words take modifiers, etc. The module should also indicate the degree of syntactic flexibility, from fixed to completely flexible.

A *metadata* module would meet the requirements identified in 2.4, allowing for a description of the source material. This could be information about the source type

(corpus, dictionary, website, etc.) and specific texts (title, author, date, etc.), and is particularly relevant for projects where MWEs have been extracted from multiple sources. The requirements pointed out in 2.1 further raise a need for *organisational data* such as the extraction method used, frequency and rank (based on the number of occurrences of the MWE in the source material), and pointers to other occurrences or entries.

The *semantic* module would be relevant for language analysis. This module should allow for an elaboration of the definition and meaning, the degree of semantic transparency, to which degree the different constituents contribute meaning to the overall expression, etc. Features relevant to terminology and multilingual resources are described in sections 2.1, 2.2 and 2.3 and include the representation of collocational features, ontological relations, etc.

5. Discussion and future work

The implementation of a flexible but standardised and agreed-upon encoding schema such as the one discussed here would ensure the scalability of lexical and terminological resources, since researchers could then take as a starting point an already developed resource and add the modules they need for their particular projects. For instance, the terminology resource described in 2.1 could be taken as a starting point for the creation of the resource under development in the project described in 2.4. Resources developed independently in different projects could easily be merged into one resource with several modules, where different modules encode the specific information for each project. Finally, in order to ensure the scalability and interoperability of the resources created, feature names, values and formats should be standardised to the extent possible and correspondences between different standards should be provided to ensure the successful merging of resources if necessary.

As a follow-up of the analysis reported here, we intend to assess the appropriateness of the different standards available for the encoding of lexicons and terminological databases, using data from our respective research projects. We may then determine to what extent these standards actually allow for encoding of the features that we have proposed as the minimal set of features to be included in the representation of any type of MWE and any type of NLP application. If we aim to develop resources which are standardised and interoperable, encoding MWEs in one of the existing standards would not be enough as it would be possible to have four different resources encoded using the same standard but providing different information or information with mismatched attribute names. In order to ensure the reusability of our resources, a compromise among all stakeholders is necessary by agreeing upon a standard set of attributes and values. This would make the mapping between different encoding formats feasible and as a result, merging, exchanging and enlarging resources would no longer be so problematic.

6. Conclusion

In this paper we have discussed the requirements for the formal representation of MWEs from different perspectives. Four projects have been presented, and their needs have been discussed to show the wide variety of projects and usage scenarios where an appropriate formal representation of MWEs may be relevant.

Despite several recent standardisation efforts and initiatives, none of the major encoding standards meet all of the requirements identified in section 2. In order to encode MWEs in lexical resources in a way that both accommodates our individual requirements and renders the resources comparable, extendable and applicable outside our own limited projects, we have thus proposed a modularised representation schema with different modules or profiles for different purposes and uses.

Importantly, this is not an attempt to define a new encoding standard. Rather, and as pointed out in section 4, we think that it is necessary to have more information considered as *mandatory* in the representation of MWEs, in particular with respect to the multilingual aspect and their unique features.

The projects described in section 2 present real usage scenarios, all of which require detailed formal representations of MWEs. From our point of view, efforts should be put into enhancing existing standards by devising DTDs with standardised sets of attributes and values for general descriptions and standardised ways for representing complex morphosyntactic information. The study reported here has highlighted the need for flexibility in the encoding of linguistic phenomena. Our recommendation is to implement specific modules for gathering and representing the specific information particular to a given topic, type or use for every MWE included within lexical or terminological resources. This will ensure their reusability and interoperability and will thus bring us closer to a proper treatment in NLP applications.

7. Acknowledgements

The research reported on in this paper has received funding from the EU under FP7, Marie Curie Actions, SP3 People ITN, grant agreement 238405, (project CLARA), and from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, grant agreement no. 270899, the University of Bergen and the Norwegian School of Economics.

The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

8. References

- Atkins, S., Bel, N., Bouillon, P., Charoenporn, T., Gibbon, D., Grishman, R., Huan, C.-R., Kawtrakul, A., Ide, N., Lee, H.-Y., Li, P. J. K., McNaught, J., Odijk, J., Palmer, M., Quochi, V., Reeves, R., Sharma, D. M., Sornlertlamvanich, V., Tokunaga, T., Thurmair, G., Villegas, M., Zampolli, A., and Zeiton, E. (2001). Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry) Deliverable D2.2-D3.2. ISLE project: ISLE Computational Lexicon Working Group.
- Baldwin, T. and Km, S. N. (2010). Multiword Expressions. In Indurkha, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.
- Bartsch, S. (2004). *Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, Tübingen.
- Borin, L. and Lindh, J. (2011). Deliverable D4.1: Metadata descriptions and other interoperability standards. Version 1.0, 2011-05-02. Deliverable in the META-NORD project (CIP 270899).
- Calzolari, N., Bel, N., Choukri, K., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., and Soria, C. (2011). Final FLaReNet Deliverable: Language Resources for the Future - The Future of Language Resources. The Strategic Language Resource Agenda. FLaReNet project.
- Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., and Flickinger, D. (2002). Multi- word Expressions: Linguistic Precision and Reusability. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands*, pp. 1941–1947.
- Cowie, A. P. (1998). *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Clarendon Press.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006a). Lexical Markup Framework (LMF) for NLP Multilingual Resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pp. 1–8, Sydney, Australia. Association for Computational Linguistics.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2009). Multilingual resources for NLP in the Lexical Markup Framework (LMF). *Language Resources and Evaluation*, 43:57–70. 10.1007/s10579-008-9077-5.
- Francopoulo, G., Declerck, T., Monachini, M., and Romary, L. (2006b). The relevance of standards for research infrastructures. In *International Conference on*

- Language Resources and Evaluation - LREC 2006*, Gênes/Italie. European Language Resources Association (ELRA).
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pp. 580–590, Avignon, France.
- Hinrichs, E. and Vogel, I. (2010). Deliverable D5C-3: Interoperability and Standards. CLARIN Project.
- ISO (2008). Language resource management - Lexical Markup Framework (LMF), ISO 24613:2008, ISO/TC 37/SC 4 N453 (N330 Rev.16).
- L’Homme, M. C. (2009). A methodology for describing collocations in a specialised dictionary. In *Lexicography in the 21st century*, pp. 237–256. John Benjamins, Amsterdam/Philadelphia.
- Litkowski, K. (2005). Computational Lexicons and Dictionaries. In Brown, K., editor, *Encyclopedia of Language and Linguistics (2nd ed.)*, pp. 753–759. Elsevier, London.
- Lyse, G. I. and Andersen, G. (2012). Collocations and statistical analysis of n-grams - Multiword expressions in newspaper text. *Exploring Newspaper Language. Using the web to create and investigate a large corpus of modern Norwegian*.
- Monachini, M., Quochi, V., Calzolari, N., Bel, N., Budin, G., Caselli, T., Choukri, K., Francopoulo, G., Hinrichs, E., Krauwer, S., Lemnitzer, L., Mariani, J., Odijk, J., Piperidis, S., Przepiorkowski, A., Romary, L., Schmidt, H., Uszkoreit, H., and Wittenburg, P. (2011). The Standards’ Landscape Towards an Interoperability Framework. FLReNet project.
- Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.
- Moreno Ortiz, A. (2000). Diseño e implementación de un lexicón computacional para lexicografía y traducción automática. *Estudios de Lingüística del Español*, 9.
- Parra Escartín, C. (2012). Design and compilation of a specialized Spanish-German parallel corpus. In Calzolari, N. C. C., Choukri, K., Declerck, T., Uğur Doğan, M., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).
- Patiño García, P. (2013). *FTA Corpus: a parallel corpus of English and Spanish Free Trade Agreements for the study of specialized collocations*, volume 3 of *Bergen Language and Linguistic Studies*, pp. 81–92. University of Bergen Library, Bergen, Norway.
- Pavel, S. (1993). Neology and phraseology as terminology-in-the-making. In

- Terminology: applications in interdisciplinary communication*, pp. 21–34. John Benjamins, Amsterdam.
- Ramisch, C. (2012). *A generic and open framework for multiword expressions treatment: from acquisition to applications*. PhD thesis, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France.
- Rirdance, S. and Vasiljevs, A. e. (2006). Towards Consolidation of European Terminology Resources. experience and Recommendations from EuroTermBank Project. Technical report, EuroTermBank Consortium.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2001). Multiword Expressions: A Pain in the Neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15.
- Sperberg McQueen, M. and Burnard, L. (2009). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical report, The TEI Consortium.
- Spohr, D. (2008). Requirements for the Design of Electronic Dictionaries and a Proposal for their Formalism. In *Proceedings of the EURALEX International Congress 2008*.
- Spohr, D. (2012). *Towards a multifunctional lexical resource design and implementations of a graph-based lexicon model*, volume 141 of *Lexicographica. Series maior*. De Gruyter.

Use of support verbs in FrameNet annotations

Kaarlo Voionmaa, Karin Friberg Heppin

Språkbanken, CLT, University of Gothenburg Address
E-mail: kaarlo.voionmaa@svenska.gu.se, karin.friberg.heppin@svenska.gu.se,

Abstract

This article discusses the frame semantic annotations done in the Swedish FrameNet (SweFN) at the Centre for Language Technology (CLT) at the University of Gothenburg. The annotations are made manually, and result in full-coded frames. These are conceptual structures representing the description of types of situations, objects or events. We focus on annotations where verbs combine with nouns to produce predicates, e.g. *göra* 'make' in *göra uppehåll* 'make a pause.' These verbs are called *support verbs*, and the corresponding constructions *support verb constructions* (SVC). Not all verb-noun-combinations are SVCs, and adequate defining features are required to identify eligible SVCs. The focus of this paper is to scrutinize the criteria through which this aim can be achieved. Working at the CLT, we have access to a variety of computational research tools and a large Swedish text corpus. These resources buttress the annotation by showing, among other things, frequential properties of verb-noun combinations. We also discuss lexico-semantic features of the Swedish language as revealed through annotations.

Keywords: support verb constructions; frame semantics; annotation; Swedish

1. Introduction

Multiword expressions are a central and well-debated topic in linguistics and computational linguistics. Among the many kinds of multiword expressions there are constructions, where the finite verbs of sentences are semantically reduced and syntactically supportive of their nominal and, occasionally, adverbial complements. It concerns collocations such as *(He) gave a lecture*, where *lecture* is the base of the collocation and *gave* is the collocating verb. *Gave* has a non-free sense in this construction and does not have the sense of transferring possession that it has in constructions like *(He) gave ice cream to the children*. We call collocations like *give (a) lecture*, *support verb constructions* (SVCs), and the verb a *support verb*.¹

¹ For a discussion on the relevant terminology, see Langer (2004b). Constructions such as or similar to the one examined in this study have been termed as *complex predicates*, *operator verbs*, *light verbs* and others.

2. Aim of the paper

The paper deals with annotations of SVCs, which are performed manually with the help of various computational tools and resources.² The SVC annotations are discussed both from a theoretical and practical point of view drawing on the relevant criteria presented by Ruppenhofer et al. (2010), and from the perspective of the tests that Stefan Langer (2004a) put forward. The focus is on the criteria used to distinguish SVCs from verb-noun combinations that are not eligible as SVCs, which has proved to be a difficult matter in practical work on annotations. A well-informed understanding of the role SVCs play in language will benefit several different areas such as text generation, information extraction and text understanding.

SVCs can often be paraphrased by monomorphic verbs, and therefore their use also concerns areas such as psycholinguistics and its applications in second language acquisition. SVCs are also interesting from a typological point of view, since they occur in many languages, though there certainly are differences in how they are construed in them. For SVC-constructions in Japanese, see Miyamoto (1999); in Korean and Japanese, Karimi-Doostan (1997); in Farsi, Goldberg (2003); in Czech and Swedish, Cinková (2009); in German, Hanks & al. (2006); and in Urdu, Butt (2003).

In their study on collocations extracted from the FrameNet corpus, Alonso Ramons et al. (2008) state that support verbs are lexically idiosyncratic, and thus hard to predict.³ In this article we discuss ways to deal with these difficulties; and in particular, the criteria for identification of support verbs and SVCs. We also examine a sample of representative SVCs in order to show how the computational research tools can be used to buttress analytical work with verb-noun combinations.

3. The Swedish FrameNet project

The present study is part of the research work currently carried out in the Swedish FrameNet++ project (SweFN++) at the Centre for Language Technology in Gothenburg.⁴ The main goal of SweFN++ is the creation of a fully integrated lexical macro-resource for Swedish for use as a basic infrastructural component in Swedish language technology research and in development of natural language processing

² The corpus search interface Korp has a central role insofar as access to and use of the corpora. It contains 146 corpora, 104 712 701 sentences and more than 1.4 billion tokens (Oct. 2013). Apart from Korp, lexicon search interfaces available to the SweFN project also include *Karp*, which comprises 21 lexica and 673,932 entries. Integrated with Korp and Karp there is SALDO (Swedish Associative Thesaurus), which is an extensive electronic lexicon resource for the modern Swedish written language.

³ Their list of English support verbs is found at <http://www1.ccls.columbia.edu/~nlp/resources/support-verbs.txt>. For computationally oriented research on SVCs, see Salkoff (y.a.), Grefenstette & Teufel (1995), and Laport et al. (2008).

⁴ Funded by Vetenskapsrådet under contract 2010-6013 (Borin et al., 2010).

applications and annotated corpora for Swedish. A second goal is to make all resources and tools developed in the project freely available under open-content/open-source licenses. One specific objective of the SweFN++ project is to create a full-scale Swedish FrameNet fully integrated into the macro-resource.

The Swedish FrameNet (SweFN)⁵ is a full-scale lexical resource with a target size of at least 50,000 lexical units which is designed to support Swedish LT applications such as machine learning, text generation, text understanding and information extraction, in all domains. In September 2013, SweFN covered 905 frames comprising over 26,000 lexical units from the SALDO (Borin et al., 2010).

The project is based on the English Berkeley FrameNet (BFN) under construction by a research team at the International Computer Science Institute in Berkeley. BFN contains over 10,000 lexical units in more than 1,000 frames, together with more than 170,000 sentences. There is a fairly big difference between the number of lexical units in BFN and SweFN, a difference which is mainly due to the specific focus in the latter project on lexical units, while in BFN the focus has been on annotated example sentences.

4. FrameNet annotation procedure

Annotation, in SweFN as in BFN, entails labeling words and phrases of a given example sentence as *frame elements* (FEs), representing different semantic roles. These elements pertain to certain *frames*, frames being script-like structures describing different types of situations, objects or events. The annotation applies frame semantic principles, and in accordance with them, the FEs divide into core FEs and non-core FEs. The core FEs are part of the definition of the frames. The non-core FEs, such as Manner, Place and Time, are elements of more general kind and exist in many frames.

The annotation is partial in the sense that the labels of FEs are applied only to the relevant words or phrases of example sentences. Moreover, annotation concerns whole constituents rather than only the heads of the constituents.

In Table 1, a simplified frame annotation is shown.⁶ It concerns the frame SPEAK_ON_TOPIC, and there is an SVC in each of the example sentences. The support verb is tagged as SUPP. The tags of the frame elements are self-explanatory, whereas the digits after the LUs (e.g. lecture...1) are indexes of the entries or word senses in SALDO. In this manner, the lexical units in SweFN are systematically connected to other resources of SweFN.

⁵ SweFN is available as a free resource (CC-BY-SA 3.0, LGPL 3.0).

⁶ For more information on frame annotations, see <http://spraakbanken.gu.se/eng/research/swefn>

In Table 1, there are two example sentences, whose predicate verbs are, respectively, **ger** 'gives' and **hölls** 'was-held'. Both of them collocate with the noun **föreläsningar** 'lectures', with which they build a SVC. One may notice that in the group of lexical units, these verbs are not included. This is because they are supportive lexical elements and not full lexical units of the frame.

Frame	SPEAK_ON_TOPIC	English translation
Core FEs	Audience (A), Speaker (S), Topic (T)	
Non-core FEs	Explanation (E), Manner (M), Medium (ME), Occasion (O), Place (P), Time (TI)	
Examples	Ja, och så [SUPP ger] [S jag] [LU föreläsningar]. [TI Igår] [SUPP hölls] [LU föreläsningar] [T om livsstil och hälsa] [P i Nordstan].	Yeah, and then [SUPP give] [S I] [LU lectures]. [TI Yesterday] [SUPP were-held] [LU lectures] [T on lifestyle and health] [P in Nordstan]
Lexical units from SALDO	vb: föreläsa..1, predika..1 nn: föreläsande..1 föreläsning..1 predikande..1	vb: lecture..1, preach..1 nn: lecturing..1, lecture..1, preaching..1

Table 1. The frame SPEAK_ON_TOPIC with annotated example sentences.

The annotation of SVCs like the ones shown in table 1, is based on the study by Ruppenhofer et al. (2010), which has in practical terms been the manual of the SweFN project.

5. Support verb constructions

In the Berkeley FrameNet project, it was noticed that the SVCs brought with them “discrepancies between syntactic and semantic structure” (Fillmore et al., 2003). These discrepancies are due to the fact that in SVCs the support verb is the syntactic head, whereas the noun is the semantic head. Fillmore et al. (2003) call the support verbs “semantically neutral.” They characterize these verbs by saying that they “turn an event noun or a state noun into a verb phrase-like predicate [...]” (op. cit.).⁷

In SVCs, the verbs are typically selected by the nouns rather than the other way around. In English, for instance, the noun **prayer** opts for the verb **say**, (**say**

⁷ Apart from SVCs there are two more verb-noun constructions that are of importance for annotation, namely, the copula-noun (or copula-adjective) combination, and the construction having a controller verb such as *merit*, *offer*, *consider* and *find* as its syntactic head. See Ruppenhofer et al. (2010: 32–33 and, 40–41) for more specific information.

prayer), while the corresponding verb for *speech* is *give* (Fillmore et al., 2003: 244). Occasionally, the choice of verbs may concern fairly fine-grained nuances as for instance in Swedish, where there is a distinction between *ha samtal* ‘have conversation, converse’ vs. *hålla samtal* ‘hold conversation(s), arrange discussion(s).’ In this case, *samtal* ‘conversation, discussion’ opts either for *ha* or *hålla*, depending on whether it concerns customary conversing or whether it is about arranging conversation(s).

In their study, Ruppenhofer et al. (2010) provide criteria for identification of the SVCs. Before presenting the criteria, they briefly discuss semantic features of support verbs. They state that these verbs do not introduce significant semantics of their own but that this does not mean that these verbs are void of semantic features altogether. This state-of-affairs is illustrated in the following examples where the support verb is in bold face:⁸

Causative support verb: **orsaka** + förstörelse ‘**cause** + destruction’

Aspectually inchoative support verb: **få** + insikt ‘**get** + insight’

Support verb indicating point-of-view: **ta** + lån ‘**take** loan’

Ruppenhofer et al. (2010) define the support verb constructions using four criteria that are listed and commented below:

- 1. The support verbs govern the nouns syntactically.** This is the case for example, in the sentence **Han gav en föreläsning** ‘He gave a lecture.’, where *föreläsning* is the object complement of the verb *gav*.
- 2. The noun denotes a state, event, or relation by itself.** This criterion excludes a number of other groups of nouns like sentient entities such as human beings and animals.
- 3. The support verb does not have the same meaning in the SVC as it has without the construction.** This criterion specifies that verbs that are used as support verbs are polysemous. The polysemy of a verb can be examined with the help of the so-called Zeugma-test as explained below and in Langer (2004a).
- 4. In an SVC, the support verb has very little meaning of its own.** The meaning of the construction relies almost entirely on the noun. This criterion must be applied with the reservation that support verbs may have semantic

⁸ In her study on light verbs (i.e. support verbs), Brugman (2001) comes to the conclusion that they not merely have function but meaning too. She suggests that light verbs are systematically related to their heavy counterparts in retaining their force-dynamic properties but drawing rather on a psychological domain than a physical domain as do their heavy counterparts.

properties of their own as shown above. Moreover, support verbs may assign semantic roles to given syntactic constituents. In the sentence **Hon genomgick en operation** 'She underwent an operation', the support verb **genomgick** 'underwent' assigns the semantic role Patient to **hon** 'she', the subject of the sentence, while in the sentence **Hon genomförde operationen** 'She performed the operation', the support verb **genomförde** assigns the role Agent to the subject noun **hon**.

In order to find a more unequivocal base to define the SVC, we may turn to the tests that Stefan Langer has presented. In his study (2004a), he puts forward a test battery to define SVCs and support verbs. In this battery, the Zeugma-test distinguishes whether a verb has more than one sense. An appropriate example is the sentence ***Hon gav en föreläsning och glass till barnen** 'She gave a lecture and ice cream to the children'. This sentence is semantically infelicitous, and as such it shows that when the verb **ge** 'give' combines with **föreläsning** 'lecture,' it has not the meaning of transfer of possession that it has in the sentence **Hon gav glass till barnen** 'She gave ice cream to the children.' In combination with the complement **föreläsning** 'lecture', the verb **ge** (and, respectively, **give**) is simply a support verb having "little meaning of its own" as required in criterion four of Ruppenhofer et al. (2010).

Another test that Stefan Langer (2004a) discusses, concerns the SVCs that can be paraphrased with a semantically equivalent monomorphic verb. For instance, the SVC **ge en föreläsning** 'give a lecture' can be paraphrased with a semantically equivalent monomorphic verb **föreläsa** 'to lecture'. By contrast, constructions consisting of non-support verbs combined with noun complements may not be paraphraseable as monomorphic verbs. See section 6.1 for further discussion on this issue.

6. SVCs in the Swedish FrameNet annotations

Below, a sample of SVCs is studied. The focus is on on three of the four SVC criteria. Criterion 1 is omitted, because all instances of SVCs examined in this paper are verb-noun combinations. The SVCs are presented in the form of verb-noun pairs. We take each of the three criteria and examine how they have been applied in the actual framenet codings. In the case of criteria 3 and 4, we shall make use of SALDO alongside Korp, the corpus search interface, and Karp, the lexical infrastructure and search tool (see footnote 2).

6.1.1 Semantic properties of the SVC-noun base

Criterion 2 requires that the noun base of the construction denotes state, event, or relation by itself (see chapter 5). Whether this requirement is realized in the annotated SVCs may be difficult to establish. Ruppenhofer et al. (2010) do not give definitions or clear guidance, either, as to how the notions in question should be interpreted. In what follows, the semantic properties of noun bases of SVCs will be

examined from the point of view of paraphraseability.

Paraphrasing is suggested by Stefan Langer (2004a) as one of the tests of SVCs, because in paraphrases both the noun base and the verb collocate are involved. Paraphrasing also reveals what the noun bases of the constructions are semantically like. In Table 2, a sample of SVCs is presented, first in Swedish in the left column, then translated into English in the middle, and paraphrased with the corresponding monomorphemic Swedish and English verbs in the right column.

Support verb construction	English translation	Monomorphemic verb
driva + jordbruk	practise + farming	bruka (jord) / to farm
begå + våldsbrott	commit + crime of violence	? våldföra / ? to violate
ge + komplimang	give + compliment	komplimentera / to compliment
göra + distinction	make + distinction	urskilja / to distinguish
göra + försök	make + attempt	försöka / to attempt
hysa + aversion	show + aversion	ogilla / to avert
hålla + överläggning	hold + discussion	diskutera / to discuss
lägga + tonvikt	lay + emphasis	betona / to emphasize
ta + hämnd	take + revenge	hämnas / to revenge

Table 2. Sample of SVCs paraphrased as monomorphemic verbs

All verb-noun pairs in the table can be paraphrased in a fairly straightforward manner except for **begå våldsbrott** ‘commit crime’, which perhaps should be interpreted as an idiom rather than a SVC. (For idioms in modern Swedish, see Sköldberg 2004.)

In regard to semantics of the noun bases, a prominent feature appears, namely, the fact that all of them denote some kind of activity or check on activity, i.e. event, state or relation. Consequently, the verb-noun pairs in table 2 meet the second criterion of SVCs as posited by Ruppenhofer et al. (2010).

6.2 SVCs and polysemy of the verbs involved

According to the third SVC criterion, the support verb does not have the same meaning in the SVC as it has without the construction as a full verb. This means that,

in effect, the support verb should be polysemous. To illustrate this criterion, the semantic features of the verb **hålla** ‘hold’ can be studied. It is frequently used as a support verb in Swedish, and it also belongs to the most frequent full verbs of the language. In table 3, the full verb **hålla** has been differentiated into its senses derived from the SALDO (see footnote 2).

Sense-ID	English	Frame
hålla..1	grab	Manipulation
hålla..2	be operational	Being_operational
hålla..3	fulfill	Meet_specifications
hålla..4	do something with X	Intentionally_affect
hålla..5	side, support	Taking_sides
hålla..6	last, persist	Duration_relation
hålla..7	remain, stay	State_continue
hålla..8	keep X V-ing	Cause_to_continue

Table 3. Senses of **hålla** ‘hold’ in the SALDO lexicon.

Table 3 illustrates the polysemy of the verb **hålla** ‘hold.’ It can be noted that none of the senses listed in the table is applicable as a sense of **hålla** when it is used as a collocate verb of SVC. It seems, then, that **hålla** as a support verb and as a full verb are mutually exclusive in semantic terms. Insofar as the tools and resources are concerned, we may note that when analyzing the semantic properties of **hålla**, SALDO as a half-automatic implement is of great help. It buttresses the reliability and validity of the analysis.

6.2.1 SVCs and semantic lightness of the support verbs

The fourth criterion states that the support verb of the SVC should have very little meaning of its own and that the meaning of the construction relies almost entirely on the noun. As a collocate in a given SVC, the verb should not be semantically specific. So, for instance, a semantically specific verb like **heed** may not be used as a support verb whereas a polysemous verb such as **hålla** will do as a support verb (see above section 6.2).

Whether or not the criterion is realized in a verb-noun combination can be ascertained in several ways with the help of available lexical resources. To begin with, one may assume that if a given transitive verb, which is the most common type of verb in SVCs, has object complements that semantically differ greatly from one another, the verb may be semantically not specific but have little meaning of its own. Based on this, it may be eligible as a verb collocate of a SVC.

In order to examine the issue, we may extract so-called *word picture* from Korp. This picture shows the lexical context of the search term as based on frequency in the

large lexical corpora. In the present case, it concerns nouns that occur after **hålla** 'hold.', which is opted as the search term. Table 4 below shows the 14 most frequent object complements of the verb.

Object complement of <i>hålla</i>	English translation	Freq
koll	control	2206
möte	meeting	1514
utkik	outlook	1419
val	choice,election	1267
tumme	thumb	1244
tävling	competition	942
väder	weather	869
förhör	interrogation	610
rättegång	trial	589
trend	trend	580
tal	speech	512
folkomröstning	referendum	512
häktningförhandling	committal proceedings	301
konferens	conference	42

Table 4. Object complements of the verb **hålla**

The first impression of the word picture shown in table 4 is that the semantic spread of the object complements of the verb **hålla** is considerable. The following nouns (here, in English) stand for some kind of event: **meeting, competition, interrogation, trial, speech, referendum, committal proceedings** and **conference**. On the other hand, **control, outlook** and **weather** denote different sorts of state, whereas **trend** denotes a certain kind of relation. **Hålla tummarna**, lit. 'hold the thumb(s)', is a Swedish saying corresponding to the English turn of phrase **cross one's fingers**.

In the word picture, a number of different nouns collocate with **hålla**, the search term, which unequivocally shows that **hålla** is a semantically non-specific verb. As such, it meets the fourth criterion of SVCs and suits well to be used as a support verb..

6.2.2 Head verbs of a given object complement

The word pictures extracted through Korp make it easy to examine various aspects of verb-noun combinations. One may take a noun as the search term, and examine what verbs may have it as the object complement. This can be illustrated with the following example, where the search term is the noun **överläggning** 'discussion, consultation'.

The word picture in table 5 shows that the verb **ha** + **överläggning** ‘have discussion’ is the most frequent verb-noun combination followed by **hålla** + **överläggning** ‘hold discussion.’ Both of these verbs are very polysemous. The verbs **inleda**, **fortsätta**, **ta**, **ta upp**, **begära** and **kräva**, on the other hand, differ from **ha** and **hålla**, since they denote a situation where discussion is being started or requested to start. Therefore they can be described as semantically specific verbs. The verb **föra** ‘conduct’ is close to **ha** and **hålla** as it also denotes continued pursuing of activity. The verb **pågå** ‘be going on’ differs from these verbs, since it takes the activity itself as its subject, typically in sentences with a preposed adverbial, e.g. **I New York pågår överläggningarna**, lit. ‘In New York the discussions are going on.’ Consequently, **ha**, **hålla** and **föra** can be used as support verbs with the noun base **överläggning** in SVCs.

Verb before the noun överläggning	English translation	Freq
ha	have	520
hålla	hold	122
inleda	open	75
fortsätta	continue	56
ta	take	50
ta upp	take up	42
begära	want, request	34
föra	conduct, pursue	32
kräva	demand	31
pågå	be going	29

Table 5. Verbs used before **överläggning** in Swedish text corpora as extracted through Korp

In this section the word picture of the search term **överläggning** has been examined. It has been found out what the verbs are semantically like that appear as its heads in various verb phrases. The word picture has also distinguished a group of verbs that may construe a SVC together with **överläggning**. These verbs are **ha**, **hålla** and **föra**. Both the present word picture and the one discussed in section 6.3 have proved to be useful for the analysis, since they have helped investigate more closely Swedish SVCs and support verbs. We may conclude that research workers' knowledge of language and her/his linguistic intuitions are buttressed and, at times, contested by the evidence shown in these pictures.

7. Conclusions

This article can be summarized in three points:

(1) Support verbs in SVCs are non-specific and polysemous verbs, and they collocate with nouns that typically denote state, event or relation. A list of verbs eligible as collocates in SVCs might be a good idea to compose in combination with a list of eligible noun bases. With the help of such lists the frame semantic annotation of SVCs could be made more consequent and, hopefully, more automatic, and thereby less time consuming. One has to keep in mind, however, that the eligibility of both nouns and verbs for SVCs may be difficult to pin down, in which case tests such as the Zeugma-test and the paraphrasing test may be helpful.

(2) SVCs and monomorphic verbs are often paraphraseable with one another. This enhances expressive resources of the language. Occasionally, fairly fine-grained distinctions emerge between SVCs themselves such as **ha** vs. **hålla samtal** ‘have vs. hold (or arrange) conversation(s)’, in Swedish.

(3) Research workers' linguistic competence, language knowledge and their linguistic intuitions are essential for successful analysis and annotation of SVCs. However, computational tools and lexical resources such as Korp, Karp and SALDO, are very much needed to buttress this work. In regard to verb-noun combinations, the aim is to establish as unequivocally as possible, their status as constructions, that is, for instance, whether they are SVCs or not. This aim should be pursued effectively and consequently, because it contributes to the value of SweFN as a reliable, adequate and rich lexical resource for linguistic research.

8. Acknowledgements

The research presented here was supported by the Swedish Research Council (the project Swedish Framenet++, VR dnr. 2010-6013) and by the University of Gothenburg, through its support of the Centre for Language Technology and Språkbanken (the Swedish Language Bank). We would like to thank our colleagues at the Department of Swedish, University of Gothenburg Lars Borin, Markus Forsberg, Håkan Jansson and Maria Toporowska Gronostaj for their comments and encouragement. We also thank the unknown reviewer of the eLex conference 2013 for many useful comments that have helped us to elaborate on the present version of the paper.

9. References

- Alonso Ramos, M., Rambow, O. & Wanner, L. (2008). Using Semantically Annotated Corpora to Build Collocation resources. In: *Proceedings of LREC 2008*. Marrakech.
- Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M. & Kokkinakis, D. (2010). *The past meets the present in Swedish Framenet++*. CLT, Gothenburg.

- Brugman, C. (2001). Light verbs and polysemy. In: *Language Sciences 23*. pp. 551–578.
- Butt, M. (2003). *The Light Verb Jungle*. Workshop on Multi-Verb Constructions Trondheim, June 26–27, 2003.
- Cinková, S. (2009). A Contrastive Lexical Description of Basic Verbs. Examples from Swedish and Czech. In: *The Prague Bulletin of Mathematical Linguistics*, number 92, Dec. 2009. pp. 21–62.
- Dixon, R. M. W. (1991). *A new approach to English grammar, on semantic principles*. Oxford: Clarendon Press.
- Fillmore, C., Johnson, C., Petruck, M. (2003). Background to Framenet. In *International Journal of Lexicography*, 16(3), pp. 235–250.
- Goldberg, Adele E. (2003). Words by Default: the Persian Complex Predicate Construction. In: Elain Francis & Lauran Michaelis (ed.). *Mismatch: Form-Function Incongruity and the Architecture of Grammar*. CSLI Publications. pp.: 83–112.
- Grefenstette, G., Teufel, S., (1995). Corpus-based method for automatic identification of support verbs for nominalizations. In: *Proceedings of the Biannual Meeting of the EACL*, pp. 27–31.
- Hanks, P., Urbschat, A. & Gehweiler, E. (2006). German Light verb Constructions in Corpora and Dictionaries. In: *International Journal of Lexicography*, vol. 19(4), pp. 439–457.
- Karimi-Doostan, G. (1997). *Light Verb Constructions in Persian*. Doctoral Thesis. Dept. of Language and Linguistics. University of Essex.
- Langer, S. (2004a). A linguistic test battery for support verb constructions. In: *Lingvisticae Investigationes*. Volume 27, Number 2. pp. 171–184.
- Langer, S. (2004b). A Formal Specification of Support Verb Constructions. In: *Semantik im Lexikon*. Tübingen: Narr.
- Laporte, É., Ranchhod, E.M. & Yannacopoulou, A. (2008). Syntactic Variation of Support Verb Constructions. In: *Lingvisticae Investigationes*, vol. 31, number 2, pp. 173–185.
- Miyamoto, T. (1999). *The Light Verb Construction in Japanese. The role of the verbal noun*. Linguistics Today, 29. Amsterdam: John Benjamins.
- Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. ICSI Technical Report.
- Salkoff, M. (year unknown). Automatic translation of support verb constructions. <http://www.aclweb.org/anthology-new/C/C90-3043.pdf>
- Sköldberg, Emma 2004. *Cards on the table. Variations in content and expression in Swedish idioms*. Meijerbergs arkiv för svensk ordforskning 31. Göteborg.

From DOC Files to a Modern Online Dictionary

Tinatin Margalitadze, George Keretchashvili

Lexicographic Centre at Ivane Javakhishvili Tbilisi State University

Address: 1 Chavchavadze av. Tbilisi, Georgia

E-mail: tinatin@margaliti.ge, contact@dictionary.ge

Abstract

The aim of this paper is to describe the process of development of software for the Comprehensive English-Georgian Online Dictionary, posted on the Internet in 2010. The Dictionary engine is built on PHP/MySQL platform and combines three major branches: user interface, administrative interface and billing system, thus making it an integrated and dynamic resource. User functionalities include: bidirectional search; auto suggestions; auto corrections; online payments, etc. The administrative interface of the Dictionary holds a number of administrative functionalities, such as: dictionary vocabulary management functionalities; generation and conversion tools necessary for editors; user registration management functionalities, etc.

The Online Dictionary databases were generated from the DOC files which contained raw text data: words, grammatical characteristics of words, pronunciations and descriptions, altogether and separated by spaces just as in any sentence. After thorough analysis and testing, a special converter was written that would automatically analyze and separate raw data input into separate rows and fields. Our experience of transformation of the DOC files into a modern online resource may be interesting for the e-lexicography community. This paper will also discuss some other applications which are under development at the Lexicographic Centre.

Keywords: data transformation; online dictionary development; control panel.

1. History of the Dictionary

Work on the Comprehensive English-Georgian Dictionary (CEGD) began in the 1960s in the Department of English Philology of Tbilisi State University. In the 1980s, a small team of editors embarked on a thorough revision of the dictionary material and launched publication of the dictionary in fascicles (1995–2012). Currently printed and published are 14 fascicles of the English-Georgian dictionary (www.margaliti.ge), which cover 2,380 pages of the printed dictionary. The online version of the dictionary, posted on the Internet in 2010, is based on the aforementioned fascicles (www.dict.ge). The CEGD comprises 110,000 entries, covering several hundred thousand English meanings, collocations, phrasal verbs, idioms, and terms from different fields (T. Margalitadze, 2012).

One of the important issues faced by the editors of the CEGD has been ‘linguistic and cultural anisomorphism’ (Hartmann and James 1998: 51) between the English and Georgian languages, resulting in semantic asymmetry of seemingly similar words of these languages. Semantic asymmetry is even wider between genetically unrelated

and structurally different languages, as is the case with the Georgian and English languages. English-Georgian lexicography is not exceptional in this respect, as it is the central problem of bilingual lexicography at large. This issue, and the treatment of equivalence in the CEGD, was presented at the XV International Congress of EURALEX in Oslo (T. Margalidze, 2012).

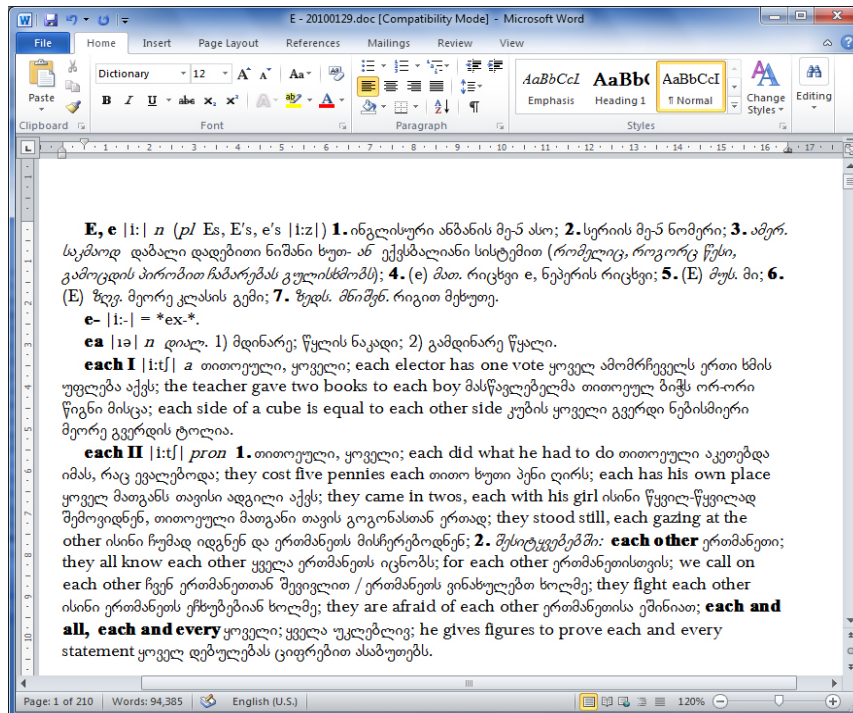


Figure 1: Text represented in MS Word document

The CEGD was not created in a Dictionary Writing System (DWS). In the 1990s, half of the dictionary, the compiled and edited entries, existed on cards (letters A–L). In 1993, the Lexicographic Centre started digitalization of the dictionary material and the first fascicle, the letter A, appeared in 1995. Back in the 1990s, there was not even a proper Georgian font with extended character support and a special font (“Dictionary”, see Figure 1) was created for the project. It is probably worth noting that the configuration of the Dictionary font was based on the Russian script, “Cyrillic”, changed into the Latin script several years later.

Dictionary cards were digitalized into the DOC files and in subsequent years the work continued in MS Word.

2. Data Transformation

As mentioned above, digitalized dictionary material, as well as the entries created later, existed in a formatted text edited by text processors like MS Word (see Figure 1).

The DOC files contained raw text data that included words, grammatical characteristics of words, and pronunciations and descriptions, altogether and separated only by spaces just as any sentence. The text was represented in a special, non-Unicode encoding and was slightly formatted (see Figure 1).

```
<p class=MsoNormal style='text-indent:14.2pt'><b><span style='font-family:DictionaryBold'>E,↓
e </span></b><span style='font-family:Dictionary'>|i:|<i> n </i><i>pl </i>Es,↓
E's, e's |i:z| </span><b><span style='font-family:DictionaryBold'>1.</span></b><span ↓
style='font-family:Dictionary'> ÈIÀÈÈÑÓÐÈ ÁIÀÁIÈÑ ÈÀ-5 ÁÑI;</span><b><span ↓
style='font-family:DictionaryBold'> 2</span></b><b><span style='font-family:↓
DictionaryBold'>.</span></b><span style='font-family:Dictionary'> ÑÀÐÈÈÑ ÈÀ-5↓
ÌÈÈÀÐÈ; </span><b><span style='font-family:DictionaryBold'>3</span></b><b><span ↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
ÑÈÑÓÀÈÈÇ (<i>DÍÈÀÈÈÚ, DÍÁÍÚ ÚÀÑÈ, ÁÀÈÍÚÀÈÑ ÌÈÐÍÀÈÇ ÚÀÁÀÐÁÁÁÑ ÁÓÈÈÑÐÈÍÁÑ</i>); </span><b><span ↓
style='font-family:DictionaryBold'>4</span></b><b><span style='font-family:↓
DictionaryBold'>.</span></b><span style='font-family:Dictionary'> (e) <i> ÈÀÇ.</i>↓
ÐÈÚÐÁÈ e, ÌÁÍÀÐÈÑ ÐÈÚÐÁÈ; </span><b><span style='font-family:DictionaryBold'>5</span></b><b><span ↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
(E) <i>ÈÓÑ</i>. ÈÈ; </span><b><span style='font-family:DictionaryBold'>6</span></b><b><span ↓
style='font-family:DictionaryBold'>.</span></b><span style='font-family:Dictionary'>↓
(E) <i>ÈÓÁ</i>. ÈÁÍÐÀ ÈÈÀÑÈÑ ÁÀÈÈ; </span><b><span style='font-family:DictionaryBold'>7.</span></b><i><span ↓
style='font-family:Dictionary'> ÈÁÑ. ÈÌÈÐÁÌ.</span></i><span style='font-family:↓
Dictionary'> ÐÈÀÈÇ ÈÁÐÓÇÀ.</span></p>
```

Figure 2: Data represented in HTML format

			"(<i>pl</i> Es, E's, e's [i:z])
			<p>1. ინგლისური ანბანის მე-5 ასო;
			<p>2. სერვის მე-5 ნომერი;
"E, e"	"n"	"i."	<p>3. <i>სამერ.</i> <i>საკომოდ</i> დაბალი დაღებიანი ნიშანი ხუთ- <i>ან</i> ექვსაბლიანი სისტემით (<i>რომელიც, როგორც წესი, გამოყენდება პირობით ჩაბარებას გულისხმობს</i>);
			<p>4. (<i>მათ.</i> რიცხვი e, ნეპერის რიცხვი;
			<p>5. (E) <i>მუს</i>. მი;
			<p>6. (E) <i>ზღვ</i>. მეორე კლასის გემი;
			<p>7. <i>ზღვს. მნიშვნ.</i> რიგით მეხუთე."
"e."	"i."	"i."	"=<i>ex</i>."
"ea"	"n"	"ie"	"<i>დიალ</i>. 1) მდინარე; წყლის ნაკადი; <p>2) გამდინარე წყალი."

Figure 3: Data ready for saving as Comma Separated Values (CSV) format

After thorough analysis and testing, a special converter was written that would automatically analyze and separate raw data input into separate rows and fields. Before dictionary data can be used for the database, the following procedures should be performed:

- A DOC file is prepared by replacing a couple of special symbols presented in the texts by other special symbols in order to be further interpreted as required;
- Then the file is converted into an HTML file, thus converting the initial text

into the data that can be parsed by converter (see Figure 2);

- The HTML file is slightly cleaned manually and submitted for conversion;
- Converter runs through the file structure and indicates errors if found;
- After the errors have been corrected, the converter parses the file and makes all necessary conversions that might include more than 20 conversions for each word set;
- Then the data is split into different fields, and special formatting is applied which outputs it in the CSV (Comma Separated Values) format (see Figures 3 and 4);
- After the CSV file is generated it can be imported into any database.

Even after inserting data into the database, several scripts are run over newly-inserted records in order to achieve the database consistency and to provide efficient search results. Final data can be later directly edited through the Dictionary Control Panel.

```
"E, e" "n" "i:" "(<i>p1</i> Es, E\'s, e\'s [i:z]) ↓
↓
<p><b>1.</b> იმელსური ანანის მე-5 ასო; ↓
↓
<p><b>2.</b> სერის მე-5 ნომერი; ↓
↓
<p><b>3.</b> <i>აგრ.</i> <i>საკმაოდ</i> დანალი დადებითი ნიშანი ხელ- <i>ან</i> ექსპალიანი სისტემით (<i>რომელიც, როგორც წესი, გამოყენის პირობით ჩანარებას ეულისხმობს</i>); ↓
↓
<p><b>4.</b> (e) <i>მათ.</i> რიგები e, შეპრის რიგები; ↓
↓
<p><b>5.</b> (E) <i>მეს</i>. მი; ↓
↓
<p><b>6.</b> (E) <i>ზეც</i>. შორე კლასის ეპი; ↓
↓
<p><b>7.</b> <i>ზედს. მნიშვნ.</i> რიგით მუხვით."↓
"e-" "n" "i:-" "= <r>ex-</r>."↓
"ea" "n" "ra" "<i>დალ</i>. 1) მდინარე; წლის ნაკადი; 2) გამდინარე წალი."↓
"each I" "a" "i:tf" "თითოეული, ყველი; each elector has one vote ყველ ამომრჩევლის ერთი ხმის უფლება აქვს; the teacher gave two books to each boy მასწავლებელმა თითოეულ ბიჭს ორ-ორი წიგნი მისცა; each side of a cube is equal to each other side კუბის ყველი გვერდი ნებისმიერი შორე გვერდის ტოლია."↓
```

Figure 4: Data in Comma Separated Values (CSV) format

3. Online Dictionary

The Comprehensive English-Georgian Online Dictionary (CEGD) is a unique, hand-written web based application easily accessible from any Internet-enabled device, such as computers, cell phones, tablets etc. The Dictionary engine is built on PHP/MySQL platform and combines three major branches: user interface, administrative interface and billing system, thus making it an integrated and dynamic resource (see Figure 5).

During the first year of the operation some new functionalities were added to the program: the user interface became bilingual, a drop-down bilingual suggestion feature was added to the search box, an auto correction/suggestion system was

implemented to correct typos, search backend was improved, entry layouts were improved for easier reading, colors and tooltips were implemented for abridgements, video tutorials were added to the user guide, etc. There is an online feedback form available to provide support for users with technical or other issues.



Figure 5: CEGD

Both the database and the engine of the CEGD are in the process of constant upgrading and improvement in order to provide the users with an up-to-date, user-friendly, safe and perfect product.

3.1 User functionalities

The bilingual user interface front- and backends hold two categories of functionalities. One combines generic system screens and functionalities like user registrations, profile editing functions, safe logins, password resets, news etc.

The other part of the system is responsible for bidirectional search (the engine includes the search functions that make it possible to look up both English and Georgian words and phrases despite the fact that the dictionary vocabulary database is one way: English to Georgian only); auto suggestions; the search engine also includes auto suggestions on spelling errors, etc. While viewing any particular word and its translation, next, previous and several nearby wordlists appear for easier navigation; words can be listed and navigated by letters, etc.

Though all the interfaces and functionalities were designed to be intuitive and easy to use, the CEGD is supplied with a user's guide with detailed textual and video instructions on how to search for English words, collocations, phrasal verbs, and idioms, as well as Georgian words and phrases. The user guide also explains the structure and organization of the entries and other details.

3.2 Administrative functionalities

Administrative screens and functions are designated for editors, managers, technical administrators and other personnel who support online dictionary operations.

The following functionalities are available through the CEGD Control Panel:

- Dictionary vocabulary management, including the functions of viewing and editing the dictionary vocabulary (see Figure 6), as well as the function of adding new entries;
- Generation and conversion tools necessary for editors (see Figure 7);
- User registration management;
- Statistics including registrations, search logs, etc.

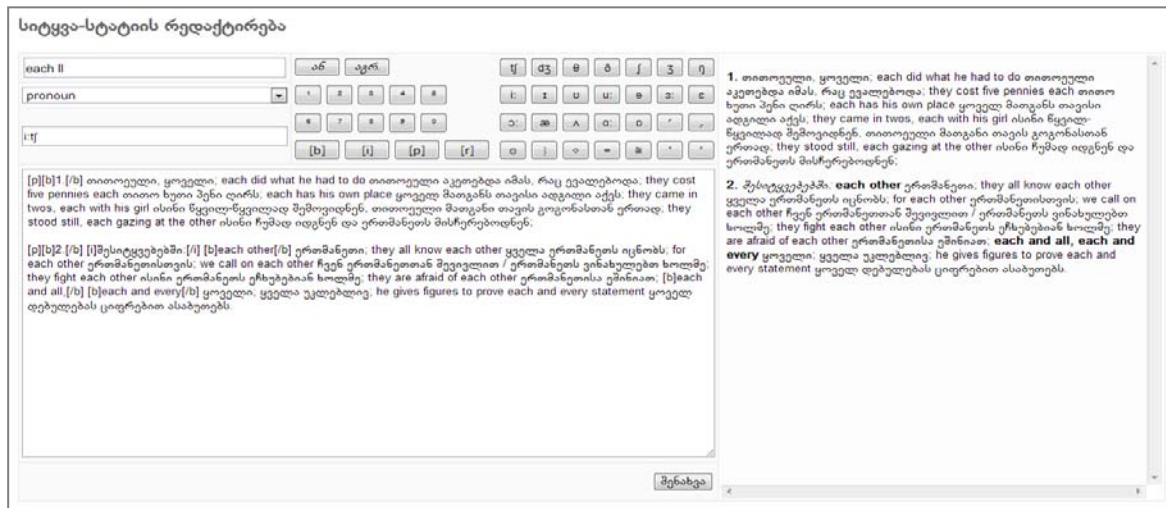


Figure 6: Editing an entry from CEGD Control Panel



Figure 7: Converter for editors

4. Specialized Dictionaries

After its successful launch, as a result of expertise gained over more than a year of operation of the CEGD, and based on accumulated experience including user feedback, a number of improvements were applied to the core engine of the CEGD: backend search functions and database extension tables were redesigned and rewritten to provide improved performance. Frontend and search result pages were also slightly modified and a clean, light version of the core engine was used for smaller specialized dictionaries of the Lexicographic Centre, namely for the “English-Georgian Military Dictionary” (<http://mil.dict.ge>), compiled in 2009 at the request of the Georgian Ministry of Defense and posted on the Internet in 2011 (see Figure 8), and the “English-Georgian Biology Dictionary” (<http://bio.dict.ge>), the current project of the Lexicographic Centre, financed by Shota Rustaveli National Science Foundation of Georgia.



Figure 8: English-Georgian Military Dictionary

Light versions of the Online Dictionary Application operate in the same way as the CEGD system.

5. Future software projects

Currently the Lexicographic Centre is working on the development and improvement of its web applications, and also on the development of new software tools and solutions for the projects of the Centre.

5.1 Desktop Application

Web applications are very common nowadays in this country. However, there are cases where web application is not the right solution and the user prefers a locally installed desktop application. This fact led to our decision to develop a desktop application for online dictionaries. Work on the first version of the electronic dictionary, i.e. the desktop application is already completed and is being tested (see Figure 9).

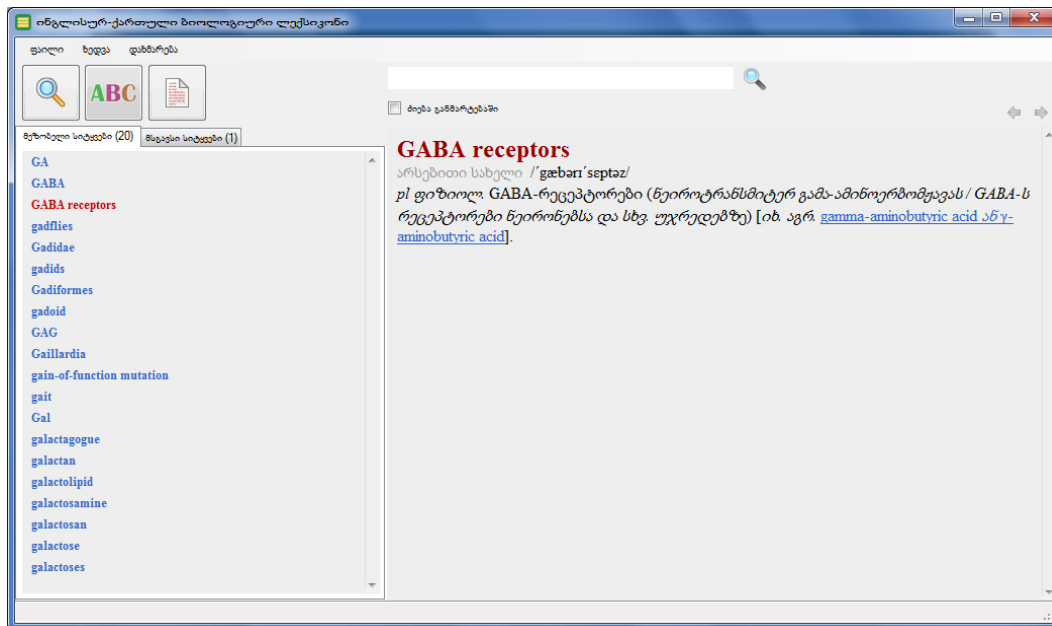


Figure 9: Desktop application (Pre-alpha version)

Desktop application includes predominantly the same functionalities as online dictionaries of the Lexicographic Centre, described above, but will be easier to access and use even in the case of slow or bad Internet connection. Unlike web applications, desktop applications will be integrated into the user's computer and will enable the addition of the functionality of directly translating words from many other applications (like word processing applications) by simple clicks or using keyboard combinations. Being offline does not mean being outdated: the desktop application database will have the functionality of being updated from the Internet, as the Lexicographic Centre regularly releases new updates of its online databases.

5.2 Dictionary writing system

As mentioned above, dictionary creating processes were conducted in the Lexicographic Centre with very limited technical resources, which required much effort to work on the data in the past. Nowadays, modern technologies offer more options and possibilities to maximize results and add more functionality and manipulation options to the dictionary data. As it was becoming more and more difficult and uncomfortable to handle Word files, the Lexicographic Centre has

launched the development of a Dictionary writing system. When this project is completed and the existing dictionaries are integrated into it, this will allow the Lexicographic Centre to add to its products synonyms, antonyms, and pictures, to apply different fonts and colors, as well as adding other functions essential to modern dictionary databases.

5.3 Mobile Application

Modern mobile devices like smartphones and tablets are becoming more and more popular in this country and are essential for students and business people, etc. In order to bring comfort and simplicity to those users, the creation of special applications are planned in order to meet mobile device requirements.

5.4 Lightweight interface of online dictionaries

Some mobile users prefer websites instead of downloading and installing applications on smartphones or tablets. As mobile devices are usually smaller in size and have limited interaction options compared to personal computers, the creation of lightweight interfaces, specially designed for mobile use are planned at the Lexicographic Centre.

6. References

- Comprehensive English-Georgian Online Dictionary. (2010). T. Margalitadze (Editor-in-Chief) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. <http://www.dict.ge>
- English-Georgian Online Biology Dictionary. (2012). T. Margalitadze (Editor) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. bio.dict.ge.
- English-Georgian Online Military Dictionary. (2011). T. Margalitadze (Editor) et al. Lexicographic Centre at Tbilisi State University, Tbilisi. mil.dict.ge.
- Hartmann, R. R. K. and G. James. (1998). Dictionary of Lexicography. London: Routledge.
- Margalitadze T. (2012). The Comprehensive English-Georgian Online Dictionary: Methods, Principles, Modern Technologies. Proceedings of the XV EURALEX International Congress. Oslo, Norway. http://www.euralex.org/elx_proceedings/Euralex2012/pp764-770Margalitadze.pdf

Online Style Guide for Slovene as a Language Resources Hub

**Simon Krek¹, Helena Dobrovoljc²,
Kaja Dobrovoljc³, Damjan Popič⁴**

¹Jožef Stefan Institute, Ljubljana, Slovenia

²Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia

³Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia

⁴Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia

E-mail: simon.krek@ijs.si, helena.dobrovoljc@zrc-sazu.si,

kaja.dobrovoljc@trojina.si, damjan.popic@ff.uni-lj.si

Abstract

There exists a long tradition of orthography guides or style manuals for Slovene dedicated to "good writing" (Slo. *pravopis*, Ger. *Rechtschreibung*), with the first one published in 1899 and the most recent in 2001. The new web portal developed within the Communication in Slovene project is taking the concept originating from the world of print one step further into the digital environment, with a question-answering system which analyses the question entered into a query window in natural language and aims to provide a three-layered answer, from a more condensed and graphical one using data from extensive corpora, lexicons, dictionaries and other online resources, to a more general user-friendly description of the problem, together with links to digitized modern and historical normative resources related to the identified language problem. The paper describes a demo version of the portal with demonstration data for 15 language problems.

Keywords: Slovene language; orthography; online style guide; language resources portal; question-answering;

1. Introduction

The basic idea of the portal¹ is to provide information about the Slovene language and the problems that average speakers have with its written norm. It is not intended only for language specialists or professionals but for all web users. The portal uses new (language) technologies now available also for Slovene and aims to complement printed orthography guides from Levec (1899) to Toporišič (2001) with a dynamic web portal based on empirical data from various extensive language resources. The concept is based on the analysis of language use in text corpora and frequent questions in web forums dedicated to language problems, at the same time also providing information from traditional orthography guides and other historical resources. The most important extensive new digital language resources used on the portal are Sloleks morphological lexicon (Grčar et al., 2013) and Gigafida corpus (Logar Berginc et al., 2012).

¹ <http://slogovni.slovenscina.eu/>

2. Background

Similar to other languages (Mønnesland 1998: 1103) Slovene has a relatively long tradition of written language codification embodied in official orthography guides in the entire 20th century. These guides have usually included an extensive dictionary section, with an emphasis on orthographically challenging vocabulary (cf. Verovnik 2004: 254). The last orthography guide in the series was published in 2001 in printed form, on CD-ROM in 2003, and has been available online since 2010. The content of the digital version replicates the printed one, the rules are available as a PDF document, and dictionary content can be searched in the search engine NEVA,² on the Termania dictionary portal,³ and in ASPplus software,⁴ all of them also allowing more complex queries.

One of the assumptions of the authors of the new portal is that the advent of the web, with the possibility of massive participation of users in the creation of texts (blogs, forums, social networks, etc.) that are immediately available to be read or commented on, radically changed the nature and dynamics of the text publication process. In post WWII Slovenia, this process has typically included the author, the publishing house with its editor, the proof-reader, and a language specialist called "lektor" responsible for the compatibility of published texts with the language norm or standard.

In the world of print, texts have traditionally been handled by a relatively narrow circle of professionals, including language specialists. However, with the possibility to publish texts online without the assumed or axiomatic interference of third parties, this cycle is now more or less broken. In addition, the time needed from the creation of the text to its publication has been reduced to just a few seconds, and numerous genres previously reserved for private communication are now part of the public sphere (Crystal, 2011). This has created the need to also present information about language standard to the general public, not just language professionals, preferably in a user friendly manner. Therefore, if previous orthography guides effectively belonged to the world of print, the new web portal aims to provide an answer to the question of how language codification should be presented in the digital (web) environment of the 21st century.

In the new environment, codification-related language help currently comes from two basic sources. The first one comprises spelling or grammar checkers and similar tools which can be seen to replace the proof-reader in the printed environment. The other sources are online portals, dedicated forums and now also social networks, or

² <http://bos.zrc-sazu.si/sp2001.html/>

³ <http://www.termania.net/slovarji/20/slovenski-pravopis/>

⁴ <http://www.amebis.si/aspplus/>

search engines, providing consultation or feedback from both peer communities and official bodies responsible for language codification. The new web portal aims to answer the need for consultation by providing standardized explanations of the most frequent problems with language or (more narrowly) spelling and orthography.

3. List of language problems

The portal consists of several parts, with a list of around 700 detected language problems functioning as the central database. The list was created by analyzing traditional orthography guides, text corpora and web forums specialized in language problems. Web forums were crawled and each question was manually assigned to a particular category. Also, special data mining procedures were established which produced lists of variant forms of words where speakers (or writers) of Slovene falter due to inappropriate, unrecognized or non-existent norms. The main task in this process was to establish a list of real language problems and balance it suitably between overgeneralization and excessive fragmentation of categories. All categories were later organized as an ontology with eight top categories: orthography (A), orthoepy (B), morphology (C), word-formation (D), vocabulary (E), syntax (F), text (G), and other (H). Current ontology extends to six levels from top to bottom, with variable granularity. Levels are formally labelled as combinations of letters and digits, as shown in Table 1.

LABEL	CATEGORY
D	word-formation
D1	adjectives
D1a	possessive adjectives from names of masculine gender
D1a1	from names ending in vowels
D1a1a	from names ending in -a
D1a1b	from names ending in unpronounced -e
D1a1c	from names ending in -y

Table 1: An example of language problems ontology

4. Three-layered configuration of answers

Each of the bottom-level categories in the ontology is linked to several elements in the database, with the “short” and “long” answers (see Figure 1) the most important ones.

4.1 Short answer

The short answer consists of text in XML format which can generate a formulaic textual answer with relevant statistical data from the corpus and the lexicon. It is designed as a universal mechanism for the (statistical) description of all possible combinations of standard and non-standard word forms belonging to one particular category. For further clarification, category D1a2e will be used as an example:

LABEL	CATEGORY
D	word-formation
D1	adjectives
D1a	possessive adjectives from names of masculine gender
D1a2	from names ending in consonants
D1a2e	from names ending in pronounced -r

Table 2: Example – category D1a2e

The full title of the D1a2e category is “Word-formation of possessive adjectives derived from names of masculine gender ending in pronounced –r”. Examples of (foreign) surnames in Slovene belonging to the category are Shakespeare, Baudelaire, etc. Most of the adjectives derived from these names have two variant forms with alternative endings *-jev* and *-ov*: *Shakespearejev* | *Shakespeareov*, *Baudelaireov* | *Baudelairejev*. Since the final unpronounced -e has to be dropped in the derivation process according to the standard, essentially changing the exact form of the original name, two non-standard forms are used frequently enough to be included in the lexicon: *Shakespearejev* | *Shakespeareov*, *Baudelairejev* | *Baudelaireov*. Therefore, there are four potential forms that have to be taken into account when creating the short answer for this category. As it is not necessary that all four forms actually appear in the corpus for all possible names in this category, a combination of 15 answers have to be included in the short answer. Table 3 shows the first four:

<pre><!-- variant 1: FOUR, standard-12, non-standard-34 --> <text var="Soo.Soo.Noo.Noo" graph="1234">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/>, <word id="3"/> and <word id="4"/> in the Gigafida corpus. Word forms in blue colour are standard, those in grey are not compatible with the current standard of written Slovene.</text></pre>
<pre><!-- variant 2: THREE, standard-12, non-standard-3 --> <text var="Soo.Soo.Noo" graph="123">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/> and <word id="3"/> in the Gigafida corpus. Word forms in blue colour are standard, the grey one is not compatible with the current standard of written Slovene.</text></pre>
<pre><!-- variant 3: THREE, standard-12, non-standard-4 --> <text var="Soo.Soo.Noo" graph="124">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/> and <word id="4"/> in the Gigafida corpus. Word forms in blue colour are standard, the grey one is not compatible with the current standard of written Slovene.</text></pre>
<pre><!-- variant 4: THREE, standard-1, non-standard-34 --> <text var="Soo.Noo.Noo" graph="134">The graph shows the data about the use of word forms <word id="1"/>, <word id="3"/> and <word id="4"/> in the Gigafida corpus. The word form in blue colour is standard, those in grey are not compatible with the current standard of written Slovene.</text></pre>

Table 3: Short answer in XML

Word forms shown in the textual part of the short answer (as opposed to the graph) are taken from the Sloleks lexicon which also contains statistical data from the Gigafida corpus. In each particular case, the system chooses the relevant short answer automatically in accordance with the lexicon data. The design of short answers therefore enables an upgrade of the corpus which is directly reflected on the portal through the upgrade of the data in the lexicon. Once the set of possible short answers is written for a particular language problem, it is not necessary to update the text of the answer again manually, as the system chooses the right answer according to the status found in the regularly updated lexicon. This makes the system dynamic and linked to external independent resources, which can be updated regularly. Where this is applicable, data from the lexicon/corpus are also shown in a graph. For visualization of the data, the portal uses Google Charts tools, as shown in the upper part of Figure 1.

4.2 Long answer

In contrast to short answers, which constitute the dynamic part of the portal linked to external resources, long answers are essentially static. Each identified problem in the ontology receives one long answer which is written in HTML format and included in the central database. When creating the system, special attention was given to wording, length, formatting and other features, to ensure that long answers are particularly useful for general users, who are the primary target audience of the portal, rather than language professionals.

Long answers (the middle part of Figure 1) can contain three types of links each with a different function:

- blue, italic, bold: link to an external resource, which can be a corpus, lexicon or other web resource such as Wikipedia, etc.
- blue, underline: pop-up window with an explanation of a linguistic term when its use is unavoidable in the long answer.
- blue, dotted underline: pop-up window with the list of words belonging to the same category, with the same orthographic problem.

Long answers are designed to provide the user with general information about the problem in lay terms, and contain links to other available resources that we consider useful for the user. This part of the portal has an explicitly educational function, as it is expected for the user to understand the problem and be able to interpret in the future.

Tvorba svojilnih pridevnikov iz moških imen, ki se končajo na govornji [r]

KRATKO IN JEDRNATO

CATEGORY: Word-formation of possessive adjectives derived from names of masculine gender ending in pronounced -r

Na grafu si lahko ogledate podatke o rabi oblik *Shakespearjev*, *Shakespearov*, *Shakespearejev* in *Shakespeareov* v korpusu Gigafida. Obliki, zapisani z modro, sta ustrezni, sivi pa nista skladni s trenutnim pravopisnim standardom.

SHORT ANSWER: explanation of the derived forms from the name "Shakespeare"

data from the lexicon and the corpus

Form	Frequency	Category
Shakespearjev	682	standard forms
Shakespearov	3,051	standard forms
Shakespearejev	26	non-standard forms
Shakespeareov	23	non-standard forms

LONG ANSWER

NA DOLGO IN ŠIROKO

Na splošno **svojilne pridevnike** iz samostalnikov moškega spola naredimo tako, da jim dodamo **-ov** ali **-ev**, pri čemer je izbira **odvisna od glasu**, s katerim se samostalnik konča. Izjema so svojilni pridevniki iz samostalnikov, ki se končajo na izgovorjeni r. Pri teh imamo dve enakovredni možnosti.

- Lahko jih podaljšamo z -j, kar pomeni, da bomo zaradi **preglasa** uporabili **-ev**, npr. *novinar* – *novinarjev*, *Gregor* – *Gregorjev*.
- Lahko pa jim dodamo **-ov**, npr. *satir* – *satirov*, *Bor* – *Borov*.

Raba ene od obeh možnih oblik se je pri večini pridevnikov ustalila in na splošno prevladuje oblika z **-jev**. Obliko z **-ov** pa skoraj vedno uporabimo v dveh primerih:

- kadar pri samostalniku pred končnim izgovorjenim r stoji **polglasnik**, zapisan s črko, ki pri sklanjanju izgine (*Peter* → *Petra*, *Petru* ... → *Petrov oče*).
- pri **enzložnih** samostalnikih (*Bor* → *Borov*).

Obstajajo tudi **posamezna imena**, pri katerih izbira precej niha. Če nas zanima, katero obliko pisci raje uporabljajo, se o tem lahko prepričamo v **korpusu**.

Poseben problem so angleška, francoska in nekatera druga **lastna imena**, ki se končajo z izgovorjenim r, vendar črki r sledi še **nemi -e**, kot na primer v priimkih *Molière* [moljêr], *Baudelaire* [bodlêr], *Saussure* [sosír], *Gilmor* [gílmor], *Shakespeare* [šékspir] in **podobnih**. Pri tvorjenju svojilnih pridevnikov iz teh lastnih imen **nemi -e** praviloma izpustimo in dodamo **-jev** ali **-ov**, kar pomeni, da sta ustrezni obe obliki, npr. *Molièrov/Molièrjev*, *Saussurov/Saussurjev*, *Baudelairov/Baudelairjev* itd. Pri nekaterih imenih se je raba ustalila pri eni od možnosti, pri drugih pa se uporabljata obe obliki, vendar ena navadno prevladuje. Podatek o tem je mogoče dobiti v **korpusu**.

Za uspešno reševanje zadrege moramo poznati pravičen izgovor tujega imena!

FOR ENTHUSIASTS": links to scholarly works related to the particular problem

ZA NAVDUŠENCE

Slovenski pravopis – pravila (2001):

- Stran 89 - Posebnosti 1. moške (o-jevske) sklanjatve – krajšanje osnove
- Stran 90 - Posebnosti 1. moške (o-jevske) sklanjatve – daljšanje z j
- Stran 114 - Težji primeri iz besedotvorja (pridevnik) – priponsko obrazilo -ov/-ev oz. -in

Preverite tudi, kaj o vašem iskalnem pogoju pravijo **digitalizirani slovenski pravopisi in starejše slovnice**, ki so izšli v obdobju od 1899 do 2001.

Figure 1: Screenshot of the query result on the portal

4.3 Links for enthusiasts

The third part of the answer is titled “For enthusiasts” and provides links to scholarly works related to the particular problem or to orthographic problems in general. The most important document in this section is the official orthographic rules book published in 2001 and available online in PDF format. Other important works include previous orthographic guides which were digitized in another project and published online independently,⁵ and are also included on the portal. This part of the portal provides more advanced users with the possibility to explore the historical background of the problem encountered.

5. Access to information on the portal

Information on the portal can be accessed in two ways: first, by entering a query in natural language which is parsed and matched with the data in the lexicon. Parsing is performed by a rule-based tagger and parser owned by the Amebis software company.⁶ Individual word forms and lemmas from the query are compared with lexicon entries that contain information about a category from the ontology of language problems. If a match is found, the corresponding answer is shown on the portal. If there is more than one match, other possibilities are shown as links in the “Did you mean?” section on the left side of the main frame. As some problems in the ontology are related to each other by default, if one is found, the others are shown in the “Linked answers” section.

The second option for accessing information is to browse the ontology on the index page which can be accessed by clicking the “See the index” link on the home page. Users who wish to go through the entire portal systematically can use this feature.

6. The corpus and the lexicon

The most important relationship, enabling the system to work as designed, is that between the ontology—with its formal hierarchy of labelled language problems—and the Sloleks lexicon containing extensive amounts of data about morphology, together with information about language norm assigned to its various elements. Gigafida corpus, on the other hand, as the source of statistical data for the lexicon, does not contain normative information. It is lemmatized and POS-tagged in a standard manner using the newly-developed Obeliks tagger and lemmatizer (Grčar et al., 2012).

⁵ Available at: <http://www.trojina.org/pravopisi/>

⁶ Web site: <http://www.amebis.si/>

The lexicon uses Lexical Markup Framework (LMF) format which allows various kinds of information to be included on every level, either assigned to the whole lexical entry or to one particular word form. These types of information can range from pronunciation or stress to normative information. One particular instance of the lexicon, i.e. lexical entry, becomes a part of the portal only when it is assigned with a particular language problem from the ontology. Without explicit information it is invisible to the system. The annotation of normative information in the lexicon is currently performed semi-automatically or manually, as this kind of information is too sensitive to be included in a fully automatic manner without checking.

6.1 Extraction of data from the corpus

In order to obtain a candidate list of lexicon entries for a particular language problem, an extraction procedure is applied to the corpus. To explain the procedure in detail, category C1a3b will be used: “Declension of (foreign) names of masculine gender with the ‘unsteady vowel’”. Examples of such names in Slovene are Russell, Powell or Robben, Bremen which lose their final [e] in some grammatical cases: *Russlla*, *Powlla* or *Robbna*, *Bremna*. Since this rule can produce rather unusual forms with a series of consonants, Slovene writers often use the final [e] in inflected forms: *Russella*, *Powella* or *Robbena*, *Bremena*.

To extract relevant names from the corpus, in order to decide which names will be later included in the lexicon, all types in the corpus are split into three parts: the root (open set), the inflections (closed set) and the variable part (closed set). Based on the variability of the middle part and the invariability of the other two, pairs of types are produced, together with frequency data. The more equally the variable part is distributed between both possible forms, the more interesting the pair. When the extracted pairs are ranked according to the combination of frequency and variability using statistical data from the corpus, a list shown in Table 4 is produced. As this category covers different combinations of an ‘unsteady vowel’ + a consonant (en/-n-, -ek/-k-, -ic/-c-, -ell/-ll-, etc.), for each consonant pair a separate list is prepared. Table 4 shows the top 20 candidates for the **en/-n-** pair. These traditionally include names of Scandinavian or Germanic origin which is also confirmed on the extracted and ranked list.

Extraction of corpus data enables the portal to offer information about the most challenging and frequent names belonging to this category, and on the other hand, long-lived examples from traditional resources can be replaced with modern and relevant ones in long answers.

Root	Lemma (artificial)	Frequency in Gigafida		Score
		root + -en- + inflection	root + -n- + inflection	
Klem	Klemen	1843	3839	0,46
Lor	Loren	908	505	0,29
Berg	Bergen	208	375	0,25
Niels	Nielsen	164	120	0,25
Test	Testen	501	2326	0,24
Robb	Robben	163	333	0,24
Natlač	Natlačen	223	147	0,23
Gold	Golden	37	29	0,21
Gall	Gallen	105	148	0,20
Ols	Olsen	112	64	0,20
Bid	Biden	102	117	0,20
Bjorndal	Bjorndalen	112	163	0,20
Franz	Franzen	117	114	0,19
Jens	Jensen	138	60	0,19
Patt	Patten	85	113	0,19
Hag	Hagen	74	120	0,19
Brem	Bremen	220	1509	0,18
Hold	Holden	60	147	0,18
Jem	Jemen	196	1319	0,18
Bed	Beden	769	164	0,18
Dresd	Dresden	194	1410	0,18

Table 4: Names extracted from the corpus and ranked according to frequency and variability

6.2 Manual analysis of corpus data

In some cases, extracted lists do not need further analysis and can be used for lexicon upgrade immediately. However, in most cases they are treated as candidate lists which have to be checked manually, either to validate data (corpus noise) or because different variants have to be attributed with unpredictable normative labels. For this purpose, the crowdsourcing platform sloCrowd (Tavčar et al., 2012) is used. The system supports annotator authentication and supervision, as well as quality control through random check based on gold-standard data. To explain the procedure in more detail we will use category C1a3f (Table 5):

LABEL	CATEGORY
C	morphology
C1	nouns
C1a	nouns of masculine gender
C1a3	nouns of masculine gender ending in vowels
C1a3f	names ending in -y

Table 5: Example using category C1a3f

This category is dedicated to (foreign) names ending in written [y] pronounced either as /ɪ/ or /e/, or a diphthong /aɪ/, /ɔɪ/, etc., such as Harry, Sydney, Playboy, Orsey, etc. In the Slovene declension system, these nouns are treated differently if they are pronounced with the final single vowel or a diphthong. In the first case, standard inflections are extended with a -j- before the inflection while in the second case this is not needed since the diphthong itself is considered to contain the sound /j/ in Slovene. Therefore, the examples mentioned above have the following forms in genitive case singular: *Harryja*, *Sydneyja*, *Playboya*, *Orseyja*. *Playboy* is pronounced with a final diphthong and has a regular inflection; others have to be extended with the medial -j.

The initial extracted list contains all names with the final written y. However, those with the consonant + y combination can be excluded from manual analysis as their pronunciation is predictable, and therefore both standard and non-standard inflectional paradigms are predictable and can be included in the lexicon automatically. With names ending in the vowel + y combination pronunciation is not predictable and manual procedure is needed to determine first the standard pronunciation of the foreign name, and based on that, the standard or non-standard inflectional paradigms.

For this purpose, a task is defined in the sloCrowd software, as shown in Figure 2, and results are obtained based on three or five decisions depending on the difficulty of the task. In the pilot project, around 100 students from the Faculty of Arts (Department of Translation) at the University of Ljubljana worked on approximately 8,000 extracted names in 10 tasks.

The screenshot shows the sloCrowd interface. At the top, the logo 'sloCrowd' is displayed with the tagline 'Sodelujte pri oblikovanju Slogovnega priročnika'. Below the logo are navigation tabs: 'ZAČETNA STRAN', 'PREVERJANJE LASTNIH IMEN', 'LESTVICA UPORABNIKOV', and 'INFO'. The main content area is titled 'Preverjanje lastnih imen'. It contains a paragraph explaining the task: 'Pri tej nalogi poskušamo ločiti **lastna imena**, torej imena oseb, krajev in stvari, pri katerih se **končna črka y izgovori kot soglasnik j** (kot pri imenu Broadway [brodvej]), od lastnih imen, pri katerih se **končni -y izgovori kot samoglasnik – bodisi kot i** (kot pri imenu Disney [diznij]) **bodisi kot e** (kot pri imenu Orsay [orsej]). Če končni y izgovorimo kot j, izberite možnost DA, če pa ga na koncu imena izgovorimo kot samoglasnik (i ali e), izberite možnost NE. Če ne veste, kako se ime izgovori, izberite možnost NE VEM.' Below this text is a question: 'Ali na koncu imena y izgovorimo kot j?'. A blue box contains the word 'beseda Sydney'. At the bottom, there are three buttons: 'DA' (with a green checkmark), 'NE' (with a red X), and 'Ne vem' (with a green question mark). A progress bar at the bottom shows '0%'.

Figure 2: Screenshot of a task in the sloCrowd crowdsourcing software

6.3 The lexicon

Sloleks lexicon is an independent language resource in the LMF (XML) format and can be found at different web addresses, both for downloading and for searching.⁷ Elements from the lexicon become part of the portal if they contain information about a category from the ontology of language problems (attribute “SPSP”), normative labels (attribute “norma”) and norm types (attribute “tip”). This additional information is added to the standard information which includes the description of formal morphological features of lemmas and word forms: morphosyntactic descriptions or MSDs.

Attribute “norma” (=norm) can have three values: *non-standard*, *variant* or *unclear*. *Variant* is used when several alternative forms can be used according to the standard, and *unclear* is used when the normative status of a lemma or word form cannot be determined due to conflicting information in the rules and dictionary parts of the official orthography guide. The absence of the attribute signifies that the lemma or word form is standard.

Attribute “tip” is used for differentiating between two or more possible morphological paradigms within one lexical entry, and related to one category, as shown in the example from lexicon in Figure 3. The lemma denotes the Slovene masculine name “Matija” which has two legitimate inflectional paradigms; therefore, the value in the attribute “norma” is *variant*. The two possible forms for genitive singular (=morphosyntactic description *Slmer* in the “msd” attribute) are *Matija* and *Matije*. The first paradigm is differentiated from the other using the attribute “tip” with the value which includes the category label, “s” for “standard form” and a sequential number for each paradigm.

Lexicon as a resource linking the portal and the corpus is used primarily for top level categories *orthography*, *word-formation*, *morphology* and *orthoepy*, and less commonly for *syntax*, *vocabulary* and *text*. For the latter three categories, data are generated either directly from the corpus or are not required, as answers are general enough to be limited to the long answer itself without the need for more detailed explanations.

⁷ Download at: <http://www.slovenscina.eu/sloleks/prenos> or search: <http://www.slovenscina.eu/sloleks>.

```

<LexicalEntry id="LE_S_Matija" xmlns:d="urn:LEKSIKON_SSJ">
  <feat att="besedna_vrsta" val="samostalnik" />
  <feat att="vrsta" val="lastno_ime" />
  <feat att="spol" val="moški" />
  <feat att="SPSP" val="C1a2a" />
  <Lemma>
    feat att="zapis_oblike" val="Matija" />
  </Lemma>
</...>
<WordForm>
  <feat att="število" val="ednina" />
  <feat att="sklon" val="rodilnik" />
  <FormRepresentation>
    <feat att="zapis_oblike" val="Matija" />
    <feat att="msd" val="Slmer" />
    <feat att="SPSP" val="C1a2a" />
    <feat att="norma" val="variantno" />
    <feat att="tip" val="C1a2a_s_1" />
    <feat att="pogostnost" val="858" />
  </FormRepresentation>
  <FormRepresentation>
    <feat att="zapis_oblike" val="Matije" />
    <feat att="msd" val="Slmer" />
    <feat att="SPSP" val="C1a2a" />
    <feat att="norma" val="variantno" />
    <feat att="tip" val="C1a2a_s_2" />
    <feat att="pogostnost" val="4018" />
  </FormRepresentation>
</WordForm>
</...>
</LexicalEntry>

```

Figure 3: Sample from the lexicon in Lexical Markup Framework format

7. Conclusion

This article describes a new web portal dedicated to problems with Slovene orthography, and includes in its demonstration version data for 15 language problems in Slovene selected from the approximately 700 problems identified by analysing traditional reference books, web forums and different extensive text corpora. The portal uses two resources to present information about real modern Slovene to the users of the portal in a user-friendly manner: the 1.2 billion-word corpus Gigafida, and the Sloleks morphological lexicon with 100,000 lemmas, together with their inflectional paradigms.

The portal is built around a central database with the 700 language problems organized in an ontology with eight top-level categories. These categories are used to identify relevant parts of the lexicon with normative information, which enables the

system to use both lexicon and corpus data to present normative information on the portal in a standardized manner. This comprises three types of answers: the short answer with statistical data, also supplied in graphical form; the static long answer for each of the bottom-level categories; and links to scholarly books and documents for experts and enthusiasts. The article describes both the portal and the extraction of relevant word forms and lemmas from the corpus, which are later assigned with normative labels and included in the lexicon, also using crowdsourcing in the process.

8. Acknowledgements

This article is based on the work of the Communication in Slovene project, which is part-financed by the European Union, the European Social Fund, and the Ministry of Education, Science and Sport of the Republic of Slovenia. The operation is being carried out within the operational programme Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013.

9. References

- Crystal, D. (2011). *Internet linguistics: a student guide*. New York, Routledge.
- Grčar, M., Krek, S., Dobrovoljc, K., (2012) Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012*. Institut Jožef Stefan, Ljubljana, pp. 42-47.
- Levec, F. (1899). *Slovenski pravopis*. Na Dunaju: cesarska kraljeva zaloga šolskih knjig.
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Trojina, Ljubljana.
- Mønnesland, S. (1998). Emerging Literary Standards and Nationalism. The Disintegration of Serbo-Croatian. *Actas do I Simposio Internacional sobre o Bilingüismo*. 1103–1113.
- Tavčar, D. Erjavec, T., Fišer, D. (2012). sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012*. Institut Jožef Stefan, Ljubljana, pp. 197-202.
- Toporišič, J. (2001). *Slovenski pravopis*. Založba ZRC, Ljubljana.
- Verovnik, T. (2004). Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX*. 241–258.

Exploring the Relationship between Language Change and Dictionary Compilation in the Age of the Collaborative Dictionary

Sharon Creese

Coventry University Priory Street, Coventry, CV1 5FB
creeses@uni.coventry.ac.uk

Abstract

The rise in collaborative ‘wiki’ dictionaries means that dictionary creation is no longer the purview solely of academics and publishing companies. Ordinary people can now create and share their own dictionary entries, whilst traditional publishing houses must compete against resources able to achieve levels of interactivity and immediacy that they simply cannot. These differences in the dictionary landscape may not be the only consequence of the rise of ‘wiki’ dictionaries, however; the very relationship between dictionary compilation and language change may be shifting, with the speed and ease of updating of ‘wiki’ dictionaries meaning that they not only reflect current use, but actually drive change.

This paper examines the possibility of this, through the findings of a pilot study featuring a new web-based corpus of youth neologisms, and media tracking of these new words. In it, I set out to determine the relationship between the *Wiktionary* definition and the grassroots use of particular words, as well as considering if and how this is changing as ‘wiki’ dictionaries become more and more firmly established.

Keywords: Wiktionary; wiki; collaborative dictionaries; language change; neologism; dictionary compilation, lexicography.

1. Introduction

Throughout history, the dictionary has always been a key tool in understanding how language should look and function. The rise of the Internet, however, and particularly the interactivity offered by Web 2.0, has fundamentally changed the dictionary landscape, with anyone now able to create and share their own ‘wiki’ contributions at the touch of a button (Meyer & Gurevych, 2012: 259; Leuf & Cunningham, 2005). The ease with which changes and additions can be made to these collaborative dictionaries means that they can be updated hundreds of times a day, offering a level of immediacy that cannot be achieved by mainstream electronic dictionaries. Though publishers may constantly add to and amend the entries in their dictionary wordlists, availability of this new information is governed by cost, meaning updates are often scheduled no more than four times a year.

The speed and ease of updating ‘wiki’ dictionaries opens up the opportunity for a more dynamic relationship between dictionary compilation and language change than has previously been the case, with the dictionary potentially not only reflecting language use, but actually driving change. Despite a growing body of literature on dictionary collaboration (see, for example, Meyer & Gurevych, 2012; Penta, 2011) this

possibility remains as yet unexplored. Evidence of such a shift in this relationship could prove valuable to dictionary publishers seeking ways to monetise and add value to their online offerings. Knowing that entry of a word into a ‘wiki’ dictionary leads to increased usage in the media could, for example, lead traditional publishers to consider working with the creators of ‘wiki’ dictionaries, in order to develop a stronger relationship with grassroots users of the language. This in turn might enable them to position themselves as more accessible than their competitors. Alternatively, a publisher that has featured new words which have gone on to be particularly active in the media might develop a marketing campaign around its success in recognising new words that stand the test of time.

This paper reports on a pilot study for a research project to examine the relationship between collaborative dictionary compilation and language change. It describes the design of a web-based corpus, WeBCoYN, to aid the identification of new words within teenage language. It subsequently discusses the pilot version of this corpus, and the process of media tracking potential neologisms in major newspapers and archives, to assess whether their use pre- or postdates appearance in the dictionary, and whether dictionary inclusion affects everyday patterns of use.

2. Wiktionary

Currently, the most influential ‘wiki’ dictionary is *Wiktionary*, launched in 2002. *Wiktionary* contributors come from all walks of life and educational backgrounds; they submit potential new entries to the dictionary by creating a new page featuring their word and its definition, which can be accepted as it is, edited and amended in the live file, or discussed in detail in the ‘Tea Room’ forum. These discussions can continue for weeks, and the entire conversation is available for others to review and join in, as is the ‘revision history’ showing changes made to the word’s *Wiktionary* page. New discussions can be started at any time if a problem with an entry is identified, or a change in definition is proposed. A historical profile of the word’s behaviour over time is therefore offered by the Tea Room combined with the ‘revision history’ page attached to each word. Revision histories comprise many lines of hyperlinks, every ‘save’ action having generated a new page in the history, accessed via a separate link. This can result in enormous amounts of loosely organised information, making it difficult to find evidence of a particular amendment.

Interestingly, although a long Tea Room discussion can provide some indication that a significant shift in meaning or usage has occurred, or that a new entry is controversial, some major changes seem to be accepted with little or no discussion, whereas minor issues can generate extensive threads. On the face of it, this often seems to depend on the individuals involved, some being more pedantic or prone to argument than others, some having a better command of English and some having more extensive knowledge of *Wiktionary* processes and content. It may also be that some people become overly concerned with the minutiae of an issue, or that

Wiktionary contributors' relative lack of linguistic knowledge and experience, as compared to that of professional lexicographers, deters them from entering into debate about complicated issues, and instead leads them to focus on less complex ones. This is an area which will be investigated in more detail during the main research project, for which this is the pilot study.

3. Materials and Methods

Online youth language was chosen for this project because young people conduct much of their lives in the electronic sphere, and may be responsible for taking neologisms coined to fill lexical gaps, for example in the technology marketplace (Janssen, 2013), and extending them into wider use. There is growing anecdotal evidence that young people play a major role in the spread and establishment of new words – see for example *Blorge*,¹ *The New York Times*,² and *Voxxi*.³ In time, WeBCoYN's corpus evidence may empirically demonstrate that this is the case.

The 16,000 word corpus used for this study is a pilot for WeBCoYN (the Web-Based Corpus of Youth Neologisms), a corpus of online 'youth' language. 'Youth' is defined in this case as those aged 12–25 (often also termed the 'teen' market, despite extending beyond the age of 19). WeBCoYN texts fall into two intersecting categories (Sinclair, 2004: 4):

- medium:
 - companion pages for 'teen' television programmes
 - online magazines/webzines
 - websites linked to trending franchise (e.g. the *Twilight* series)
 - independent 'teen' blogs
- type:
 - articles/features
 - biographies
 - personal comments (short entries referring to a previously mentioned topic)
 - blog posts (longer pieces on a new topic, possibly generating comments).

All texts are categorised according to the intersection between medium and type (see Figure 1). In the pilot study, each cell is approximately 1,000 words long. Contextual information was collected for each text, including date of collection, original

¹ Accessed at: <http://tech.blorge.com/Structure:%20/2009/08/31/teens-who-use-twitter-and-facebook-add-new-words-to-dictionary/>

² Accessed at: http://www.nytimes.com/2012/02/28/science/young-women-often-trendsetters-in-vocal-patterns.html?_r=2&

³ Accessed at: <http://www.voxxi.com/new-times-new-generations-new-words-genya-mujer/>.

publication date, and, where possible, the author’s age, gender, location, and education level.

Texts for the pilot corpus were collected using Google searches and manual reading of websites to identify suitable sections (for the full study, a web ‘crawler’ programme will be used to automate this process [Fletcher, 2013: 5]). Texts were then POS (part of speech) tagged using Wmatrix software (Rayson, 2008), and manually tagged for potential neologisms, that is, words that looked ‘new’.

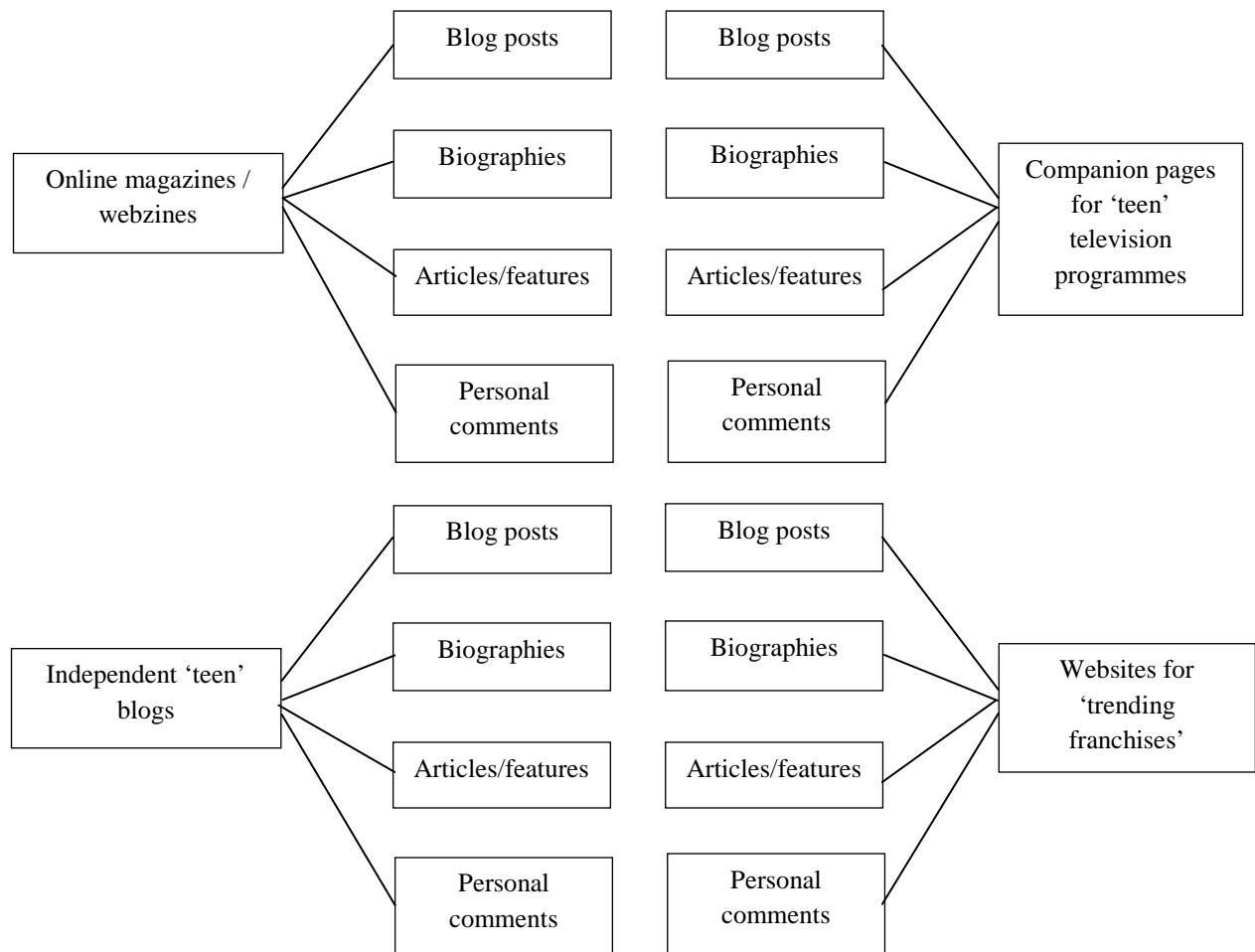


Figure 1. WeBCoYN cell structure (each cell contains an approximately equal number of words).

Once compiled and checked, the pilot corpus of 16,567 words was run through the Range programme,⁴ to exclude the most common 2,000 everyday words, and the top 1,000 academic words, as defined by the GSL (General Service List) and AWL (Academic Word List) (West, 1953; Coxhead, 2000). The resulting list of 2,452 words was then manually filtered, removing duplications, proper, place and trade names, obvious misspellings, and words which were clearly already well established. New

⁴ Accessed at: <http://www.victoria.ac.nz/lals/about/staff/paul-nation>.

word senses initially tagged as potential neologisms but which had been excluded during the Range filtering process on the basis of the original sense (for example ‘fetch’ and ‘genius’ [see Table 1]) were returned to the list, and the remaining 289 words were checked against eight dictionaries, to determine approximately when they entered the lexicon.⁵

Dictionaries from a number of different sectors were chosen here, in order to see whether new words appeared more quickly in standard reference works, in those aimed at second language learners, in non-British English dictionaries or in collaborative ‘wikis’:

- *Concise Oxford English Dictionary* (Soanes & Stevenson, 2006)
- *Macmillan English Dictionary for Advanced Learners of American English* (Rundell, 2002)
- *Macmillan English Dictionary for Advanced Learners* (Rundell, 2007)
- *Merriam-Webster* (2013) (online, accessed at: <http://www.merriam-webster.com/>)
- *Oxford Dictionaries* (2013) (online, accessed at: <http://oxforddictionaries.com/>)
- *Macmillan Dictionary* (British English/American English) (2013) (online, accessed at: <http://www.macmillandictionary.com/>)
- *Oxford English Dictionary* (OED) (2013) (online, accessed at: <http://www.oed.com/>)
- *Wiktionary* (2013) (accessed at: http://en.wiktionary.org/wiki/Wiktionary:Main_Page)

All words appearing in a dictionary before 2008 were deleted, since they can no longer be considered ‘new’, as were terms used only in an Internet context, unless they had already entered the dictionary and become established beyond their original sphere (e.g. ‘LOL, which entered *Wiktionary* in 2003, and appears in all of the dictionaries above). Google searches were then conducted to find evidence of use of the remaining 43 words. Evidence of significant usage, generating, for example, several pages of valid search results in multiple mass-media/social media/gaming contexts, or use in three or more ‘reputable’ sources (for example, websites produced by legitimate publishers), was deemed sufficient to consider the word ‘in use’. From the original 2,452 potential new words, 24 met these criteria, however only 14 had made it into a dictionary (in most cases, *Wiktionary*) and can therefore be considered established neologisms (see Table 1).

⁵ Unfortunately, online dictionaries produced by traditional publishers do not feature inclusion dates, and unlike *Wiktionary*, details of their lexicographical processes are not available to readers, meaning there is no way to know how, why or exactly when they were included.

Neologism (created since 2008)	Meaning	Part of Speech	Date entered Wiktionary	OED (online) 2013	OD (online) 2013**	MW (online) 2013**	MED (online) 2013**
fav	favourite	adjective	Feb-08	N	N	N	N
fetch	cool	adjective	Jan-13	N	N	N	N
genius	impressive	adjective	N	N	Y	N	N
girlchild	female child	noun	Feb-10	Y	N	N	N
gravatar	automatic avatar	noun	Mar-12	N	N	N	N
homeschooler	someone who is homeschooled	noun	Mar-09	Y	Y	Y	N
liveblogging	writing a real-time blog	verb	Dec-08	Y	Y	N	N
mischief- maker	creator of mischief	noun	N	Y	Y	N	Y
OMG	oh my god	exclamation	May-08	Y	Y	Y	Y
pre-visualize	imagine something before creating it	verb	N	Y	N	N	N
quick-release	single action release mechanism	adjective	N	Y	Y	N	N
sooo(o)*	emphatic version of 'so'	adverb	Aug-08	N	N	N	N
teared up	started to cry	verb	Jun-09	N	N	N	N
teenhood	period spent as a teenager	noun	Dec-09	Y	N	N	N

*This entry includes the variant 'soooo' found in the corpus, as indicated in brackets.

**OD = Oxford Dictionaries online; MW = Merriam-Webster online; MED = Macmillan Dictionary online

Table 1. Neologisms identified through analysis of WeBCoYN.

Twelve of these neologisms returned a frequency of one, with only 'sooo' (and the variant 'soooo') and 'OMG' appearing more than once (see Table 2). Given the size of the pilot corpus, it is perhaps unsurprising that the frequencies are so low. To get a wider view of the use of these words, they were also examined in Sketch Engine's SiBol/Port newspaper corpus.⁶

⁶ SiBol/Port draws data from three specific years: 1993, 2005 and 2010. Accessed at: https://the.sketchengine.co.uk/bonito/run.cgi/first_form?corpname=preloaded/sibolport_1;

Frequencies:	WeBCoYN	SiBol/Port
s000	3	65
s0000	2	63
s000(o) total	5	128
OMG	4	154

Table 2. Neologism frequency comparison – WeBCoYN and SiBol/Port.⁷

Five newspapers were chosen for media tracking, to cover the broad spectrum of target audiences (in terms of education level and socio-economic group) for this medium within the UK. In all cases, it was the online version of the newspaper that was consulted:

- *The Independent* (<http://www.independent.co.uk/>)
- *The Guardian* (<http://www.guardian.co.uk/>)
- *Daily Mail – Mail Online* (<http://www.dailymail.co.uk/home/index.html>)
- *The Sun* (<http://www.thesun.co.uk/sol/homepage/>)
- *Daily Express – Express* (<http://www.express.co.uk/>)

In addition to the main media tracking, digital newspaper archives were also interrogated, through the British Newspaper Archive.⁸

Newspapers were chosen to provide evidence of the use of new words/meanings in this pilot study because they are more able to keep pace with language change than books or magazines, since they are produced daily. They are also aimed at a wide cross section of the population – different ages, education levels, income brackets and social groups – meaning that new or amended words that appear in newspapers can be deemed to have moved beyond their original sphere, and become established within the language.

In all five newspapers, a search was conducted for the neologism, using the paper's online search engine.⁹ A number of problems were encountered, for example concerning the lack of consistency in how results are presented. *The Guardian* presents a list of the number of articles featuring the search word, broken down by

⁷ Size of corpora – WeBCoYN: 16,567 tokens, 3785 words; SiBol/Port: 387,585,716 tokens, 327,025,669 words.

⁸ See <http://www.bl.uk/>.

⁹ Since conducting the pilot study in April 2013, *The Independent* and *The Guardian* have changed their search functions. At the time of the original media tracking, the latter only searched mainstream articles; it now also includes data from interactive pages like blogs and comments. *The Independent* no longer limits initial search results to post-2010, and some articles included in the April 2013 results list are now excluded (presumably because the articles have been removed). *The Sun*, meanwhile, has rebranded its online presence to *Sun+* and now no longer allows for searching without creating a subscriber account. All of this means that conducting the same study in August 2013 could lead to different results than those reported here.

year, which the user can then click through to read.¹⁰ The other newspapers do not provide numerical results lists; *The Independent* simply says that the word has appeared ‘x times since 2010’, and provides links to the relevant articles, whilst *Mail Online*, *The Sun* and the *Express* merely give the total number of results, plus links, with no indication of the time frame. This problem was largely overcome by conducting manual year-on-year searches in *The Independent*, *Mail Online* and *The Sun* (using the ‘advanced search’ function), in order to obtain results comparable with those from *The Guardian*. The lack of an advanced search facility in the *Express*, however, meant it was not possible to do the same, and hence only flat figures were available, with no date context.

A further difficulty was that where the term under investigation is a new sense of an existing word (for example ‘fetch’ or ‘genius’ above), the number of search results is unmanageably high, since the search functions offered by these newspapers have no POS filter, and hence every instance of the word appears. Each article must then be individually examined, to determine if it contains the correct sense of the word. For the pilot study, this problem was resolved by selecting newly created words for media tracking, rather than new senses of existing words. This returned few enough results that each article could be individually checked, using corpus query software to generate concordance lines of all of the instances of the word, which could then be analysed for sense and meaning to ensure they were, indeed, the word under investigation.

For the main WeBCoYN study, a three-stage process will be employed to create a searchable corpus of newspaper articles containing the neologisms being media tracked. Firstly, files identified by the newspaper’s search engine as containing the relevant neologism will be automatically downloaded, to create a corpus of HTML files. A script will then be used to remove all HTML tags and output the files as pure text. Finally, these text files will be run through a POS tagger such as Wmatrix to add the structural mark-up required to enable identification of the correct use of the word, for example, the adjective form of ‘genius’ as opposed to the noun.

The two words chosen for media tracking in the pilot study (from the list in Table 1) were ‘gravatar’ (an automatic avatar) and ‘teenhood’ (the period of being a teenager).

Although both these words are in *Wiktionary*, neither of them has a discussion page in the Tea Room. This suggests that no-one has objected to the original definitions, and there has been no further development of the words. The use of these two words by teenagers, and their appearance in *Wiktionary* but their absence from most traditional dictionaries (online or print), makes them ideal candidates for

¹⁰ Following the changes to its search function, *The Guardian*’s results are now also presented differently, appearing as a chronological list, instead of year by year. This is less user-friendly than the previous format, and could hamper media tracking for the main WeBCoYN study.

examination of the impact of collaborative dictionaries on language change. They are new words which are still in the process of establishing themselves in the lexicon. By examining the frequency, date and context of their use in the media, we can consider the possible impact of entry into a ‘wiki’ dictionary on everyday use of a word.

4. Findings

Gravatar

‘Gravatar’ entered *Wiktionary* in March 2012. A blend of ‘globally recognised avatar’, it began life as a trade name, but is rapidly becoming a generic term. ‘Gravatar’ refers to an avatar linked to an email address via a central registration point; wherever that email address is used to post a comment on a website, the ‘gravatar’ is automatically imported.¹¹ Plugins are now available to allow ‘gravatars’ to be incorporated into independent sites.

‘Gravatar’ is a new word at the beginning of its lexical journey. So far, it has only entered *Wiktionary* (2013) (and *Wikipedia* [2013]); it does not yet appear in any of the other collaborative dictionaries, such as *The Free Dictionary* (2013) or the *Urban Dictionary* (2013) (although a film of the same name is included in the latter [2010]). ‘Gravatar’ is beginning to be used as an alternative to ‘avatar’, and it is possible that this may become more common as its use spreads from social networking and blogging sites, to more mainstream ones. Similarly, as users of the term grow older, they will likely carry the word with them, so we could reasonably expect to see ‘gravatars’, rather than ‘avatars’ on the comment pages of newspapers or other news outlets in the future.

Media tracking ‘gravatar’ in the five target newspapers returned no results, which is unsurprising given how new the word is and the fact that at present it remains firmly within the online sphere of use. (The British Newspaper Archive returned one result, but it was the name of a school, featured in an advertisement in 1883.) A Google search for ‘gravatar’, returned 170 million hits,¹² the first few pages being mostly blogging sites, Internet forums and compatible software.

1.	enabled sites such as this one. Using	gravatars	helps make our weblog a more friendly and personal
<i>Source: http://www.synchronoustechology.net/blog/how-to/set-up-your-gravatar/.</i>			
2.	activate the plugin, and it will add	gravatars	to your blog template and admin panel automatically
<i>Source: http://wordpress.org/extend/plugins/gravatar-favicon/.</i>			
3.	anyone know how I get my	gravatar	on my battlelog I've got it set up but just don't know
<i>Source: http://battlelog.medalofhonor.com/mohw/forum/threadview/2832654490161464530/.</i>			
4.	here Just wanted to say about your	gravatar.	We both have sketched birds out there. => Best
<i>Source: WeBCoYN pilot study, April 2013.</i>			

Table 3. Online concordances of ‘gravatar’.

¹¹ See <https://en.gravatar.com>.

¹² As at 18.4.13; by 18.8.13 this had risen to 211 million.

Table 3 shows concordance lines for ‘gravatar’ taken from these online sources, and the WeBCoYN pilot study. Lines 1 and 2 demonstrate the use of ‘gravatar’ in blogging contexts, giving guidance on how ‘gravatars’ can enhance websites. Both assume a level of understanding of what a ‘gravatar’ is and its purpose.

Concordance line 3 is taken from a gaming forum, and features a player who is seeking help with his ‘gravatar’. Many of the instances of ‘gravatar’ online are in forums or ‘help’ pages like this, offering advice on how to get the best out of a ‘gravatar’. Concordance line 4 (taken from the WeBCoYN pilot study) is slightly different, in that the reader of a blog is commenting on the ‘gravatar’ used by the blog’s author. Again, it is clear that the writer of the comment understands and is familiar with ‘gravatars’, and it is implied that the blog author is in a similar position. The fact that the comment is being made, suggests that the writer has seen and recognised the ‘gravatar’ from elsewhere on the web. This ‘transferability’ is, according to the company behind them, one of the key functions of ‘gravatars’.¹³

Teenhood

Although it also occurred only once in the pilot WeBCoYN study, ‘teenhood’ is a more established word than ‘gravatar’, having entered *Wiktionary* in December 2009 and already featuring in *OED* and the *Urban Dictionary*. It also appears in *UKWaC*, with concordances dating back to 2003.¹⁴ Despite this, media tracking of ‘teenhood’ returned only 20 results (from 2000–2012), the majority of which were confined to *The Guardian* (see Table 4), suggesting that it is still not particularly well established in the lexicon.

‘Teenhood’ also appears seven times in the British Newspaper Archive, with all instances carrying the same meaning. All of these come from the late 1800s, however (see Table 5). This suggests that while we may think of ‘teenhood’ as new, it is actually a word which enjoyed a brief period of use over a century ago, fell out of favour and was then reinstated, or was perhaps even created anew without awareness of its earlier existence. Unlike other reinstated words, such as ‘truthiness’, there is no indication in *Wiktionary* of this previous incarnation of ‘teenhood’.¹⁵

¹³ See <https://en.gravatar.com/>.

¹⁴ See <https://the.sketchengine.co.uk/bonito/run.cgi/first?iquery=teenhood&queryselector=iqueryrow&corpname=preloaded%2Fukwac2>.

¹⁵ See <http://en.wiktionary.org/wiki/truthiness>.

Year	The Guardian	Month	The Independent	Month	Mail Online	Month	The Sun	Express
2013	0		0		0		0	
2012	1	Dec	1	Nov	0		0	
2011	4	Mar, May*, Jun, Nov	0		0		0	
2010	1	Aug	0		1	Apr	0	
2009	1	Sep	2	Aug, Dec	0		0	
2008	2	Oct, Nov	0		1	Feb	0	
2007	2	Jun, Jul*	0		0		0	
2006	0		0		0		0	
2005	1	Aug	0				0	
2004	1	Jun*			0		0	
2003	0				0		0	
2002	1	Jul			0		0	
2001	0				0		0	
2000	1	Aug			0		0	
Total	15		3		2		0	0

*alternate spelling used: ‘teen-hood’

Table 4. Appearances of ‘teenhood’ in target media.¹⁶

Word	British Newspaper Archive	Date
teenhood	7	1870s
	1	1890s

Table 5. Appearances of ‘teenhood’ in digital archives.

Whilst ‘teenhood’ may not be strictly speaking ‘new’, the number of appearances found during media tracking was lower than expected, and was unexpectedly biased towards a single newspaper, *The Guardian*. Comparing the dates of these instances with the date of entry into *Wiktionary* – December 2009 – shows a marked increase in usage after inclusion. ‘Teenhood’ appeared in the five target newspapers 11 times between January 2000 and November 2009, and nine times from December 2009 to December 2012. Thus we see a doubling of the frequency of appearances, from an average of 1.2 per year pre-*Wiktionary*, to 3, post-*Wiktionary*. It is also interesting to note that, outside of *The Guardian*, only two appearances of ‘teenhood’ occurred before the word entered *Wiktionary*: one in *The Independent* just four months beforehand, and the other in *Mail Online* in February 2008. After its entry into

¹⁶ Some newspaper archives are more comprehensive than others, leading to gaps in the search results prior to 2005. Blank cells indicate that it was not possible to search that period. *The Express* search engine does not facilitate year-by-year searching.

Wiktionary, *The Independent* used ‘teenhood’ twice more, the *Mail Online*, only once.¹⁷

Examining sample concordance lines for these uses of ‘teenhood’ (see Table 6), we can see that there has been no change in the use or meaning of the word over this time. From 2000 through until 2010, ‘teenhood’ is used in the context of ‘adolescence’, with a sense of nostalgia for an earlier time in life. All of the concordances either refer to or imply powerful relationships, along with the sense of a journey, sometimes physical (concordance lines 1 and 2) sometimes emotional (3 and 4).

1.	about a car-crazy, rock ‘n’ roll midwestern	teenhood	in the late 50s and early 60s, Lucas made a movie
Source: <i>The Guardian</i> , August 2000			
2.	on a roadtrip to find their boyfriends from	teenhood?	It doesn’t mean love is dead: it merely means
Source: <i>The Guardian</i> , August 2005			
3.	early films captured the exquisite pains of	teenhood	growing in popularity to achieve cult status.
Source: <i>The Independent</i> , August 2009			
4.	and it’s not like you spend childhood and	teenhood	preparing for adulthood and then everything is
Source: <i>The Guardian</i> , December 2012			

Table 6. Concordance lines for ‘teenhood’ – media tracking April 2013.

The lack of any discussion over ‘teenhood’ in the Tea Room, and the mere five entries in its revision history (almost all of which occurred over a ten minute period) all indicate that the *Wiktionary* populace is happy with its definition of ‘teenhood’:

- ‘1. adolescence
- 2. state of being a teenager’ (2009).

The media’s corresponding use of the word, and its growing popularity, suggests that non-*Wiktionary* users are similarly satisfied with it.

5. Conclusion

The speed and ease of updating ‘wiki’ dictionaries opens up the opportunity for a more dynamic relationship between dictionary compilation and language change, with the dictionary potentially not only reflecting language use, but actually driving change. Whilst several authors are already working on the implications of ‘wiki’ dictionaries (see for example Meyer & Gurevych, 2012; Gurevych & Wolf, 2010; Penta, 2011), following on from earlier works on the wider field of electronic, and collaborative but non-interactive dictionaries (see Nesi, 2008 and de Schryver, 2003), this relationship has, as yet, gone unexplored.

¹⁷ Although one of *The Independent’s* post-*Wiktionary* uses was in a round-up obituary article featuring quotes from the earlier August 2009 piece. Both have since been removed from the site.

Of course not every word that enters *Wiktionary* will stand the test of time (Algeo, 1993). Whilst ‘gravatar’ is at too early a stage in its linguistic development to predict its future with any certainty, it appears that ‘teenhood’ is surviving and may, in fact, thrive. Media tracking in five newspapers identified that ‘teenhood’ was used only 11 times in the media prior to inclusion in *Wiktionary*, whilst it appeared nine times in the three years afterwards. It will be interesting to see over the coming months and years whether this increase is sustained and will lead to a successful reincarnation for ‘teenhood’, and whether this ultimately leads to recognition by traditional publishers of works other than *OED*, and incorporation into new editions of other mainstream dictionaries.

If it does, it may be that the project following this pilot study will reveal a similar pattern of entry into *Wiktionary*, followed by an increase in use and faster establishment of a place in the lexicon. This could suggest a new role for *Wiktionary* as an early predictor of successful neologisms. Determining this will require analysis of both successful and unsuccessful new additions to *Wiktionary* (defined by the longevity of the word).

Evidence of a new relationship between dictionary-making and language change would not only satisfy academic curiosity, but could prove useful to dictionary publishers seeking innovative ways to monetise their online offerings and set themselves apart from the competition. A clearer understanding of the behaviour of new words once they have entered the dictionary and begun to spread into wider spheres of use could enable these companies to better tailor their time and resources, whilst building a stronger relationship with grassroots language users. Collaboration with the producers of ‘wiki’ dictionaries could present traditional publishers with a unique selling point around which to promote their products.

6. Acknowledgements

My thanks to my PhD supervisor Professor Hilary Nesi (Coventry University) for her help and guidance on this paper and on the corpus pilot study. Thanks also to Sian Alsop (Coventry University) for her assistance in devising the automated media tracking process proposed for the next stage of this research.

7. References

- Algeo, J. (1993). Desuetude among New English Words. *International Journal of Lexicography* 6(4), pp. 281-293.
- Coxhead, A. (2000). An Academic Word List, *ELI Occasional Publications #18*, Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Daily Express – Express* [online]. Accessed at: <http://www.express.co.uk/>.
- Daily Mail – Mail Online*. Accessed at: <http://www.dailymail.co.uk/home/index.html>.

- de Schryver, G-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2), 143-199.
- Fletcher, W.H. (2013). Corpus Analysis of the World Wide Web. In Chapelle, C.A. (ed.) *The Encyclopedia of Applied Linguistics* [online] Blackwell Publishing. Accessed at: <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0254/full>.
- Gravatar* [online]. Accessed at: <https://en.gravatar.com/>.
- Gurevych, I. & Wolf, E. (2010). Expert-Built and Collaboratively Constructed Lexical Semantic Resources. *Language and Linguistics Compass*, 4(11), pp. 1075-1090.
- Janssen, M. (2013). Lexical Gaps. In Chapelle, C.A. (ed.) *The Encyclopedia of Applied Linguistics*. [online] Blackwell Publishing. Accessed at: <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0693/full>.
- Leuf, B. & Cunningham, W. (2005). *The Wiki Way: Quick Collaboration on the Web*, USA: Addison-Wesley.
- Macmillan Dictionary (British English / American English)* [online] Accessed at: <http://www.macmillandictionary.com/>.
- Merriam-Webster* [online]. Accessed at: <http://www.merriam-webster.com/>.
- Meyer, C. & Gurevych, I. (2012). Wiktionary: a New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography. In Granger, S. & Paquot, M. (eds.) *Electronic Lexicography*. Oxford: Oxford University Press, pp. 259-292.
- Nesi, H. (2008). Dictionaries in Electronic Form. In Cowie, A.P. (ed.) *The Oxford History of English Lexicography*. Oxford University Press, pp. 458-478.
- Oxford Dictionaries* [online]. Accessed at: <http://oxforddictionaries.com/>.
- Oxford English Dictionary* [online]. Accessed at: <http://www.oed.com/>.
- Penta, D.J. (2011). *The Wiki-fication of the Dictionary: Defining Lexicography in the Digital Age*. Media in Transition 7 Conference, Massachusetts Institute of Technology, Cambridge, MA, USA, 13 May 2011.
- Rayson, P. (2008). From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics* 13(4), pp. 519-549.
- Rundell, M. (2002). *Macmillan English Dictionary for Advanced Learners of American English*. Oxford: Macmillan Education.
- Rundell, M. (2007). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan Education.
- Sinclair, J. (2004). Corpus and Text – Basic Principles. In Wynne, M. (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. [online]. Oxford: Oxbow Books. Accessed at: www.ahds.ac.uk/linguistic-corpora/.
- Sketch Engine: SiBol/Port Corpus* [online]. Accessed at: https://the.sketchengine.co.uk/bonito/run.cgi/first_form?corpname=preloaded/sibolport_1;
- Sketch Engine: SiBol/Port Corpus – ‘teenhood’* [online]. Accessed at: https://the.sketchengine.co.uk/bonito/run.cgi/first?corpname=preloaded%2Fsibolport_1&reload=&iquery=teenhood&queryselector=iqueryrow&lemma=&lpo

s=&phrase=&word=&wpos=&char=&cql=&default_attr=word&fc_lemword_wi
ndow_type=both&fc_lemword_wsize=5&fc_lemword=&fc_lemword_type=all
&fc_pos_window_type=both&fc_pos_wsize=5&fc_pos_type=all&usesubcorp=
&sca_doc.author=&sca_doc.date=.

Sketch Engine: ukWaC (old WSG) – ‘teenhood’ [online]. Accessed at:
<https://the.sketchengine.co.uk/bonito/run.cgi/first?iquery=teenhood&queryselector=iqueryrow&corpname=preloaded%2Fukwac2>.

Soanes, C. & Stevenson, A. (2006). *Concise Oxford English Dictionary*. Oxford:
Oxford University Press.

The British Newspaper Archive [online]. Accessed at:
<http://www.britishnewspaperarchive.co.uk/>.

The Free Dictionary [online]. Accessed at: <http://www.thefreedictionary.com/>.

The Guardian [online]. Accessed at: <http://www.guardian.co.uk/>.

The Independent [online]. Accessed at: <http://www.independent.co.uk/>.

The Sun [online]. Accessed at: <http://www.thesun.co.uk/sol/homepage/>.

Urban Dictionary (2013). [online]. Accessed at: <http://www.urbandictionary.com/>.

Urban Dictionary – ‘Gravatar’ [online]. Accessed at:
<http://www.urbandictionary.com/define.php?term=gravatar>.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green
& Co.

Wikipedia – ‘gravatar’. [online]. Accessed at: <http://en.wikipedia.org/wiki/Gravatar>.

Wiktionary (2013). [online]. Accessed at:
http://en.wiktionary.org/wiki/Wiktionary:Main_Page.

Wiktionary – ‘teenhood’ [online]. Accessed at:
<http://en.wiktionary.org/wiki/teenhood>.

Wiktionary – ‘teenhood’ revision history [online]. Accessed at:
<http://en.wiktionary.org/w/index.php?title=teenhood&action=history>.

Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic MediaWiki

Bruno Bon¹, Krzysztof Nowak²

¹Institut de recherche et d'histoire des textes (CNRS), Paris, France

²Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland

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

Abstract

This paper reports the results of a survey of which the main aim was to scrutinize consequences of adopting a wiki model in alignment of Medieval Latin dictionaries. In the first section, the objectives as well as the methodology of the project are presented. As a framework, we used Semantic MediaWiki (SMW), and for the purpose of the research several entries from four dictionaries were selected. In the following sections we scrutinize the presentation, search, and collaboration features provided by SMW. We demonstrate how intrinsic wiki concepts, such as namespaces, templates, property-value pairs etc., may be employed in macro- and microstructure display. Next, alternative modes of accessing lexicographical data are presented such as maps, timelines, charts etc. After that search capabilities are analyzed, among which the most important appear to be semantic properties search and faceted browsing. Lastly, the paper considers on different ways which SMW can encourage researchers to collaborate and enrich dictionary content.

Keywords: Medieval Latin; wiki-interface; multilingual dictionaries linking; dictionary alignment

1. Introduction

In 1913 the idea of a Pan-European dictionary of Medieval Latin was clearly expressed by the research community, but not until the early 1920s did work begin on preparing *Novum Glossarium Mediae Latinitatis* which covered four centuries (IX–XII) of Latin language use (Langlois, 1924), replacing the older and already obsolete *Glossarium* of Charles du Fresne, sieur du Cange (Du Cange, 1883). From the very beginning it was also clear that, due to various periodization of the Middle Ages, the compilation of national dictionaries was necessary. This is the reason why there now exist a dozen dictionaries which vary not only in their chronological (500–1600 AD) and regional (from Spain to Poland; from Sweden to Italy) coverage, but also in their advancement (three were completed, but the majority of projects are works in progress).

Yet, with the advent of e-lexicography the founding idea of the common dictionary of European Latin should again be considered. A first step was made during the congress of Medieval Latin lexicography in Barcelona in 2004, where several elements of microstructure were proposed as a basis for dictionary alignment, among them headword, etymology and sense definitions (Heid, 2004). This proposal,

however, was put forward without major consideration of such “technical” issues as software framework, encoding schema or data structure. Over subsequent years the community witnessed the emergence of several e-lexicography and e-corpora projects, among which one should mention:

(1) electronic editions of Du Cange’s *Glossarium*¹, *Novum Glossarium Mediae Latinitatis*², dictionaries of medieval Latin from Polish³ and Catalan⁴ sources;

(2) corpora of medieval Latin in Catalonia⁵, Galicia⁶, and Poland⁷.

This rapid development, in turn, has raised an interest in encoding standards and lexicographical data interoperability. At the same time, several institutional enterprises have been launched in order to foster research collaboration and sustain data exchange, one of them being COST Action 1005 “Medioevo Europeo”⁸. Its goal, as the project’s description says, is the development of a so-called “Virtual Centre of Medieval Studies”, a common interface for querying until now dispersed databases, text collections, library catalogues, etc. After being appointed as experts of the project on behalf of Medieval Latin dictionary teams, we advanced the idea of a wiki-based tool. In the present paper, we discuss a working prototype of such a wiki, an interface and research environment which could potentially serve as a unified edition of Medieval Latin dictionaries and lexical databases.

2. Objectives and procedures

As a framework for our survey we choose MediaWiki (MW)⁹ which is best known as an application running in the background of Wikipedia. Once installed, it was subsequently supplemented with a bunch of plugins of which the essential one was Semantic MediaWiki (SMW)¹⁰, an extension that enhances MediaWiki with semantic dimension, enabling advanced data annotation and as a consequence finer data retrieval. MW enables an explicit declaration of the exact meaning of the data

¹ Accomplished, available at <http://ducange.enc.sorbonne.fr/>.

² In progress, due to be finished in 2013, more information on <http://glossaria.eu>.

³ *eLexicon Mediae et Infimae Latinitatis Polonorum*, in progress, due to be finished in mid-2014, <http://scriptores.pl>.

⁴ In progress, more information at <http://gmlc.imf.csic.es/>.

⁵ *CODOLCAT. Corpus Documentale Latinum Cataloniae*, <http://gmlc.imf.csic.es/codolcat/index.php>.

⁶ *CODOLGA. Corpus Documentale Latinum Gallaeciae*, <http://www.cirp.es/codolga/>.

⁷ *Fontes Mediae Latinitatis Polonorum*, in progress, due to be finished in 2016, more information at <http://scriptores.pl>.

⁸ <http://www.medioevoeuropeo.org/>

⁹ <http://www.mediawiki.org>

¹⁰ <http://semantic-mediawiki.org/>

contributed to the wiki page by annotating it with the property name:

[[Property_name::Property_value]]
eg. [[Headword::Mandragora]].

Software choice was driven by the project goals and objectives, which can be summarized as follows:

1. Software should already exist and be free. There was no funding envisaged in the project for writing software from scratch, since it is treated as a means of fostering discussion rather than as a goal of the project.
2. Software should be open-source. The lexicographical and corpus data in the emerging projects, in the majority of cases, are (or soon will be) available under liberal licensing models, as should be therefore the tools used in their retrieval. Since the goal of the project is to foster collaboration and data exchange, participating projects cannot be excluded or limited by the use of binary file formats or infrastructures closed to further refinement.
3. Stable development, and community support. In order to ensure project longevity, the tool should be actively developed and supported by a stable number of code contributors.
4. Compliant with dictionary-type data.
5. Multilingual interface.
6. Collaboration-oriented.
7. Easy to use.

MW and SMW are not only free and fully open-source, but they have also been created with encyclopedia-like data in mind to provide an internationalized interface. Thanks to its popularity, MW may also encourage less advanced users to actively collaborate.

SMW, although steadily gaining in popularity, has not yet been employed in vast lexicographic projects. According to the list of sites using SMW¹¹, extension has been implemented in such projects as *Liddell-Scott-Jones Ancient Greek Lexicon Edition*¹², *An interactive online etymological dictionary of Lepontic*¹³, or *Neuroscience Lexicon*¹⁴. None of them were known to the authors in early 2012 when

¹¹ <http://smw.referata.com/wiki/Special:BrowseData/Sites>.

¹² <http://lsj.translatum.gr/>.

¹³ <http://www.univie.ac.at/lexlep>.

¹⁴ <http://neurolex.org/>

works on integrated query interface were launched.

For the purpose of the present paper, 4–6 entries from four dictionaries were chosen and subsequently encoded by typing wiki syntax code. Whenever possible, lexicographical content was passed to the formerly created templates¹⁵, which automatize not only text formatting, but also semantic annotation of data. For instance, when a content author types `{{headword|mandragora}}`, a template “headword” is called upon with the first argument set to “mandragora”. Once triggered, the template:

(1) sets property “headword” to “mandragora” and displays text string “mandragora” on entry page¹⁶;

(2) sets property “headword_canonical” to “mandragora” without displaying the word itself on the entry page¹⁷.

The annotation task was primarily conducted by the authors of this paper with the help of Renaud Alexandre (IRHT CNRS). Subsequently, several members of other lexicographical teams have become familiar with the wiki editing interface (especially wiki syntax) and have been asked to correct or edit entries from scratch.¹⁸

3. Macrostructure

the main goal of the present database was to enable a unified retrieval of dispersed dictionaries, the provenance of lexicographical data is that they should always be easily traceable. Firstly, this enables an acknowledgement of the institutions and research teams which have developed the machine-readable dictionaries. Secondly, it offers users the possibility of limiting their search results. In our prototype, separation of dictionary entries has been assured by resorting to the mechanism of wiki namespaces¹⁹, each entry being preceded by a 2-letter prefix indicating the dictionary from which it originates: namespace:entry_headword, e.g. for Latin word *decipula* ‘a snare, trap’, a full page title is *PL:Decipula* which results in the following entry link: `.../index.php?title=PL:Decipula`. This separation allows users to browse each dictionary in a traditional way by referring to the entry list (`.../index.php?title=PL`).

¹⁵ <http://www.mediawiki.org/wiki/Help:Templates>.

¹⁶ The code in template is `[[Headword::{{{1}}}|{{{1}}}]]`, where number stands for argument order number.

¹⁷ The code being `{{#set:Headword_canonical={{#regex:{{{1}}}/\w+\s*+/.}}}`. Canonical form of entry headword is computed by applying to a full headword a simple regular expression which gets rid of symbols, numbers etc.

¹⁸ Their names can be found in Acknowledgment section of the present paper.

¹⁹ <http://www.mediawiki.org/wiki/Namespaces>.

Main namespace has been reserved for so-called “super-entries”, i.e., entries of the unified dictionary which serve as an index for all headwords. A super-entry page for the headword *depost*, for example, will provide a list of the dictionaries in which the word is attested (with appropriate links), as well as other information about the word in question that can be retrieved by means of the embedded queries. This information is presented in the form of timelines and maps of the attested word occurrences which have been extracted from respective dictionary entries:

Figure 1: Super-entry page

Spatio-temporal information retrieval, as at the heart of the WikiLexicographica, was possible, because each source quotation is stored as a so-called “semantic internal object” (SIO)²⁰, a complex data structure which permits encapsulation of multiple property-value pairs. SIOs include in particular:

- (1) reference to the entry to which they belong;
- (2) source reference abbreviation;
- (3) bibliographical data (page, verse etc.);
- (4) proper citation;
- (5) date of text composition;
- (6) geographical provenance of the text.

²⁰ http://www.mediawiki.org/wiki/Extension:Semantic_Internal_Objects.

```

{{#set_internal:CitInt
|Cit_siglum={{1}}
|Cit_ref={{2}}
|Cit_datum={{3}}
|Cit_citatio={{4}}
|Cit_start={{#if:{{5}}|{{5}}|{{#show:{{NAMESPACE}}:{{1}}|?Fons start}} }}
|Cit_end={{#if:{{6}}|{{6}}|{{#show:{{NAMESPACE}}:{{1}}|?Fons end}} }}
|Cit_genus={{#show:{{NAMESPACE}}:{{1}}|?Fons genus}}
|IsPartOfLemme={{NAMESPACE}}:{{PAGENAME}}
|Cit_geo={{#show:{{NAMESPACE}}:{{1}}|?Fons geo}}
|Cit_coordinates={{#geocode:{{#show:{{NAMESPACE}}:{{1}}|?Fons geo}} }}
}}

```

Figure 2: SIO structure

Since in Medieval Latin dictionaries neither chronological (5) nor geographical (6) data are explicitly declared for each quotation, values of these properties are usually computed from information provided in source description pages which form an essential part of integrated dictionary macrostructure.

Source pages belong to the same namespaces as the entries themselves. They are distinguished from them by category attribution: whereas entries belong to the category *Voces* (*lat.* ‘words’), source description pages are marked as *Fontes* (*lat.* ‘sources’). Source description pages consist of manually typed or database-extracted metadata which are subsequently passed to a bunch of embedded queries. So, for instance, the wiki syntax:

```

{{fons|EU|France, Lille||1150|1200|Alan. Ins. elucid.|Elucidatio in Cantica canticorum. – PL 210 col. 51-110|commentarius}}

```

results in a page (Figure 3) which shows a source provenance map, a bibliographical record and so on. The most interesting item, however, is the section *Citationes* (*lat.* ‘quotations’) where the user can find a list of headwords in which the source in question is referenced with its appropriate quotations. As long as we do not have at our disposal a complete, research-driven corpus of Medieval Latin, dictionaries can only be considered provisional corpora including a selection of Medieval Latin literature in seemingly good editions.

Sources and their quotations can subsequently be browsed, traditionally in the form of alphabetically-ordered lists. However, the user can also:

- (1) sort by frequency in dictionaries;
- (2) browse them on a timeline;
- (3) access them on a google map.

Informatio

Lexica in quibus invenitur:

EU

Ubi situs est?

France, Lille



Quando?

Quando incipit?

1150

Quando finitur?

1200

Quo siglo praedatur

Alan. Ins. elucid.

Genus

commentarius

Descriptio

Elucidatio In Cantica canticorum. - PL 210 col. 51-110

Citationes

Lemma	Reference	Texte
Mandragera	col. 103 ^a	per -as, herbam scilicet medicinalem et odoriferam, nisi perfe

Figure 3: Source description page

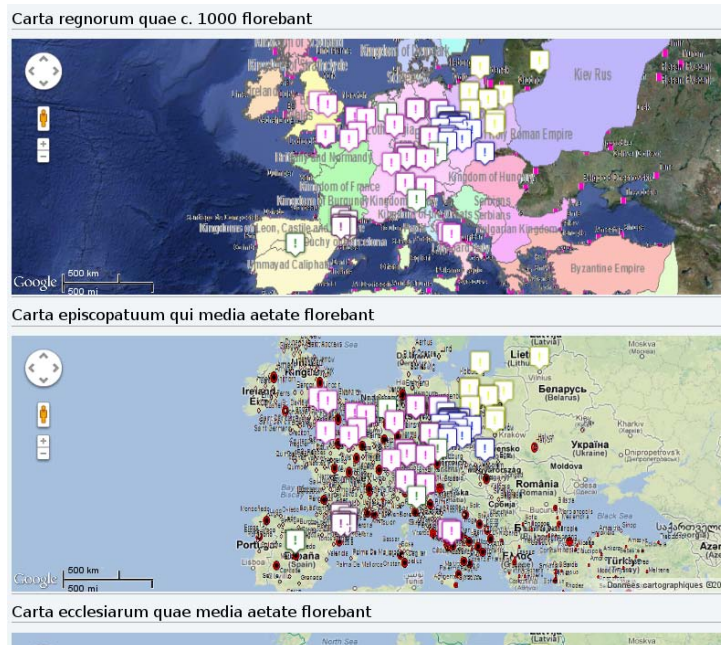


Figure 4: Map layers

In our survey, the last form of lexicographical data analysis has been enriched (Figure 4) thanks to the map layers provided by the project *Digital Atlas of Roman*

*and Medieval Civilization*²¹. One is now available for viewing source citations in the context of administrative boundaries of the medieval world and in the light of regional variation of medieval intellectual culture.²²

4. Microstructure

Medieval Latin dictionaries have as their primary public the research community; a fact which too often means that their entry structure is far from being user friendly. In our wiki we attempt to address this problem by providing two parallel access points to the dictionary microstructure. The first perspective presented to the user visiting the entry page, is a basic one (Figure 5). It comprises essential lexicographical information, such as graphic forms, inflection type, gender, and abbreviated sense definitions. The basic view tab, though, is also a place where the user is offered an overall picture of word occurrences. Entry source citations here are conveniently epitomized in text type chart, timeline, and map. Therefore, a quick glance should suffice to estimate in what Medieval genres, when and where the word in question would be cited most frequently.

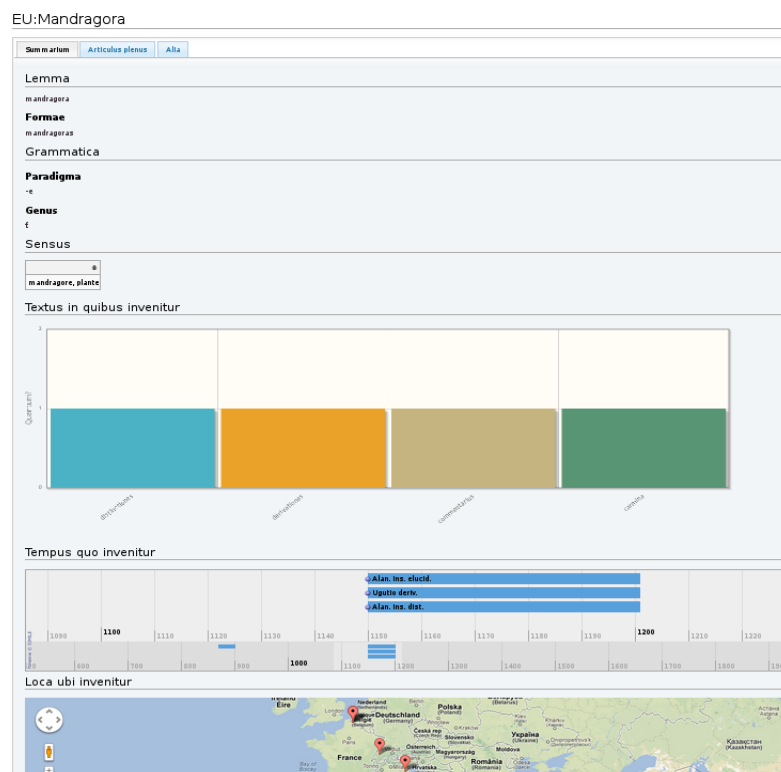


Figure 5: Dictionary entry (basic view)

²¹ <http://darmc.harvard.edu/icb/icb.do>.

²² This was possible due to the use of such layers as „medieval kingdoms”, „universities C12-C15” etc.

In the subsequent tab, the user can consult full entry with all idiosyncrasies that each dictionary editorial system is affected with. The relative variety of typographic conventions, as well as different levels of data explicitness, renders the preservation of original entry display for each dictionary a very difficult, if not impossible, task. This is one of the reasons why, in our opinion, the wiki interface should not be considered a means of text-oriented digitization of lexicographical work. The other major problem stems from the relatively flat textual data representation in Semantic MediaWiki: it is not easy (if at all possible) to properly reflect the nested tree structure of the dictionary entry by means of wiki syntax only. It can seem a serious limitation, considering that Medieval Latin lexicographers tend to make a heavy use of sense nesting in order to account for semantic change. Thus, a more appropriate approach seems to be data-oriented recompilation of source data into the desired output format, even if some of the original data (in particular, formatting) are lost. The burden of preserving original work in its typographic, sequential, etc., order may, then, be shifted towards each separate project rather than the integrated query interface.

Huius wiki alia vocabularia

[CZ:Mandragora](#) [DE:Mandragora](#) [PL:Mandragora](#)

Alia vocabularia

Latinitatis antiquae aetate florentis

- [Lewis-Short](#)
- [Georges \(ed. 1913\)](#)
- [Gaffiot](#)

Latinitatis media aetate florentis

- [Vetus DuCange](#)
- [Lexicon musicum Latinum medii aevi](#)

Latinitatis aetatis recentis

- [Ramminger](#)
- [Camena](#)

Corpora

- [Wikisource](#)
- [Perseus](#)
- [Patrologia Latina Database \(payant!\)](#)
- [Brepols Databases \(payant!\)](#)
- [CODOLCAT](#)
- [monasterium.net](#)
- [CroALa](#)
- [Chartae Burgundiae Medii Aevi](#)

Ad rem

[Mandragora at Wikipedia](#)

[The International Plant Names Index](#)

Figure 6: Dictionary entry ('Other resources' tab)

The next tab of each entry comprises links to other linguistic resources (Figure 6). Firstly, users can easily verify whether the same headword exists in other dictionaries included in the wiki. Secondly, it is proposed to search the headword in

other Latin dictionaries. Lastly, links to textual corpora and text collections are provided. This is also where lexicographical data may be enriched with world knowledge. The example of *mandragora* ‘mandragora’ shows possible fields of lexicon-encyclopaedia interface enrichment: here, links point to plant taxonomy pages, to the Wikipedia entry on *mandragora*, and to the images accessible in Wikimedia Commons.

5. Search and Browse Capabilities

Search and browse capabilities of the presented infrastructure are partly known from Wikipedia and its derivatives. It comes as no surprise that entries may be retrieved by means of a simple full text search. As in Wikipedia, when typing a word beginning in an ajax-based search form, the user is given suggestions. Naturally, it is also possible in the advanced mode to limit search results to specific namespaces, i.e. dictionaries.²³

The framework, which is the object of the present study, seems to reach its full potential, however, thanks to the semantic layer provided by SMW. Semantic properties embedded in each entry can be browsed, for instance, thanks to the factbox displayed on the bottom of the entry page (Figure 7).

Faits relatifs à Mandragora — Recherche de pages similaires avec		Voir comme RDF
Art Définition	mandragore, plante	
Art Genre	f.	
Art Paradigme	-e	
Art forme	mandragoras	
Art thema		
Has image	commons:NaplesDioscuridesMandrake.jpg, commons:File:Mandragora Tacuinum Sanitatis.jpg, commons:File:Mandragora autumnalis 056.jpg	
Lemme	mandragora	
Lemme canonical	mandragora	
Possède un sous-objet	EU:Mandragora, EU:Mandragora, EU:Mandragora, and EU:Mandragora	

Figure 7: Entry page factbox

After clicking the “magnifying glass” icon near the value of each property, the user is taken to the page where all the entries with the same value set for the selected property will be listed. For example, in the case of *mandragora*, if one clicks on the zoom icon near the value *f.* (*femininum*, *lat.* ‘feminine’) of property “gender”, one is redirected to the page where all feminine substantives from all dictionaries included in the wiki are listed. Similar results can be obtained from the “Special Page”, where users can process a simple semantic search, by directly specifying in a two-field form, the property and its value they are looking for (Figure 8).

²³ Full-text search capabilities may be enhanced by using plugins list at http://www.mediawiki.org/wiki/Fulltext_search_engines. They have not been subject to the tests in the present survey.

Search by property

Search for all pages that have a given property and value.

Property: Value:

Figure 8: Direct search for values of semantic properties

More advanced semantic queries can be formulated from within two other search interfaces available in SMW-based wiki, accessible from “Special Pages”: Special:Ask and Special:BrowseData. The first (Figure 9) requires of the users a basic knowledge of SMW syntax, but it also provides them with numerous output formats from which they can choose, e.g., different types of charts, timelines, maps, tables, slides etc.²⁴

Semantic search

Query

```
[Category:Voces][[Art Genre::f.]]
```

Format as: For a detailed description, please visit the [Broad table \(default\)](#):

Sorting
[\[Add sorting condition\]](#)

Other options

limit: <input type="text"/>	The maximum number of results to return	offset: <input type="text"/>	The offset of the first r
sort: <input type="text"/>	Property to sort the query by	<input type="checkbox"/> descending <input type="checkbox"/>	<input type="checkbox"/> rand <input type="checkbox"/> random
mainlabel: <input type="text"/>	The label to give to the main page name	intro: <input type="text"/>	The text to display bef
searchlabel: <input type="text" value="\\x26hellip; autres résultats"/>	Text for continuing the search	default: <input type="text"/>	The text to display if tl

[Hide query](#) | [Show embed code](#) | [Querying help](#)

Previous **Results 1 - 6** Ne

Decipula
Decipula
Decipula

Figure 9: Semantic search (Special:Ask page)

“Special:BrowseData” (Figure 10), on the other hand, includes search patterns envisaged by each wiki creator and depends only on their creativity, user requirements, and last but not least, time or funding limitations. It enables faceted browsing of semantic properties of wiki data. In the case of our framework, the data in question are source and entry pages. The latter can be browsed according to the

²⁴ Display of search results is provided by Semantic Result Formats plugin (http://semantic-mediawiki.org/wiki/Semantic_Result_Formats).

part of speech they represent, inflectional type, gender, domain of use, etc., while the first can be browsed according to all the metadata previously mentioned.

Browse data: Voces

Voces
Click on one or more items below to narrow your results.

Choose a category:
Fontes (120)
Voces (11)

▼ Genus:
f. (6) · f.? (1) · m. (2) · n. (1)

▼ Paradigma:
-ari, -atus sum (1) · -ae (6) · -are, -avi, -atum (5) · -ari, -atus sum (1) · -e (1) · -i (1) · -ntis (1) · 1. (2) · Part. praes. (1) · adv. (2) · coni. (1) · praep. (1) · praep. c. abl. (1) · praep. c. acc. vel abl. (1) · prp. c. abl. (1) · -e (1)

Verba in definitione:

▼ Usus:
bot. (1) · christ. (1) · in (1) · iur. (1) · nat. (2) · qui (1) · spec. (2)

Showing below up to 20 results starting with #1.
View (previous 250 | next 250) (20 | 50 | 100 | 250 | 500)

A	D cont.	M cont.
• Abbatizo	• Depost	• Medico
n	i	• Medicor

Figure 10: Semantic faceted browsing (Special:BrowseData page)

6. Collaboration

From its beginning, the wiki-based interface that is the subject of the present study has been conceived as a means of promoting collaboration between researchers of different expertise in medieval studies. Lexicographical data enriched with encyclopaedic information extracted from knowledge databases may be a good starting point for a prospective framework of medieval culture research. User contributions should be encouraged by the reuse of a Wikipedia-like interface, with its well-known collaboration feature, namely “Discussion page”. Despite the fact that MediaWiki was created for projects in which anonymous editing is welcomed, one can, however (1) impose access and edition limitations in order to get rid of the acts of vandalism and (2) provide admin users with the right to accept or deny any changes. Users who are familiar with wiki syntax can be assigned edit rights and contribute to entries or source pages without any difficulty. However, even non-technical oriented users may contribute to the wiki, if given the chance to use simple edit forms. In the framework demonstrated in the present paper, this is the case of the source pages which can be modified in a traditional way, by entering the wiki syntax code, or by filling in forms provided by wiki developers (Figure 11).²⁵

One can expect the wiki to be fed at first with data from ongoing lexicographic projects, and later enriched by the users themselves. Batch import of lexicographical information from existing dictionaries may be carried out, e.g., by means of RDF

²⁵ This is possible thanks to the plugin called Semantic Forms (http://www.mediawiki.org/wiki/Extension:Semantic_Forms).

import plugin. In spite of the fact that the dictionaries under analysis may seem to differ essentially, at least as far as their micro- and macrostructures are concerned, their electronic versions are considered to be TEI compliant, and follow rules indicated in the chapter of TEI Guidelines devoted to the encoding of machine-readable dictionaries (TEI Consortium, 2013).²⁶ Since WikiLexicographica has to serve as a common interface for data retrieval, shared information schema should be conceived as well. The extent of data extraction could then be decided according to time or financial limitations; however, the burden of mapping between particular schemas and the common one needs to be shifted to each lexicographic team.

The second main contributor of WikiLexicographica is expected to be the research community, namely philologists, linguists, historians, palaeographers; briefly, all those who work with Medieval Latin texts. Apart from simple form or meaning corrections and additions, users may be encouraged, e.g., to propose the addition of new words found in their sources or the deletion of existing ones if manuscripts deny lexicographers' reading; to supply entries with world knowledge which in turn can greatly support text comprehension; to create links between words by making their relations explicit, and so on.

7. Conclusions

MediaWiki, the software underlying Wikipedia, is enhanced with semantic data annotation capabilities offered by Semantic MediaWiki extension, and appears to be a tool mature enough to serve as an interface for lexicographical data retrieval. It provides presentation and collaboration features with which an average Wikipedia user can already be familiar. Interface popularity itself is likely to encourage contributions, even from those less technical-oriented researchers. It is, however, the retrieval of semantic properties that should attract a major interest of researchers since charts, timelines and maps, as well as embedded queries, offer a fresh and inventive look at lexicographical data.

8. Acknowledgment

This work was supported by the following grants: ANR Omnia "Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins"; NCN 3736/B/H03/2011/40 "eLexicon Mediae et Infimae Latinitatis Polonorum".

Collaboration on this paper was made possible by support of COST Action 1005 "Medioevo Europeo" (www.medioevoeuropeo.org).

Content of the WikiLexicographica was partially typed by members of respective

²⁶ So far no attempt has been made in order to standardize medieval Latin dictionaries schema according to, for instance, Lexical Markup Framework.

dictionaries teams: Renaud Alexandre (Novum Glossarium), Susanna Allés Torrent (Glossarium Mediae Latinitatis Cataloniae), Pavel Nývlt (Lexicon Bohemorum), Marta Segarrés Gisbert (GLMC).

9. References

- Du Cange, C.D.F. (1883). *Glossarium mediae et infimae latinitatis (Editio nova aucta pluribus verbis aliorum scriptorum a Leopold Favre) conditum a C. Du Fresne, domino Du Cange, auctum a monachis ordinis sancti Benedicti; cum supplementis integris D. P. Carpenterii, Adelungii, aliorum suisque digessit G. A. L. Henschel*, Niort: L. Favre.
- Heid, C. (2004). Table ronde “Lexicographie et informatique” (Barcelone, 2 juin 2004). *ALMA, Bulletin du Cange*, 62, pp.327–332.
- Langlois, C.-V. (1924). Historique sommaire de l’entreprise de 1920 à janvier 1924. *ALMA, Bulletin du Cange*, 1, pp.1–15.
- TEI Consortium (2013). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version: 2.5.0. Last modified: 26th July 2013. Accessed at: <http://www.tei-c.org/Guidelines/P5/>.

The modern electronic dictionary that always provides an answer

Daiga Dekšne, Inguna Skadiņa, Andrejs Vasiļjevs

Tilde, Vienības gatve 75a, Rīga, Latvia

E-mail: daiga.deksne@tilde.lv, inguna.skadina@tilde.lv, andrejs@tilde.lv

Abstract

This paper presents the Tilde Dictionary Browser (TDB), an innovative dictionary browsing environment for a wide range of users: language learners, language teachers, translators, and casual users. We describe several techniques to maximise the likelihood of providing users with a useful result even when searched items do not have a direct match in the dictionary due to misspellings, inflected words, multi-word items or phrase fragments, or where there is a lack of data in the main dictionary. TDB is targeted for broad use on multiple platforms and is implemented as desktop software, and a Web and mobile application. The desktop version of TDB currently contains dictionaries for more than 20 language pairs, including the languages of the Baltic countries, and is easily extendable to other languages. Besides the data from translation dictionaries, TDB also provides information from different online resources, such as terminology dictionaries, as well as integrates the machine translation facility.

Keywords: electronic dictionaries; machine translation; spelling checker; morphological analyser; text-to-speech synthesis.

1. Introduction

In the last two decades, electronic dictionaries have been established among the most widely used software applications for non-English speakers, and the majority of users prefer electronic dictionaries to printed ones (Koren 1997). Different models for electronic dictionaries have been of interest to researchers for a long time (for an overview, see de Schryver, 2003). In their work, Oppentocht and Schutz (2003) describe the advantages of electronic dictionaries (e.g., explicit information, consistency, reusability, etc.). Detailed analysis of electronic dictionaries from different viewpoints is presented by Müller-Spitzer; her later findings are related to user needs and usage scenarios of electronic dictionaries (Müller-Spitzer et al., 2011). There has also been a lot of research on the typology of electronic dictionaries (e.g., Ide 1993; Sharpe 1995; Lehr 1996) and the different types of users.

When using a paper dictionary, the user usually must flip through pages to find the sought-after entry, whereas when using an electronic dictionary, the user can type the word in a search field or choose an entry from a word list. However, several authors (Měchura 2008; Nessi and Hail 2002) point out that users often fail to locate the information that they need. Users often search dictionaries for words that cannot be found in them, or cannot be found in the form in which they have typed them:

misspellings, inflected words, multi-word items, phrase fragments or even whole sentences. Many electronic dictionaries fail to return useful results when being searched for anything other than exactly matching units.

The aim of our work was to develop a dictionary software that is able to provide useful information for all types of search queries and information needs, including many problematic cases, i.e., when searched items do not have any direct matches in the dictionary data.

In the dictionary software, Tilde Dictionary Browser (TDB), that is presented in this paper, we have applied several techniques to maximise the likelihood of providing users with useful results:

- The entries from a main dictionary and possibly several terminology and explanatory dictionaries are merged in a single list, allowing users to get consolidated information from **several dictionaries simultaneously**.
- In the case of **incorrect spelling**, TDB suggests possible corrections and provides their translations.
- **For languages with rich morphology**, users can find translations for words that are not in base form, as usually dictionary entries are. With the help of the morphological analyser, possible base forms are obtained and their translations are displayed.
- Users can also **see** all of the **inflectional forms** for a particular word.
- If a user wants to see **usage examples** for a particular word, the search engine will show all dictionary entries containing this word, even if it is not a headword or translation, but part of a longer multi-word phrase.
- Users can also **search** terminology dictionaries **in the Web**, and the results will be displayed in the same uniform way along with the local dictionary entries.
- If there is **no entry in lexicon** to a user's request, the request can be redirected to a machine translation (MT) system on the Web, which will then translate and present the translation in TDB translation view.
- For those who are learning a language, TDB provides a **text-to-speech facility** that allows to hear the pronunciation of the selected dictionary entry.

Currently, TDB includes numerous general and specialised dictionaries for 19 translation directions: from English, French, German and Russian into Latvian and vice versa, from English, French, German and Russian into Lithuanian, as well as Latvian-Lithuanian, Lithuanian-Latvian and Estonian-Latvian. More than 25 terminology dictionaries are integrated into the TDB.

The dictionary content is licensed from leading lexicographers (authors of printed

dictionaries). The cooperation with authors goes beyond the licensing of existing content of printed dictionaries: using corpora processing techniques we provide lexicographers with lexical items that are not included into dictionaries as they have appeared recently. Such lexical items are then investigated by lexicographers and after validation added to the corresponding TDB lexicon. As a result TDB allows the location of lexical items that are not yet available from any printed dictionary.

TDB has been incorporated into several commercial products (Tildes Birojs, Tildes Biuras) and is also extended (while maintaining the same functionality) for dictionary look-up on the Web and on mobile phones. It is one of the most popular software applications in the Baltic countries, with about 400 000 users.

In this paper, we describe the functionality of the Tilde Dictionary Browser in detail, demonstrate the importance of language technologies in a modern electronic dictionary, discuss scalability and interoperability issues in different media, and present common application scenarios for a modern electronic dictionary.

2. Consolidation of data in dictionary entry creation

While a printed dictionary limits a search to the particular dictionary, electronic dictionaries can provide users with the ability to work with **several dictionaries simultaneously**. For this, entries from a main dictionary, and possibly several terminology, explanatory and synonym dictionaries, are merged in a single alphabetical list. Users can browse the entry just by clicking on a particular word in a list or search for a particular word or phrase by typing or copying it in a search field.

2.1 Forming a lexical entry: merging different sources

A logical part of a dictionary is an entry. However, dictionary entries may have very diverse formats. Some entries are very simple – just a word in a source language and a single or several translations in a target language.

More complex entries may contain translations grouped into several meanings, pronunciations, grammatical information, comments, usage samples and their translations, and explanations. Explanatory and synonym dictionaries usually have entries in a single language, while entries in translation and terminology dictionaries usually are in two or more languages.

The original formatting of dictionary entries is also very different: from simple tab or space separated words to entries with a rich formatting. Some samples of diverse dictionary formats are shown in Figure 1.

dangerous [ˈdɛndʂoros] *a* pavojīngas, ~ *illness* pavojīga/sunki liga; ~ *driving* pavojīngas važiavimas; **to look** ~ atrodyti suerzintam/pavojīngam

dangle [ˈdæɪrɟl] *v* tabaluoti, kyburuoti, kaboti, karoti; pakabinti; **to** ~ *one's legs* tabaluoti/maskatuoti kojas/kojomis Δ **to** ~ *smell in front of, or before, smb* sūlyti kam ką gundančio, vilioti ką kuo

all # viss # все
 all over # visam pāri # весь
 all over # visam pāri # полностью
 allow # atļaut # позволять

ader der Pflug, -"e
administraator der Administrator, -en
adressaat der Adressat, -en
adreseerina adressieren (an A), richten (an A)
advokaat der Rechtsanwalt, -"e
advokatuur die Anwaltschaft, -en

Figure 1: Samples of different dictionary formats in printed dictionaries

The task of a modern electronic dictionary browser is to present the entries from different sources in a uniform way. This is achieved by parsing original dictionaries and internal representation of their entries in an XML format.

We have developed a special XML format for dictionary entry representation (Figure 2). This format differs from Text Encoding Initiative (TEI) guidelines, however, it can be transformed to TEI rather easily. About twenty different XML tags mark the different semantic parts of an entry, but not all of them are used in every dictionary.

<pre> <entry title="ābece"> <title>ābece</title> <gram>n</gram> <mean digits="1" symbol="."/ > <transl>ABC</transl><comment>(book)</comment> <transl>primer</transl> <mean digits="2" symbol="."/ > <transl>ABC</transl> <usage>pārn</usage> <transl>the rudiments</transl> <from_sample>ābeces patiesība</from_sample> <to_sample>platitude</to_sample> <to_sample>self-evidence</to_sample> <to_sample>truism</to_sample><comment>(man)</c omment> <idiom /> <from_sample>tā ir ķīniešu ābece</from_sample> <to_sample>it is all _Greek</to_sample><comment>(to me)</comment> </entry> </pre>	<p>ābece <i>n</i> 1. ABC (<i>book</i>), primer; 2. ABC <i>pār</i>n the rudiments; ābeces patiesība - platitude, self-evidence, truism (<i>man</i>); ♦ tā ir ķīniešu ābece - it is all Greek (<i>to me</i>)</p>
---	--

Figure 2: Sample of dictionary entry in printed dictionary (right) and XML format (left) for the dictionary entry *ābece*.

Each entry is included in <entry> tag. Every entry starts and must have at least one <title> tag that represents the lexical entry. Other possible tags include:

- part of speech and other grammatical information, enclosed by a <gram> tag;
- in bi/multi-lingual dictionaries, there usually are one or several <transl> tags which are used to describe the translation;
- <link> tag, used to point at another related entry;
- <from_sample> tag, enclosing a sample in the source language, and the following <to_sample> tag, enclosing its translation into the target language. In case of a monolingual dictionary, only the <from_sample> tag is used;
- <comment> tag, enclosing additional contextual information that is specific to the entry, its translation, or sample phrase.

Diversity of XML tags helps to preserve the rich content of a dictionary, very close to its original view. When a dictionary entry is presented to a user, the dictionary entry is transformed from XML format to HTML view, and different XML tags are specifically formatted: bold, italic, different font size and different font colour (Figure 3).

ābece
 1. **ABC** (*book*), primer;
 2. *pār.* **ABC**, the rudiments;
 ābeces patiesība - platitūde, self-evidence, truism;
 ◆
 (*man*) tā ir ķīniešu ā. - it is all Greek (*to me*) ..

Figure 3: Dictionary entry in the electronic dictionary for the word *ābece*.

Although dictionary entries are merged, a user still has the possibility to search in a single dictionary (or several dictionary sources), as TDB allows all dictionary sources to be seen for each translation direction, or select a particular dictionary (or dictionaries).

2.2 Adding terminology data

In addition to general language dictionaries, terminological data is another type of resource that can be very useful for translation or comprehension of lexical units, particularly if a user is dealing with a text in a specialized domain.

TDB provides two options for integrating terminological data. A terminology resource can be added as an additional local terminology dictionary or accessed as a remote online resource.

Local terminology dictionaries are provided in a similar manner, as lexical dictionaries. Terms are automatically added to the list of all headwords for the source language that is displayed on the left side pane of the Dictionary Browser (excluding

duplicates, in case some similar general language headword is already present in the list). Users can also access a terminological entry using the search feature.

Terminology entries that match the selected headword or a search query are displayed in a separate terminology section on the right side pane of TDB.

Although representation of terminological entries is similar to that of lexical entries, there are important conceptual differences. While in a lexical entry all of the meanings are grouped under one headword, in the case of terminology data, there are separate entries displayed for each term corresponding to the search criteria. This approach is chosen because we follow the concept based principle for the organisation of terminological data. According to this approach, every terminological entry corresponds to one concept. One concept may have several lexical units denoting it, but a single terminology entry may not depict more than one concept.

Figure 4 shows this approach for an example of terminology data found for the search-word *communication*. Several terminology entries are displayed from a number of terminology dictionaries on different subject fields.

The screenshot shows a search interface titled "Terminu vārdnīcas" (Terminology Dictionaries). It displays several entries for the search term "communication":

- Entry 1:** LV **sakari** ekon. Sk. *arī sakars*
EN communication, ties
Ekonomikas, lietvedības un darba organizācijas (ELDO) termini
- Entry 2:** LV **satiksme** ekon.
EN service, communication
Ekonomikas, lietvedības un darba organizācijas (ELDO) termini
- Entry 3:** LV **komunikācija** muzeol.
EN communication
Muzeoloģijas terminu vārdnīca
- Entry 4:** LV **saskarsme** ped.
LV Cilvēkdarbības procesos cilvēku un to grupu mijiedarbībā, informācijas apmaiņā (saziņā) u. tml. uz savstarpējām attiecībām balstīta garīgā saskare, iekšējā saikne. Saskarsmes raksturs atkarīgs no personas īpašībām, saskarsmes prasmes. Nonākot saskarsmē ar lielām personībām, cilvēks bagātinās. Uz personas attīstību savu ietekmi atstāj arī saskarsme ar parādībām un notikumiem apkārtējā sabiedrībā, dabā.
EN communication, interaction, interface
Pedagoģijas terminu skaidrojošā vārdnīca
- Entry 5:** LV **saziņa, sazināšanās** ped.
LV Process, kurā dara zināmas vienam otra, citam cita domas; savstarpēja informācijas apmaiņa starp personām tieši vai izmantojot dažādus tehniskos līdzekļus. Datortehnikas un datortīklu attīstība devusi iespēju cilvēkam sazināties ar augstas tehnoloģijas ierīci – datoru – un saņemt informāciju, kāda pieejama datoru tīklā (tīklā Internet). Sk. *arī komunikācija* communication
Pedagoģijas terminu skaidrojošā vārdnīca

Figure 4: Representation of terminological data from multiple resources and domains

Terminological data of multiple domains can be very voluminous with many specific and rarely used terms. This makes it impractical to provide all of these data locally. Our approach is to limit the data stored on a user's computer only to the most-used domains, such as economics and finance, law, legislation and information technology. Other terminological resources are accessed through dynamic querying of online

sources. This also ensures the up-to-datedness of information, as new terms are being introduced, and some older terms become depreciated or changed.

For TDB, such an external terminology resource is EuroTermBank¹. It provides free web-based access to the richest collection of European multilingual terminology from a variety of collections and domains (Vasiljevs et al. 2008). Its database currently contains approximately 2.6 million terms from 137 terminology resources in more than 30 languages. EuroTermBank provides not only terms stored in its repository, but also matching terms retrieved from external online terminology databases, such as the database of the Terminology Commission of Latvia² and EU inter-institutional terminology database IATE³.

EuroTermBank provides a common application programming interface (API) to query its data by external systems. This API returns terminology data in the TBX format. TBX (TermBase eXchange) is a standard format for terminology exchange developed by the Terminology Special Interest Group of the recently dissolved Localization Industry Standards Association (LISA). In 2008, this format was adapted by ISO as international standard ISO 30042:2008. Terminological data is organized in data categories that are compliant to ISOcat data category registry as defined in ISO 12620.

TDB queries EuroTermBank for the word searched by the user and processes the received result to represent it in a way similar to that of terminology data from locally stored resources. As online querying of EuroTermBank may take some time depending on the speed of the user's Internet connection, it is optional, and the user can easily switch it on or off.

The terminology entry represented to a user includes such data as the term in the source language, its equivalent in the target language, subject domain, definition (if provided) and the source of data, e.g., information about the terminology resource from which this particular entry originates.

3. Integration of language technologies

While the basic functionality of the electronic dictionary is realized through a common data format and efficient search algorithms, the more advanced and important features are realised through integration of several language technology solutions. For different tasks, TDB uses spelling checker, morphological analyser, text to speech engine, and machine translation services.

¹ <http://www.eurotermbank.com>

² <http://termnet.lv>

³ <http://iate.europe.eu>

3.1 Language technologies that enrich search facilities

The integration of **spelling checker** into TDB plays an important role for users in two cases: (1) for a language with rich diacritics, a spelling checker helps to correct mistakes of forgotten diacritics (see Figure 5), and (2) for users with insufficient knowledge of a language (e.g. a foreign language learner or a child), spelling checker helps to correct errors in words with complicated spelling. In both cases, the task of spelling checker is to help the user find a translation in cases when an incorrect lexical entry is requested.

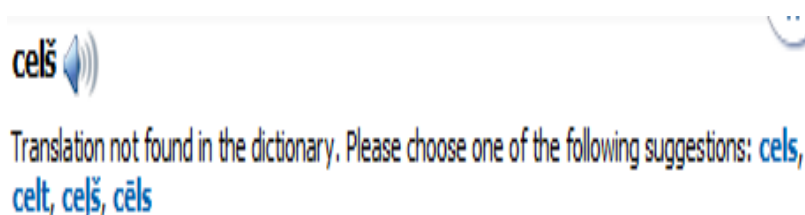


Figure 5: Suggestion from spelling checker for incorrect Latvian word *celš*

More advanced, but similar functionality is provided by the **lemmatizer** and **morphological analysis tools**. These tools allow a user to find translations for forms that differ from the lexical entry. This feature is very useful for highly inflected languages where word form can vary significantly from the base form, as illustrated in Table 1 for the verb *iet* (*to walk*).

	<i>Present</i>	<i>Past</i>	<i>Future</i>
<i>1st pers. sing.</i>	<i>eju</i>	<i>gāju</i>	<i>iešu</i>
<i>2nd pers. sing.</i>	<i>ej</i>	<i>gāji</i>	<i>iesi</i>
<i>3rd pers. sing.</i>	<i>iet</i>	<i>gāja</i>	<i>ies</i>
<i>1st pers. plur.</i>	<i>ejam</i>	<i>gājām</i>	<i>iesim</i>
<i>2nd pers. plur.</i>	<i>ejat</i>	<i>gājāt</i>	<i>iesiet</i>
<i>3rd pers. plur.</i>	<i>iet</i>	<i>gāja</i>	<i>ies</i>

Table 1: Inflected forms for verb *iet* (*to walk*)

The morphological analyser can also play the role of disambiguator in a dictionary. In the case of the entered word form corresponding to several base forms, the morphological analysis tool allows to choose between them and leads to the most appropriate translation (see Figure 6).

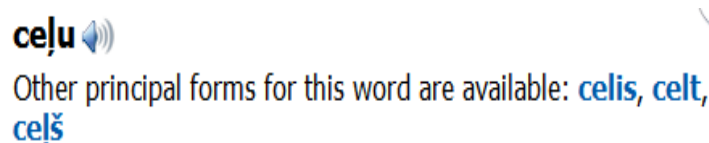


Figure 6: Suggestions of the morphological analyser for word form *ceļu*.

Finally, the morphological analyser is used as a reference tool that allows all inflectional forms of the word to be seen. As mentioned before, this is an important feature for inflected languages with a rich morphology. For instance, in the Latvian language, many palatalised forms occur for nouns. Although palatalisation rules are rather regular, some exceptions exist for each particular case, forming a set of exceptions, words which in many cases are spelled incorrectly even by native speakers.

3.2 Content enrichment through machine translation

The language technologies described above enrich search facilities in dictionary content and help users find a necessary dictionary entry. However, all dictionaries are limited in size and content and no dictionaries contain all possible words for a particular language and their translations. One possibility of how to extend coverage of translation dictionary content is to apply machine translation. Translations suggested by the machine translation system are not always perfect, but in many cases, they provide an added value for the user. Moreover, integration of the machine translation system into the dictionary software allows a user to translate a phrase or sentence with a particular word, thus allowing the user to find its contextual meaning (Figure 7).

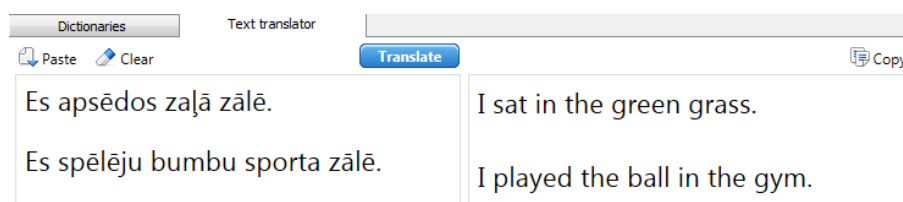


Figure 7: Machine translated samples for word *zāle* (*grass/hall*)

4. Dictionary content in different media

As there are more and more different devices where dictionaries could be presented, it is important to develop a dictionary browser that is interoperable between different platforms and devices. TDB is implemented not only as a desktop application, but also as a Web dictionary and mobile application. The same data modules are searched to translate a word or phrase upon user request. Only the way in which results are presented differs. The form in which results are presented depends on the size of the device, Internet access and other limitations.

As a desktop application, TDB has no limitation in the presentation of results. If a result does not fit on a visible part of the window, the result window has a scroll bar. The results from main dictionaries, term dictionaries, and synonym dictionaries are on separate foldable panels, which, if opened, show translations of particular types

and while in a folded state, do not take up much space in the result window (see Figure 8).

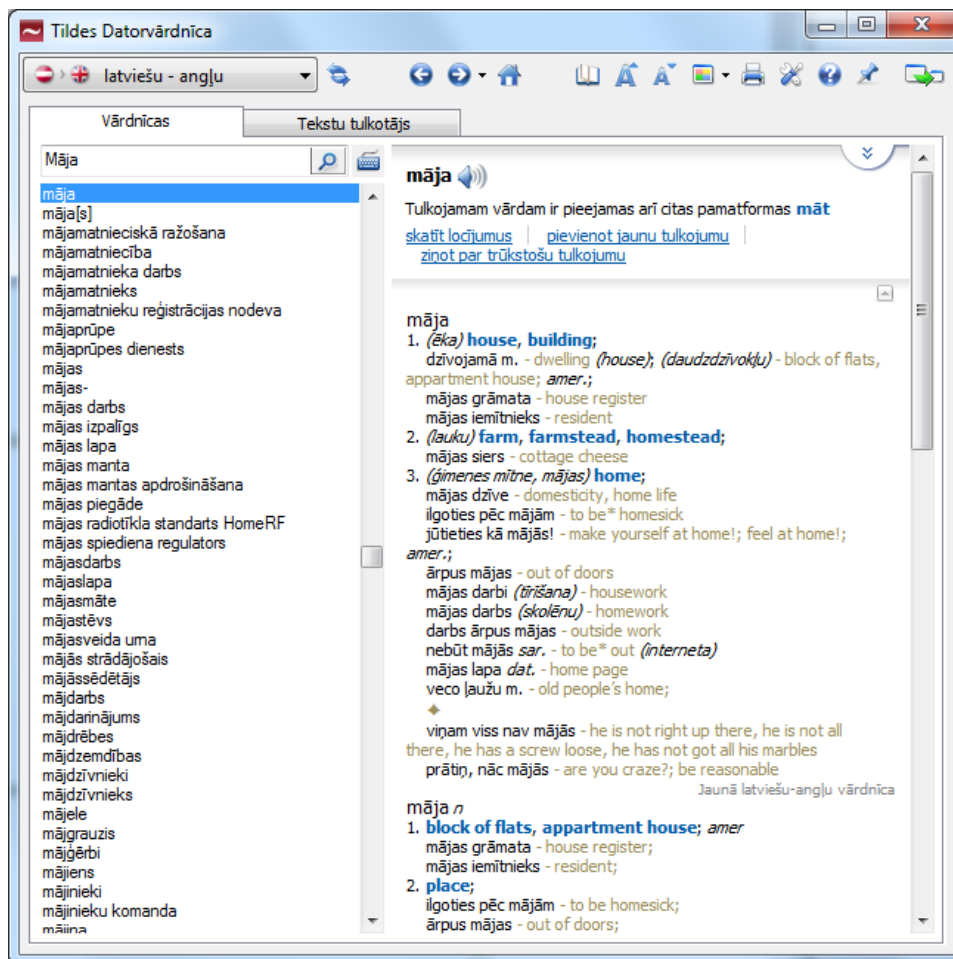


Figure 8: Results for word māja in Tilde Dictionary Browser

In TDB, a user can click on a link and add a new translation to the current entry or send a report to dictionary creators about a missing translation. A user can also switch to the Text translation tab, which allows the user to translate texts with an online Machine Translation service.

All dictionaries available from TDB are also available from the Web portal *letonika.lv* (Figure 9). Here, advanced search options are also available.

In mobile devices, the window for result presentation is much smaller than for a computer screen, and accordingly, less information can be displayed. Therefore we show a limited number of translations from the main dictionary and a limited number of usage samples (see Figure 10).

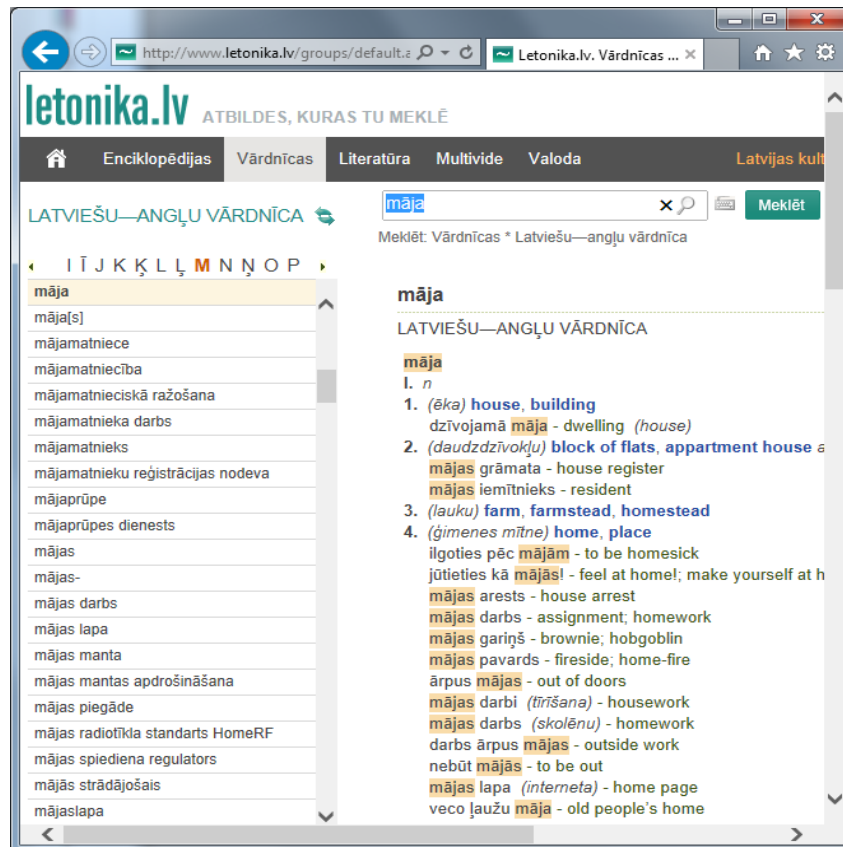


Figure 9: Results for the word *māja* in Web application



Figure 10: Results for word *māja* on a mobile phone

5. Other features to increase applicability

A number of usability features are implemented to facilitate fast and efficient work with TDB. Users can switch between full view and compact view that provides only the essential translation information in a smaller window. Compact view is particularly useful if a user needs to consult a dictionary very often. Then, TDB can stay open as a foreground application (always on top of other open windows) that occupies relatively little space on the screen.

If a user is reading text in a Web browser, text editor, or some other application and needs to quickly find a translation of a particular word, then TDB can be easily accessed by pressing a hot-key combination. In several applications like popular Web browsers and MS Word versions, the translation command is also included in the context menu evocable by the right-click of a mouse.

To facilitate the typing of search words, the keyboard is automatically switched to the target language layout. Special characters can also be typed by using an integrated on-screen keyboard.

A user can also create user dictionaries that can be local or shared throughout an organization. New entries in a user dictionary can be created from the TDB interface or by directly writing into the dictionary file that has a simple to understand text-based format.

Besides phonetic transcription of headword pronunciation, TDB makes it possible to listen to a particular translation, a sample of usage, or even a fragment of text. This feature is enabled through the integration of a **text-to-speech** engine. Currently, TDB integrates Latvian TTS developed by Tilde (Goba and Vasiljevs 2007) and English TTS provided by Microsoft. Microsoft Speech API is used for the TTS integration making it easy to extend language support with other MS SAPI compliant TTS engines.

6. Conclusion and tasks for the future

In this paper, we presented the electronic dictionary software TDB, that, in addition to simple search and browsing, also supports different language technology driven services that facilitate better retrieval of requested entries in non-trivial cases.

TDB can be used on different platforms, including mobile devices and the Web. Currently, 20 language pairs are supported for general content dictionaries. However, more language pairs can be easily incorporated, and additional dictionaries for current language pairs can be added.

Development of a user-friendly dictionary is a never-ending process. Our development plans include two directions: extension in content and extension in

functionality.

With respect to functionality, two major extensions are planned. Firstly, we plan to support specialists and language learners with extended context for a selected lexical item by providing concordances from corpora. Secondly, closer integration with machine translation is planned, thus allowing users to translate a full document instead of a phrase, sentence, or small fragment of text.

7. Acknowledgements

The research leading to these results has received funding from the research project “Information and Communication Technology Competence Center ” of EU Structural funds, contract nr. L-KC-11-0003, signed between ICT Competence Centre and Investment and Development Agency of Latvia, Research No. 2.8 ”Research of automatic methods for text structural analysis”.

8. References

- Bogaards, P. (2003). Uses and users of dictionaries. In van Sterkenburg, Piet (ed.), *A practical guide to lexicography, Terminology and Lexicography Research and Practice 6*, pp. 26-33. Amsterdam: John Benjamins.
- Burke, S. M. (1998). *The Design of Online Lexicons*. Master's thesis: Northwestern University, Evanston, IL.
- de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. In *International Journal of Lexicography*, 16(2), pp. 143-199.
- Deksne, D., Skadiņa, I., Skadiņš, R., Vasiļjevs, A. (2005). Foreign Language Reading Tool – First Step Towards English-Latvian Commercial Machine Translation. In *Proceedings of Second Baltic Conference „Human Language Technologies – the Baltic Perspective”*, Tallinn, 2005.
- Goba, K., Vasiļjevs, A. (2007). Development of Text-To-Speech System for Latvian. In Joakim Nivre, H.-J. Kaalep, K. Muischnek, & M. Koit (Eds.), In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, Tartu, pp. 67–72.
- Ide, K. (1993). A Catalogue of Electronic Dictionaries. *Language* 22.5, pp. 42-49.
- Koren, S. (1997). Quality versus convenience: comparison of modern dictionaries from the researcher's, teacher's and learner's points of view. In *TESL-EJ* 2 (3).
- Lehr, A. (1996). Electronic Dictionaries. In *Lexicographica* 12, pp. 310-17.
- Lew, R. (2004). Which dictionary for whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English. Poznan: Motivex.
- Měchura, M. B. (2008). Giving them what they want: search strategies for electronic dictionaries. In *Proceedings of the 13th Euralex International Congress*, pp.

1295-1299.

- Müller-Spitzer, C. (2011). Textual Structures in Electronic Dictionaries compared with Printed Dictionaries. A Short General Survey. In: Gouws, Rufus H./Heid, Ulrich/Schweickhard, Wolfgang/Wiegand, Herbert Ernst (Hgg.): *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter.
- Müller-Spitzer, C., Koplenig, A., Töpel, A. (2011). What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project. In *Proceedings of eLex 2011*, pp. 203-208.
- Nesi, H. and Hail, R. (2002). A Study of Dictionary Use by International Students at a British University. In *International Journal of Lexicography*, 15.4: 277-305.
- Oppentocht, L. and Schutz, R. (2003). Developments in electronic dictionary design. In van Sterkenburg, Piet (ed.), *A practical guide to lexicography, Terminology and Lexicography Research and Practice* 6, 215-227. Amsterdam: John Benjamins.
- Sharpe, P. (1995). 'Electronic Dictionaries with Particular Reference to the Design of an Electronic Bilingual Dictionary for English-speaking Learners of Japanese. *International Journal of Lexicography* 8.1, pp. 39-54.
- Skadiņa, I., Vasiljevs, A., Dekšne, D., Skadiņš, R., Goldberga, L. (2007). Comprehension Assistant for Languages of Baltic States. In *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*, Tartu, pp. 167-174.
- Vasiljevs, A., Rirdance, S., & Liedskalnins, A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, Hong Kong, pp. 213–220.

Graphical representation of the web of knowledge. Analyzing the local hierarchies and the global network of connections in a specialized encyclopedia.

Daniele Besomi

Centre Recherches Interdisciplinaires Walras Pareto, University of Lausanne

E-mail: daniele.besomi@unil.ch

Abstract

Encyclopedias were originally intended to present knowledge in an organized way. Historically, this was often attempted by means of graphical metaphors, but strangely enough the advent of IT and the Internet seems to have hindered, rather than boosted, the original aspiration. This paper presents a tool that aims to visually place encyclopedic entries in their context. It takes the form of a working sample of a specialized ‘En-cycle-pedia’ concerned with the theories of economic cycles and crises, addressed to a learned readership. Its main feature is the representation of its articles as nodes of a directed graph, dynamically centred around the article under examination, linked by structural relationships representing different kinds of connections between articles (e.g., analytical or methodological), internal cross-references, and bibliographical references. The graph is customizable, as typologies of links and articles to be included can be selected by the user. Besides illustrating the relational structure of the encyclopedic contents, the graph also acts as a navigation tool. Moreover, users can experiment by editing both articles and the system of links, thereby turning the encyclopedia into an active analytical tool that permits the reader to compare different interpretations and their implications and premises.

Keywords: knowledge organization; Diderot & d’Alembert’s *Encyclopédie*; connected graph as rendition of connections between encyclopedic articles; macrostructural graphical representation.

1. Introduction

This paper propounds an encyclopedic approach that descends from a long tradition of thought concerning the organization of knowledge, in particular with the aid of graphical metaphors and tools. Before describing the apparatus in section 3, I will briefly recall some relevant reflections on these issues by classical encyclopedists (especially Diderot and d’Alembert) and their recent reappraisal in an epistemological perspective.

2. Graphical metaphors for the organization of knowledge

The etymology of the word indicates that an encyclopedia is a ‘circle of learning’ or a ‘chain of knowledge’. Thus, encyclopedias are meant not only to collect and present information, but to organize knowledge and order it in some way. Early encyclopedists often tackled this problem in graphical terms reflecting the epistemology that guides the chosen organization form. Medieval encyclopedias, such as Vincent de Bouvais’s *Speculum Majus* (1240–1260), the ‘great mirror’ consisting

in a *Speculum naturale*, a *Speculum doctrinale* and a *Speculum historiale*, or Honorius Augustodunensis's *Imago Mundi* (12th century), referred to the image of the mirror: the book discovers and reflects the order of nature and human affairs as created by God (Clark, 1992: 99–101; Van Ewijk, 2011: 208). Later, stairways and ladders arranged bodies in order of increasing perfection, describing both a path of improvement for man and the structure of knowledge, as exemplified by Lull's *Liber de ascensu et decensu intellectus* (written in 1305 but published in 1512) and Carolus Bovillus's *Liber de sapiente* (Paris, 1510) (Quaggiotto, 2011: 5–7). With the Renaissance, and in particular with Francis Bacon's classification of all human knowledge, the image of the tree served as an organizing metaphor for knowledge, following the order of reason rather than the order of the world.¹ Chambers's *Cyclopædia*, for instance, incorporated a table showing how knowledge branches from a common stock, depicting how its several parts relate to each other (Chambers, 1728, vol. 1: ii); similarly, Diderot and d'Alembert's *Encyclopédie* portrayed a 'tree of knowledge' (1751, vol. 1: xlvi–liii), likewise represented as a table and later illustrated by means of a tree by Roth in 1769. The plan of the *encyclopédistes*, however, was more radical: while depicting the tree as a key for the organization of the sciences, they also rejected a systematic approach that fixed knowledge into an unchangeable scheme, and used the tree and the system of cross-referencing as a guideline capable of encompassing science in its dynamics (see Salsano, 1977; Zimmer, 2009; Yeo, 2001).

As a reaction to the form of 'reasoned dictionary' adopted by the *Encyclopédie*, with its revolutionary implications, there was for a while a return to a more systematic approach. In particular, the methodic (unfinished) *Encyclopaedia Metropolitana* and the early editions of the *Britannica*, which introduced within the alphabetic sequence long essays on about 45 principal subjects, each of which was supported by 30 more lengthy articles to which the shorter articles on specific subjects referred. During most of the nineteenth century, however, encyclopedias largely turned into repositories of knowledge. Only later in the twentieth century did the older encyclopedic spirit enjoy a revival. However, the problem of the organization of knowledge and of the relationship between different branches of science gave way to the related, but by no means identical, issue of reconstructing the unity of science fragmented by the breaking down of subjects into semi-monographic articles of manageable size arranged in alphabetical, rather than thematic, order. As summarized in the 'Encyclopedia' article of *Britannica* "Even a brief survey of encyclopaedia publishing during the second half of the 20th century is enough to make it clear that ... a number of modern encyclopaedists [are] concerned with the importance of making a restatement of the unity of knowledge and of the consequent interdependence of its parts. Though most encyclopaedists were willing to accept the

¹ Besides his stairway, Lullus also produced an *arbor scientiae*, which, however, identified the order of knowledge with the order of creation (Salsano, 1977: p. 35).

essential reference-book function of encyclopaedias and the role of an alphabetical organization in carrying out that function, they became increasingly disturbed about the emphasis on the fragmentation of knowledge that such a function and such an organization encouraged. A number looked for ways of enhancing the educational function of encyclopaedias by reclaiming for them some of the values of the classified or topical organizations of earlier history” (Collson and Preece, 2013).

In its 15th edition (1974), *Britannica* carried a *Micropaedia*, with short definitions, a *Macropaedia*, or ‘knowledge in depth’, with longer entries, and a *Propaedia*, a topical guide to the opus, in its bid to be at once a reference work and an instrument of learning. The *Encyclopaedia Universalis* (1968–75) also focused more on science as a problem-solving activity than on the organized retrieval of results, and presented a series of *tableaux de relations* where it suggested by means of convoluted graphs different kinds of relationships (formal, methodological, ...) between concepts, some of which rather loose but evocative of connections worth exploring, beyond disciplinary boundaries. The *Enciclopedia Einaudi* explicitly declared its role as organizing knowledge rather than acting as a storehouse of notions, and decided to focus on a limited number of “concepts capable of organizing the knowledge and the life of mankind as a whole that ... revolve around very general problems” (Einaudi, 1977: xvi, xvii and xiii). It also offered in graphical form a grouping of entries representing ‘reading areas’, based on a logical reconstruction of the network of relationships between entries.

This was all very promising, but instead of being further elaborated by means of the possibilities offered by IT and the Internet, in the web versions of *Britannica* and *Universalis* all attempts to show the intricate relationships between branches of knowledge have been ditched altogether. In its most up-to-date entry on ‘Encyclopaedia’, *Britannica* is rather reticent on Internet encyclopedias and itself effaces any trace of the *Outline of knowledge* of the 15th edition. The omologous article in *Universalis* explicitly worries that “IT technologies and the internet are destructors of the encyclopedic spirit”.² The Internet version of *Universalis*, however, offers at least a thematic tree-index categorized by discipline branching into three further sub-levels.

²“Il est évident que l’existence d’Internet, où d’autres encyclopédies se créent et se créeront, où celles du passé peuvent être consultées en ligne, va dans le sens du projet encyclopédique de l’humanité, entrée dans l’ère du clavier. Mais l’instantanéité de l’électron, rendant accessible une accumulation de données et de liens jamais atteinte, ne donne aucune garantie de valeur, d’ordre ni de hiérarchie. En cela, l’informatique et l’Internet sont destructeurs de l’esprit encyclopédique incarné par Aristote, saint Augustin, Bacon, Locke, Leibniz, Condillac, Hegel, Coleridge ou Auguste Comte (pour s’en tenir à l’Occident), ce qui est au moins préoccupant. Dans *encyclopédie*, le ‘cycle’, le cercle est devenu sans limite, son centre étant partout et sa circonférence nulle part, et la ‘pédagogie’ que suscite *paideia* relève du self-service le plus hâtif. En même temps, la diffusion du savoir encyclopédique s’est largement accrue. Le présent nous lègue ce paradoxe; l’avenir ne le résoudra pas facilement” (Rey, 2013).

Even the Internet encyclopedia *par excellence*, Wikipedia, is rather modest in its claims regarding the organization of its materials – which is not surprising, given its essentially anarchic format. In its article on ‘Encyclopedia’, it recalls that “Some systematic method of organization is essential to making an encyclopedia usable as a work of reference. There have historically been two main methods of organizing printed encyclopedias: the alphabetical method (consisting of a number of separate articles, organised in alphabetical order), or organization by hierarchical categories. The former method is today the most common by far, especially for general works. The epigraph from Horace on the title page of the 18th century Encyclopédie suggests the importance of the structure of an encyclopedia: ‘What grace may be added to commonplace matters by the power of order and connection.’” The article continues by claiming that “The fluidity of electronic media, however, allows new possibilities for multiple methods of organization of the same content. Further, electronic media offer previously unimaginable capabilities for search, indexing and cross reference”. However, in presenting the influence of the Internet on encyclopedias it only stresses its “ever-increasing effect on the collection, verification, summation, and presentation of information of all kinds” and that “On-line encyclopedias offer the additional advantage of being dynamic: new information can be presented almost immediately, rather than waiting for the next release of a static format, as with a disk- or paper-based publication. The 21st century has seen the dominance of wikis as popular encyclopedias, including Wikipedia among many others” (Wikipedia, 2013). In truth, Wikipedia offers a number of cladistic ‘portals’ aiming at systematizing some fields; but it is precisely in some of these portals (e.g., business and economics) that the absence of co-ordination by an encyclopedist is most notable.

2.1 Trees vs. networks

Meanwhile, however, some of the issues raised by Diderot and d’Alembert have been taken up in the literature in an epistemological perspective. Umberto Eco noted that the *Philosophes* themselves made the tree metaphor inadequate (Eco, 1984: 80–84; see also Cernuschi, 1996; Bates, 2002; Chauderlot, 2002; Zimmer, 2009: 104; and Weigel, 2013: §21). While their *Système figuré des connaissances humaines* summarized the “genealogy and the filiation of the parts of our knowledge”³ and introduced the examination of “the causes that brought the various branches of our knowledge into being, and the characteristics that distinguish them” (d’Alembert, 1751, Eng. transl.: 5), they were fully aware that the image of the “encyclopedic tree” would be “disfigured, indeed utterly destroyed” if one were to take into account the actual intricacies, discontinuities, obstacles, U-turns and crossroads of thought processes. A more appropriate metaphor would be the labyrinth, to reflect the tortuous roads followed by the intellect, or the map (ibid.: 46–47). The encyclopedic

³ D’Alembert described the *Encyclopédie*’s “genealogical or encyclopedic tree” as gathering “the various branches of knowledge together under a single point of view and [serving] to indicate their origin and their relationships to one another” (1751: 45–46).

tree provides no more than “a kind of world map”, where only “the principal countries, their position and their mutual dependency, the road that leads directly from one to the other” are shown (ibid.: 47). Individual articles are placed on such a world map by means of a direct reference to the discipline(s) to which they pertain.⁴ The intricacies are shown by means of cross-references to other articles. D’Alembert is anxious to stress that “such references in this Dictionary are unusual in that they serve principally to indicate the connections of the materials, whereas in other works of this type, they are intended only to elucidate one article by another” (ibid.: 57). Diderot goes further, explaining that there are four kinds of links: explanatory (‘verbal’) references; the ‘material references’ that indicate close and distant relationships between objects, establishing analogies and consequences or, on the contrary, denying them; the ‘references of genius’ that suggest “new speculative truths” by imagining “fanciful conjectures” and drawing suggestive connections between distant fields; and the ‘satirical or epigrammatic references’ that deride “certain kinds of foolishness” and prejudices (Diderot, 1755: 642–644; see Anderson, 1986: 922–926 and 1990: Ch. 3; Zimmer, 2009; and Le Ru, 2002). The graphical representation by means of the tree image, with its linear branching, is naturally unsuitable to map these cross-references, for which the encyclopedists’ metaphor of the map is surely better fitted, as are the modern analogies of the rhizome (Eco) or the network.

A related issue discussed by the *philosophes* concerns the different understandings of phenomena or concepts according to the point of view from which they are examined. D’Alembert stressed that “as, in the case of the general maps of the globe we inhabit, objects will be near or far and will have different appearances according to the vantage point at which the eye is placed by the geographer constructing the map, likewise the form of the encyclopedic tree will depend on the vantage point one assumes in viewing the universe of letters. Thus one can create as many different systems of human knowledge as there are world maps having different projections, and each one of these systems might even have some particular advantage possessed by none of the others. There are hardly any scholars who do not readily assume that their own science is at the center of all the rest, somewhat in the way that the first men placed themselves at the center of the world, persuaded that the universe was made for them. Viewed with a philosophical eye, the claim of several of these scholars could perhaps be justified by rather good reasons, quite aside from self-esteem” (d’Alembert, 1751: 48).

Diderot similarly stated that “In general the description of a machine can begin with any part at all. The larger and more complicated the machine, the more connections there are between its parts, the less we know these connections, the more different perspectives for description there will be. What then if the machine is in every sense

⁴ The article ‘Eau’ (Water), for instance, refers to various domains including physics, medicine, hydraulics, pharmacy and chemistry.

infinite; if we are speaking of the real universe and the intelligible universe, or a work which is like the imprint of both? Either the real or the intelligible universe has infinite points of view from which it can be represented, and the possible systems of human knowledge are as numerous as those points of view” (Diderot, 1755).

This multiplicity of paths concerns both the encyclopedists and the readers. The very ‘world map’ of the *Système figuré des connaissances humaines* is an arbitrary construction, for other systems of classification could have been devised (d’Alembert, 1751: 49–50), while every local map – that is, any individual article in the *Encyclopédie* – offers many cross references to other articles, thus building innumerable possible paths that cannot be followed simultaneously, so that different minds will chose different routes at each crossroad (ibid.: 47).

3. The En-cycle-pedia: Connections between entries as a dynamically constructed graph

Constrained by the paper medium, Diderot and d’Alembert could hardly explore in full the implications of the challenge they set themselves. As we have seen, the approach of the *encyclopédistes* has remained largely underinvestigated by later encyclopedists, notwithstanding the increased potential opened by electronic editing and the Internet.⁵ It has, however, inspired the construction of the encyclopedic scheme presented here.

3.1 A graphical tool for organizing knowledge

This En-cycle-pedia ⁶ offers a tool for organizing knowledge, in the form of dynamically constructed graphs built around the article being read that places it in its context. Articles are represented as nodes, connected by directed arrows standing for various kinds of links between them. The graph, besides showing the paths branching off from each node, also acts as a navigation tool, each node being an active hyperlink redirecting to the corresponding article to which a new graph is associated. The tool is exemplified by means of a sample of a specialized encyclopedic dictionary concerned with the theories of economic cycles and crises, offering for the time being about 80 articles. The choice of the topic was determined by the fact that the subject is circumscribed and that the propounder of this project has sufficient expertise in the field to have formed an epistemological view as to its representation. The entries are drawn from his previously published writings, with the double purpose of illustrating the working of the encyclopedic structure and of (literally) connecting the dots between the various topics of his research: hence the name En-cycle-pedia,

⁵ The system of cross-references in the *Encyclopédie*, however, has been studied and represented by a graph depicting the connections between categories of entries (but not article by article): Blanchard and Olsen, 2002.

⁶ www.en-cycle-pedia.ch. Access is not yet public as the work is still in progress, but readers can test the tool by entering the following: ID: `elex.user`, password: `elex.user`.

emphasizing at once the focus on cycles and the encircling scope of the project. The chosen target audience consists of graduate students and researchers. Such an expert readership has been selected in order to experiment with a large number of variables in the graph, to which corresponds some complexity in the management of control parameters. The objective is to explore the possibilities opened by the tool; simplifications for a more generic audience can be introduced at any time.

3.2 Structure and cross-references

Like the *Encyclopédie*, the En-cycle-pedia starts from a general structure representing the ‘big picture’ resulting from the encyclopedist’s understanding of the En-cycle-pedia’s subject matter. This is constructed by dynamically linking together, one by one, the various articles in a hierarchical, or genealogical, order. The resulting graph thus shows how the encyclopedist orders knowledge on the basis of his interpretation of the connections between individual topics within the general scheme.

There are, however, a number of differences between such an arrangement and that of the *Encyclopédie*. Firstly, while the *philosophes* could only resort to the tree metaphor for their basic classification of knowledge, the En-cycle-pedia has no such constraint. Its graph is a complex network, constructed beginning from a meta-classificatory project⁷ and implemented by linking each entry to ‘genealogically’ connected articles. Locally, therefore, the structure is hierarchically organized. Globally, however, the tree soon turns into a non-linear network because lines of descent and ascent can be multiple and intersecting.

Secondly, while the encyclopédistes’ graphical representation in the *Système* could only envisage one kind of link, indicating the division of a topic into various sub-topics, links in the En-cycle-pedia can have various attributes. The first is strength (represented by lines of different thickness): some connections are more forceful than others, and it makes full sense to recognize this and to allow the user to decide whether to focus only on stronger links or also to examine less cogent connections—in the map metaphor, one can choose whether to depict only motorways or also national and local roads.

The second attribute is qualitative (represented by lines of different colours). Relationships between topics have different natures. In the En-cycle-pedia, which is concerned with the history of economic thought, one can distinguish relationships between entries based on the discipline’s methodology, or the general way of thinking

⁷ The En-cycle-pedia has a core article focusing on the “Classifications of crises and cycles theories”, discussing ten or so modes of classification suggested in the literature since the 1840s. Each mode of classification is discussed in detail in specific articles. Different specific theories are, of course, treated in different ways by each classificatory scheme, and are therefore linked to several of these schemes at once. The result is of necessity a rather intricate network.

about the subject; some articles are related at the analytical level, others by factual connections, or yet other connections characterize worldviews. These may be the encyclopedist's broad understanding of the relationships between the subject matters of the encyclopedia, or how different interpretations of the nature of the subject reflect into theoretical schools, approaches, etc. Naturally, knowledge in different domains suggests to focus on different qualities: an encyclopedia of jazz musicians, for instance, is more likely to be concerned with connections of the kind "plays compositions by ...", "has played with ..." or "is inspired by ...".

Thirdly, links have a preferred direction reflecting the hierarchical ordering of topics, which translates in directed or bidirectional arrows (see Figures 2 and 3). Other attributes could be added, making of course the network even more convoluted, and one should balance the benefits of finer representation with the increasing difficulty in usage and interpretation. Links appearing in the En-cycle-pedia's graph carry information on the reasons why they are set as they are.

Again like the *Encyclopédie*, the En-cycle-pedia superimposes the connections represented by the cross-references inserted in each article onto the systemic, hierarchical links, distinguishing the references to 'further information' from those inviting to consult 'in depth treatment' (see Figure 4).⁸ A finer division, such as Diderot's four categories described above, can naturally be envisaged, but again one should balance advantages, increased complexity and risk of overlapping, with the scope of the hierarchical links.

The En-cycle-pedia also treats the literature cited as 'bibliographic objects' connected to each article; such objects can be visualized as nodes in the graph, so that the mutual relationship between the references (alone, or in connection with the links between the En-cycle-pedia's articles) can also be explored, distinguishing, if desired, between primary and secondary sources. Each reference in the bibliographies directly offers the link to a graph representing the network of citations (Figure 5).

The information carried by the links can be visualized selectively in the graph: by link type (structural, cross-reference or bibliographic), and by selecting any of the attributes (quality and strength, type of cross-reference, primary or secondary literature); the depth (that is, the number of 'generations') can also be adjusted, showing longer or shorter chains of nodes, and the visualization can be further

⁸ A graphical tool depicting the 'main' cross-references between an entry and other articles is offered by the online version of Gabler's *Wirtschaftslexikon*. It is built starting from the system of cross-references to and from the article under examination: an unspecified algorithm selects the five most important entries, and represents them in their connections to the central article by means of incoming, outgoing or bidirectional arrows, depending on whether the entry refers to, is referred from, or both refers to and is referred from, the central article. Any cross-references between these entries are also represented in a lighter colour. Each node in the graph is a link to the corresponding article. The reader can also examine second-order connections, resulting from the iterated application of the algorithm to each of the five entries, giving rise to a 26-nodes graph (at most).

restricted by relevant dates (Figure 1). The result both places articles in their context, selectively defined, and suggests reading paths, which naturally can be explored by navigating the graph itself (one can visualize the articles' abstracts directly from the graph by positioning the cursor on the articles' labels). In contrast with the *Encyclopédie*, where a reader could visualize the (unique) 'world map' but could only follow the cross-references one at the time, the En-cycle-pedia offers at a glance a global perspective at the desired depth, enabling the reader to see distant connections otherwise hidden—concealed perhaps to the encyclopedist him- or herself. Each new article visualized forces a recalculation and re-drawing of the graph, with the new entry at its center. The user can thus personalize his or her reading experience, and save graphs for later usage.

3.3 Points of view: the En-cycle-pedia as an analytical instrument

The En-cycle-pedia radically interprets the second issue raised by Diderot and d'Alembert – that of the different perspectives from which one examines a certain problem. The encyclopedist offers a personal (hopefully well-informed) interpretation of the connections between topics, which is visually translated into the graphs. There is no reason, however, why the reader should necessarily agree with the encyclopedist's view. Users are therefore allowed (and invited) to implement their own interpretation by changing as extensively as they wish the system of structural links, editing the cross-references, modifying articles (including creating and deleting any), or revising the links to references—naturally on a separate, personalized copy of the encyclopedia, the settings of which can be saved and retrieved, shared with others and publicly discussed in a forum.

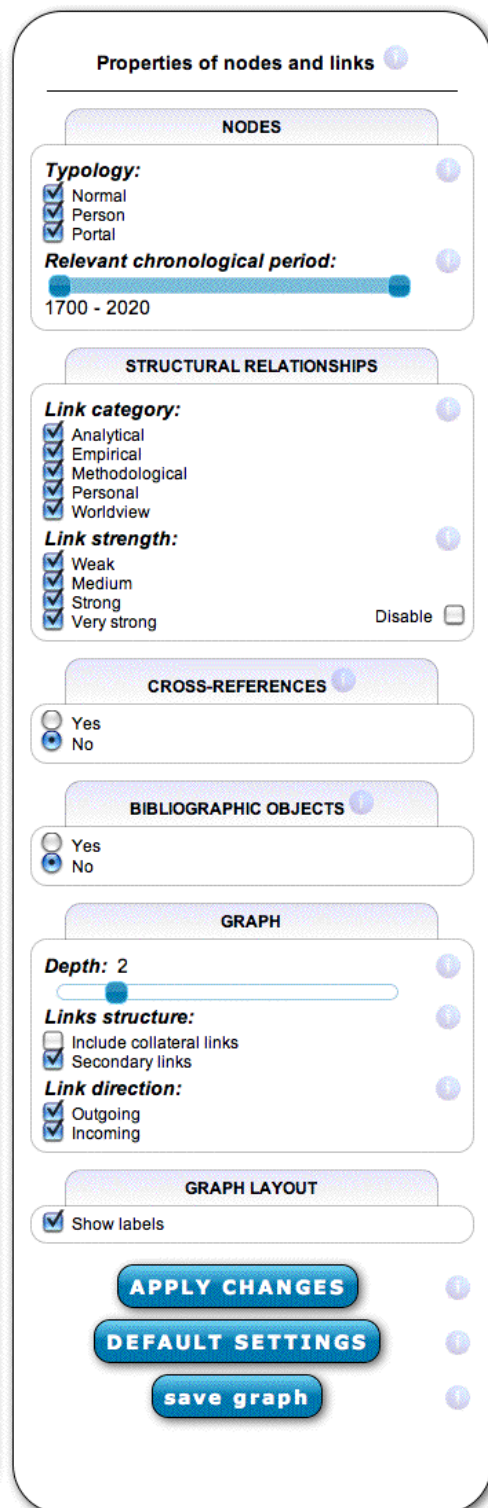
The purpose of this innovation is not only to allow for pluralism in the En-cycle-pedia, but to turn it into an active analytical instrument: the reader can compare the global consequences of switches to different epistemological views or changes of perspective on a point of detail; conversely, the reader can examine which fundamental assumptions have to be changed in order to build (or solidify) bridges between previously unconnected (or loosely linked) topics or, on the contrary, cutting existing connections that one feels should not be there. Again, if the reader thinks that some unconnected nodes should be linked, he or she can explore ways of building the necessary bridges and visualize the consequences of doing so.

By these means, the En-cycle-pedia is not a mere repository of notions or just a flexible organizer of knowledge, but becomes an analytical tool enabling the encyclopedist and readers to assess the implications of different interpretative schemes. Moreover, the tool helps to reveal gaps and inconsistencies in the planning of the list of entries, and is therefore of definite support to the work of the encyclopedist.

3.4 Searches

The graphical tool is also applied to search results. These are thus grouped in clusters of connected articles, thus forming islands of sense that can be explored separately (Figures 6 and 7).

Fig. 1: The filter system, enabling users to select the properties of nodes and links to be shown in the graphs.



The search page offers the possibility of finding the shortest directed path between two articles or any pair of items in the literature cited in the En-cycle-pedia, thereby enabling users to inquire into the connections between concepts, themes, people and the literature. It is also possible to find the common ‘ancestors’ and ‘descendants’ of any pair of articles, thereby examining whether there is a shared source or implication between concepts or people.

The filters illustrated in Fig. 1 can be applied to all these searches, thus restricting the query to specific domains, chronological periods or degree of significance.

3.5 Article attributes

The En-cycle-pedia’s articles also have attributes that can be used to filter the elements appearing in the graph. Similar to the major articles in the old *Britannica*, some articles act as portals in providing general overviews of relatively large topics and redirecting the reader to more specific articles (which form a second category) or to other portals. A third category of articles are of a bio-bibliographical character, and also act as mini-portals redirecting to the various articles discussing that person’s work. In the graph, one can select the kind of articles that should be shown.

Articles are also associated with a specific time frame where relevant, so that navigation and searches can be directed to the desired chronological period.

3.6 Limitations

While the graphical tool works as requested, and thus satisfies the purpose of exploring the features imagined for the En-cycle-pedia, it is not yet aesthetically very appealing and can be rather slow in representing large graphs. Although the logical engine determining the structural components of the graph works fairly fast and efficiently, the actual drawing of the graph takes too long to manage a large number of nodes and edges. The En-cycle-pedia's content is presently limited to about 80 articles, while one can imagine that the network's complexity would grow exponentially with the increase in the number of nodes. Nevertheless, in graphical terms the problem should affect mainly the 'central' articles, not the 'peripheral' ones (that is, those dealing with specific terms, concepts or facts); those more likely to be visited by users of an encyclopedia.

Generally, the tool is still under construction, as several details need to be sorted out,⁹ and for this reason the En-cycle-pedia will remain, for some time, accessible only by invitation (see footnote 6).

4. Conclusion

For the time being, the En-cycle-pedia is intended as a proof of concept rather than an attempt at providing the contents of a full specialized dictionary. For that to be possible, the limitations of the graphical tools must be overcome—probably by choosing different graphical software to represent the structural matrix. Meanwhile, however, the tool offers the possibility of experimenting with a flexible organization of knowledge by enabling the reader to examine encyclopedic articles in their multiple mutual relationships: as interpreted by the encyclopedist with respect to their place in the construction of the discipline (pre-analytical, methodological, analytical, empirical, personal); as reconstructed by the article's author when cross-referring to other entries; and as emerging from the literature on which it is based. The En-cycle-pedia is also an analytical instrument, as users can modify any of the above connections and explore the consequences of such changes. These features are not limited to the academic field chosen for this example, but can be applied to any other domain, as the distinction of different numbers and kinds of nodes, links and other attributes, and their corresponding labelling, is fully customizable.¹⁰

⁹ Among these, the search engine does not yet allow Boolean searches.

¹⁰ Although the tool was developed to be applied to an encyclopedic project, its educational implications are rather straightforward, as the instrument is fully open to interaction when associated with appropriate management of access. Indeed I use a simplified version in my teaching.

5. Acknowledgements

The encyclopedic project is a creation of the author. The technical realization of the En-cycle-pedia is mainly due to the work of Stefano Pettorossi, with the collaboration of team Dedalos R&D, which is financing the project and is collaborating with the SSIG School in Bellinzona, Switzerland. Naturally, the original ideas co-evolve with the actual implementation of the tool, which makes apparent new potentialities. Stefano's role thus transcends mere technicalities. I am grateful to Cécile Dangel Hagnauer and two anonymous referees for comments and suggestions on the first draft, and to Roberto Baranzini, Giorgio Colacchio and Giorgio Rampa for discussions on the project; the usual caveats apply.

6. References

- Alembert, J. Le Rond d' (1751). Discours préliminaire. In D. Diderot & J. Le Rond d'Alembert (eds), *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers*. Paris: Briasson vol. 1, pp. i–liii, transcription online at <http://encyclopedie.uchicago.edu/node/88>. English translation by R. N. Schwab as *Preliminary discourse to the Encyclopedia of Diderot*, Indianapolis: Bobbs-Merrill, 1963, online transcription at <http://quod.lib.umich.edu/d/did/>).
- Anderson, W. (1986). Encyclopedic Topologies. *MLN*, 101: 4, French Issue (September), pp. 912–929.
- Anderson, W. (1990). *Diderot's dream*. Baltimore and London: John Hopkins University Press.
- Bates, D. (2002). Cartographic aberrations: epistemology and order in the encyclopedic map. In D. Brewer and J. Candler Hayes (eds.), *Using the Encyclopédie. Ways of knowing, ways of reading*. Oxford: Voltaire Foundation, pp. 1–20.
- Blanchard, G. & Olsen, M. (2002). Le système de renvoi dans l'Encyclopédie: Une cartographie des structures de connaissances au XVIII^e siècle. *Recherches sur Diderot et sur l'Encyclopédie* 31-32, April, pp. 45–70.
- Cernuschi, A. (1996). L'arbre encyclopédique des connaissances. Figures, opérations, métamorphoses. In R. Schaer (ed.), *Tous les savoirs du monde. Encyclopédies et bibliothèques, de Sumer au XXI^e siècle*. Paris: Bibliothèque Nationale de France/Flammarion, pp. 377–382.
- Chauderlot, F.-S. (2002) Encyclopédismes d'hier et d'aujourd'hui: information ou pensée? Une lecture de l'*Encyclopédie* à la Deleuze. In D. Brewer & J. Candler Hayes (eds.), *Using the Encyclopédie. Ways of knowing, ways of reading*. Oxford: Voltaire Foundation, pp. 37–62.
- Collson, R. L. & Preece, W. E. [2013]. Encyclopedia. In *Encyclopædia Britannica*.

- Encyclopædia Britannica Online Academic Edition*. Encyclopædia Britannica Inc., 2013. Accessed on 19 Mar. 2013 at <http://www.britannica.com/EBchecked/topic/186603/encyclopaedia>
- Diderot, D. (1755). Encyclopédie. In D. Diderot & J. Le Rond d'Alembert (eds), *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers*. Paris: Briasson vol. 5, pp. 635–649 (English translation at <http://quod.lib.umich.edu/d/did/>).
- Eco, U. (1984). *Semiotics and the philosophy of language*. Bloomington: Indiana University Press.
- [Einaudi, G.] (1977). *Premessa dell'Editore* In R. Romano (ed.), *Enciclopedia*. Torino: Einaudi, Vol. I, pp. xiii–xix.
- Gabler Wirtschaftslexikon online, accessed 12 August 2013: <http://wirtschaftslexikon.gabler.de>
- Le Ru, V. (2002). Un exemple d'utilisation du CD-rom de Redon ou comment faire mouche dans la toile des renvois. *Recherches sur Diderot et sur l'Encyclopédie* 31-32, April, pp. 169–176.
- Quaggiotto, M. (2011). Images of knowledge. Interfaces for knowledge access in an epistemic transition. In *Proc. of the First International Workshop on Knowledge Federation*, Dubrovnik, Croatia, October 2010. CEUR Workshop Proceedings Vol. 822. Accessed <http://ceur-ws.org/Vol-822/MQ.pdf>.
- Rey, A. [2013]. Encyclopédie. In *Encyclopædia Universalis* [online], accessed on 31 March 2013 at <http://www.universalis-edu.com/encyclopedie/encyclopedie>.
- Roth, [Ch. F. W.] (1769). *Essai d'une distribution généalogique des sciences et des arts principaux selon l'explication détaillée du système des connaissances humaines dans le Discours préliminaire des éditeurs de l'Encyclopédie publiée par MM. Diderot et M. d'Alembert, à Paris, en 1751, réduit en cette forme pour découvrir la connoissance humaine d'un coup d'oeil*, Weimar: E.L. Hoffmann. (Accessed at <http://encyclopedie.uchicago.edu/content/arbre-généalogique>).
- Salsano, A. (1977). Enciclopedia. In R. Romano (ed.), *Enciclopedia*. Torino: Einaudi, vol. 1, pp. 3–76.
- Van Ewijk, P. (2011). Encyclopedia, network, hypertext, database: The continuing relevance of encyclopedic narrative and encyclopedic novel as generic designations. *Genre* 44:2, pp. 205–22.
- Weigel, S. (2013). Epistemology of Wandering, Tree and Taxonomy. The system figuré in Warburg's Mnemosyne project within the history of cartographic and encyclopaedic knowledge. *Images Re-vues*, hors-série 4 (2013).
- Wikipedia 2013, 'Encyclopedia', <http://en.wikipedia.org/wiki/Encyclopedia>, accessed 8 August 2013, 15:57.
- Yeo, R. (1991). Reading encyclopedias: science and the organization of knowledge in

British dictionaries of arts and sciences, 1730–1850. *Isis*, 82: 1, March, pp. 24–49.

Zimmer, M. (2009). *Revois* of the past, present and future: hyperlinks and the structuring of knowledge from the *Encyclopédie* to Web 2.0. *New Media & Society* 11(1–2), pp. 95–114.

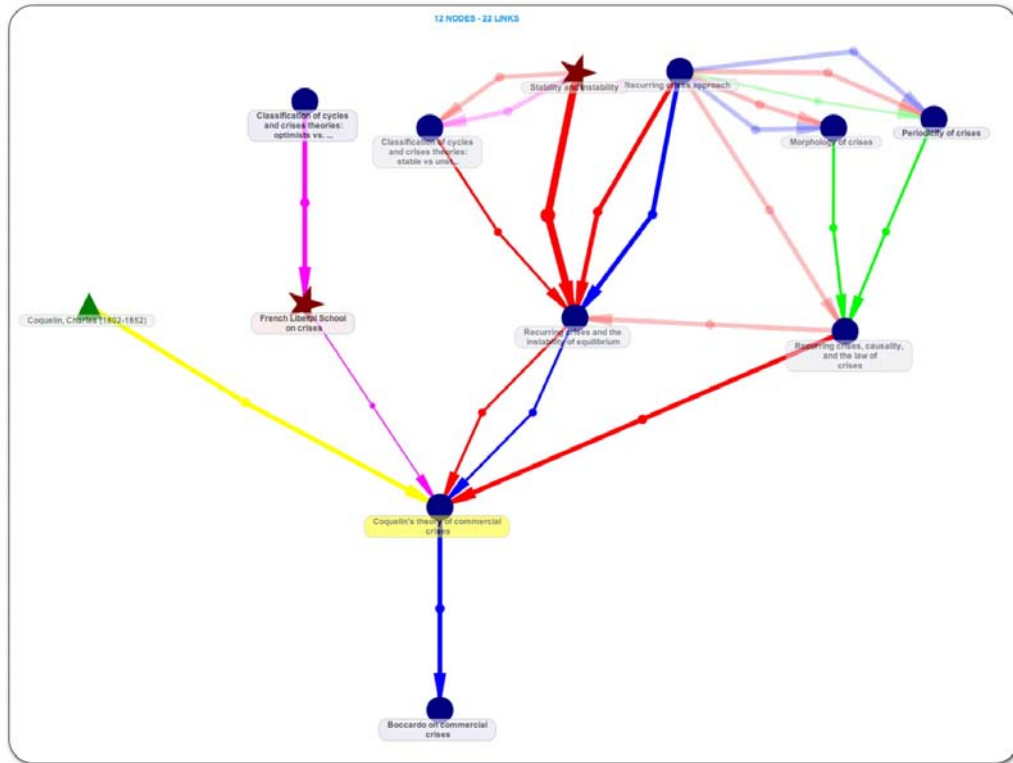


Figure 2: Graph relating to article on “Coquelin’s theory of commercial crises”, at depth = 2. Arrows of different colours represent various kinds of relationships (analytical, methodological, personal, etc.), while strength is represented by different thickness. Triangular nodes indicate bio-bibliographical entries, stars represent portal entries, while the ordinary articles are represented by circles.

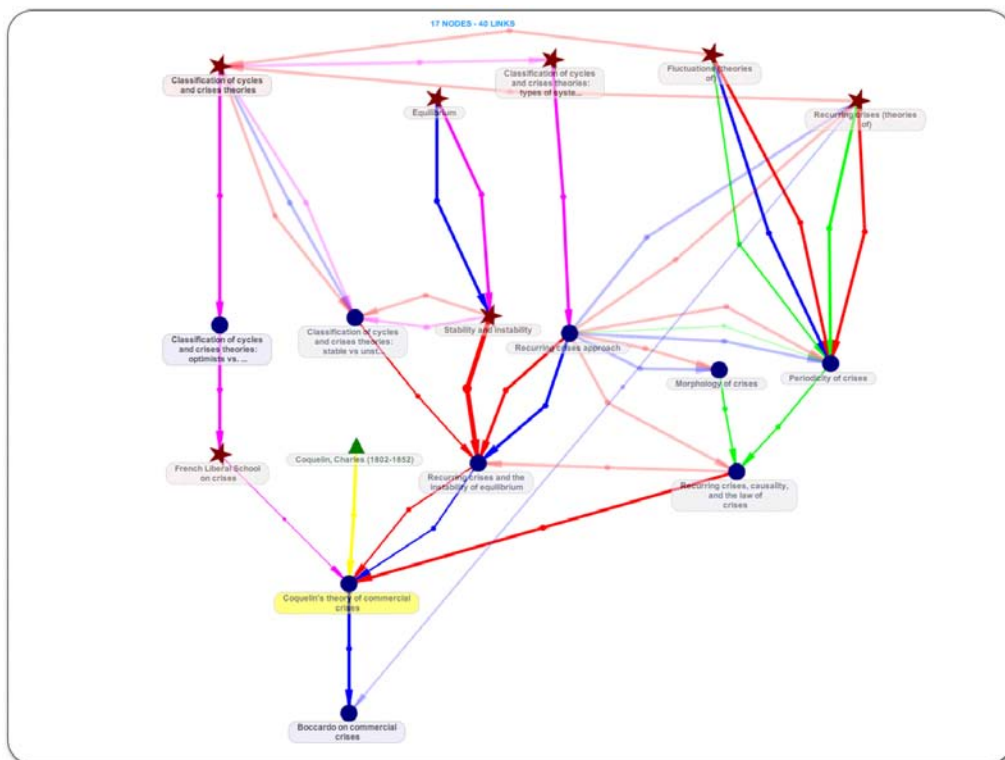


Figure 3: Third order structural links. Here, as in all graphs, any of the filters indicated in Figure 1 can be applied.

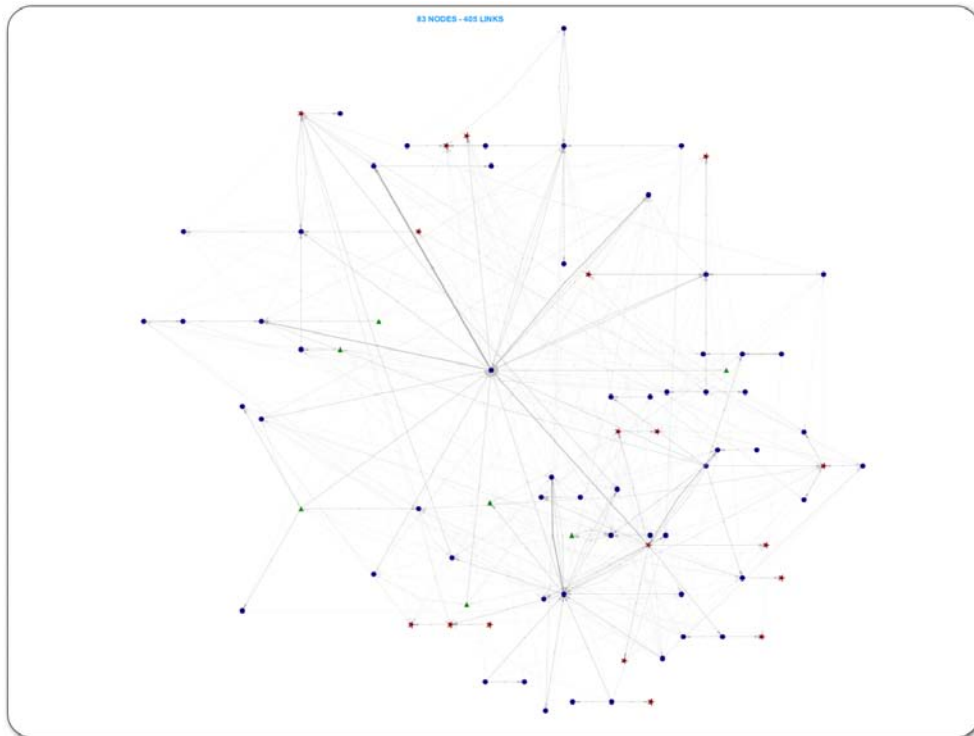


Figure 4: Graph representing the 3rd degree of cross-references relating to the same article on “Coquelin’s theory of commercial crises”. Line thickness is proportionate to the number of links to the same article.

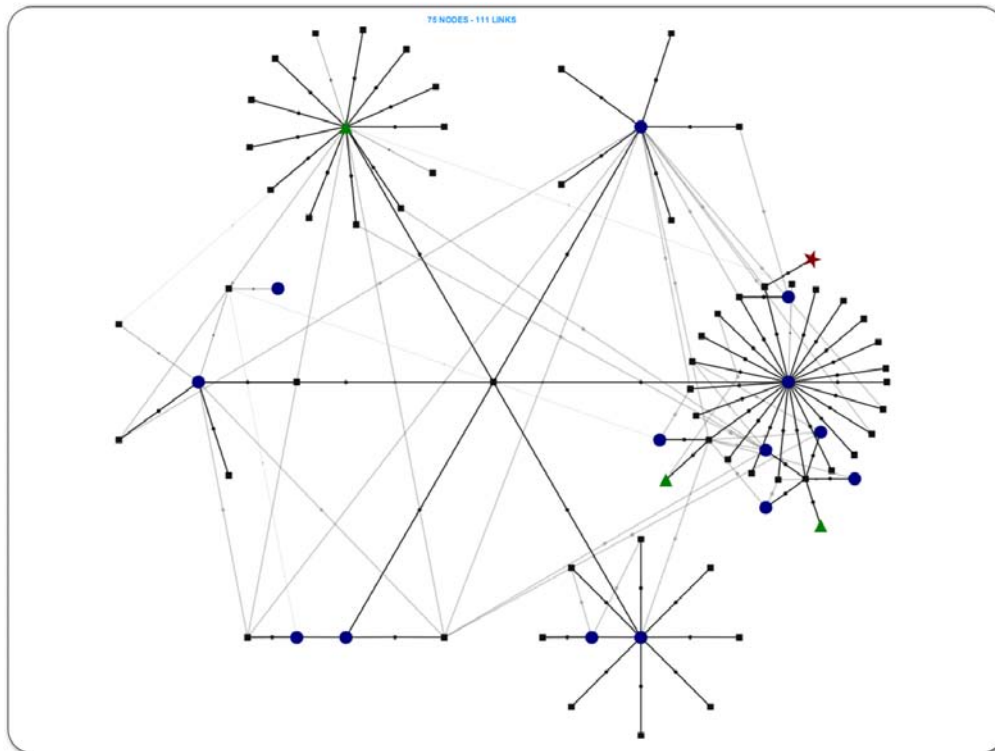


Figure 5: Graph showing the bibliographic relationships within the En-cycle-pedia: the literature item at the center of the graph is cited in the 6 articles connected with black lines; the remaining black square nodes represent the references cited by these articles, each of which is again connected to the articles citing it. The depth can be adjusted at will.

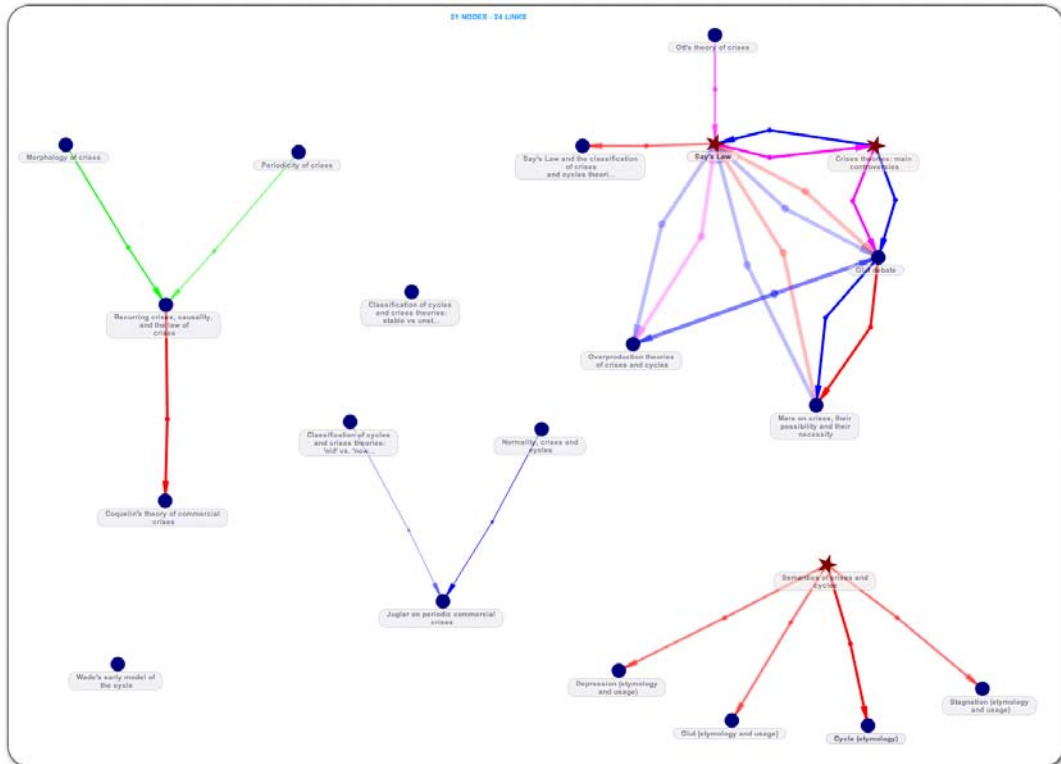


Figure 6: Graph representing a search result. Articles containing the term 'Glut' are clustered according to the existing structural links connecting them.

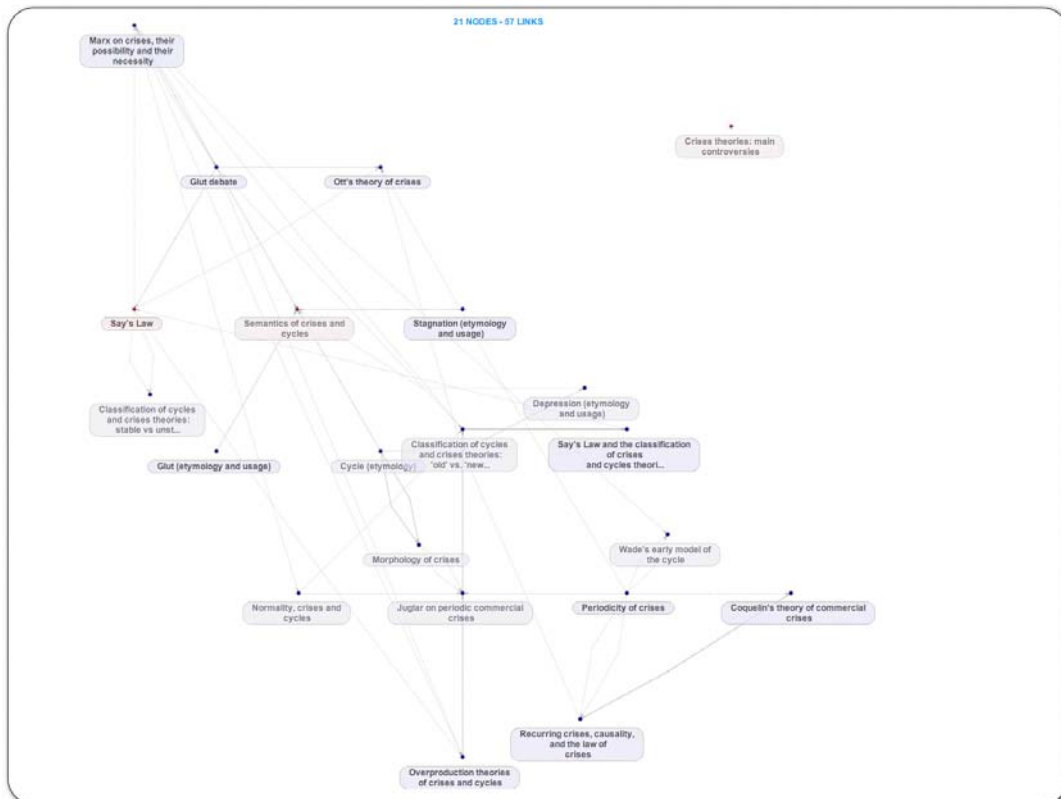


Figure 7: The same search results as in Figure 6, but clustered according to their mutual cross-references.

How preferred are preferred terms?

Gintare Grigonyte¹, Simon Clematide², Fabio Rinaldi²

¹Computational Linguistics Group, Department of Linguistics, Stockholm University
Universitetsvagen 10 C SE-106 91 Stockholm, Sweden

²Institute of Computational Linguistics, University of Zurich
Binzmuhlestrasse 14, CH-8050 Zurich, Switzerland

E-mail: gintare@ling.su.se, siclemat@cl.uzh.ch, rinaldi@cl.uzh.ch

Abstract

We present a novel approach for synonymous term preference detection that relies on chronological text analysis. Our approach analyses the use of synonymous term entries in a chronological reference corpus. As a result of preference evaluation, a ranking of preference between all the synonymous term entries belonging to the same concept is established.

Keywords: automatic terminology curation; synonymous terms; term preference; chronological corpus.

1. Introduction

This article discusses the problem of automatically determining preferred terms in terminological databases. The notion of a preferred term becomes important for automatic domain text processing. We have experimented with biomedical terminology; however the approach presented in this paper can be extended to other domains and terminologies.

Terminological entries in databases like Unified Medical Language System (UMLS) contain manually assigned tags denoting which synonym among all listed synonyms is the preferred one.

To illustrate the impact of the UMLS, consider the largest database of biomedical domain literature PubMed. PubMed publishes more than 500,000 documents each year and its publications are indexed with UMLS terms.

The UMLS (Bodenreider, 2004) is a human-expert curated terminological resource that has the following micro-structure:

ConceptID

Synonym 1

Synonym 2... PreferredTerm

Synonym n

The conceptID is a conceptual identifier for all subsumed terms. The conceptual identifier is similar to a synset identifier in WordNet. Just like a synset contains synonymous interchangeable expressions, so a concept in the UMLS also has synonymous terms. The preferred term tag is reviewed periodically and assigned manually by domain experts who curate terminological entries.

Domain terminology is extremely responsive to changes and new developments inside the respective domain, which motivates the development of automatic approaches for terminology maintenance. We view term preference in domain texts as a usage-based, and thus dynamic, phenomenon. An automatic preference detection is important if we want to take into account how terms are actually used in domain literature.

2. Data and tools

We used a subset of the UMLS terminology covering the topic of diseases. This subset contains over 90,000 concepts. The total number of terms is over 500,000. As a chronological reference corpus to study the usage of domain terms, we used all publications of the PubMed¹ January 2012 release. The 2012 PubMed dataset release contains over 22 million documents consisting of titles and some abstracts between 1881 and 2012.

In order to consistently detect occurrence of terminology in the PubMed2012 corpus we have used a specialized tool MetaMap², developed by the National Library of Medicine, which identifies biomedical concepts from unstructured texts and maps them into concepts from the UMLS (Pratt and Yetisgen-Yildiz, 2003).

3. Possible approaches

A terminological concept in UMLS contains multiple synonyms expressing the same concept and one of those synonyms is marked as a *preferred term*. For instance, the *C0008049* concept in UMLS has 16 synonyms, of which one is marked as preferred: ‘*varicella infection*’.

This paper proposes a corpus-based approach for automatically detecting preference among synonymous terms in terminologies such as UMLS. We see term preference as a usage related, dynamic phenomenon. The simplest way of automatically measuring term preference is counting the number of occurrences in a reference corpus:

¹ <http://www.ncbi.nlm.nih.gov/pubmed>

² <http://metamap.nlm.nih.gov>

chickenpox varicella	11
varicella infection	346
varicella	3820
chicken pox	1767

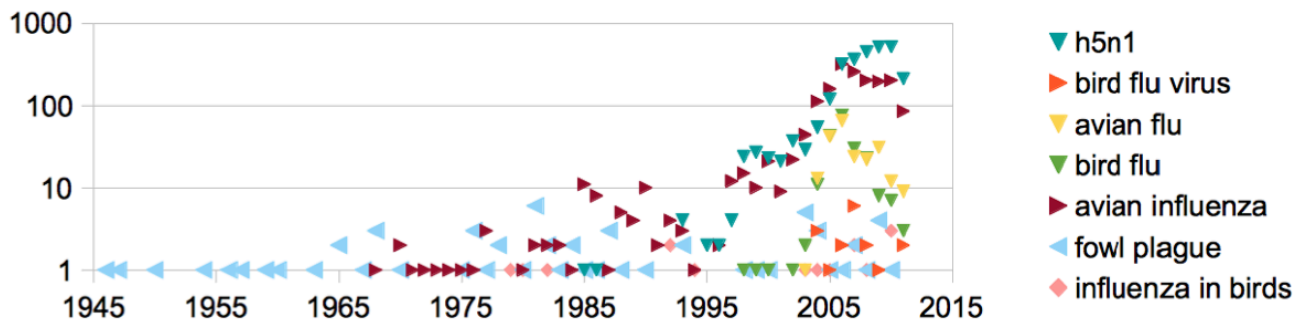


Figure 1: Chronological occurrences of synonyms of the concept 'COO16627'.

However, in case of recently emerged and topical terms, like '*h5n1*' in the concept *COO16627* (see Table 1), we find that their frequency is overwhelming and that this criteria for determining term preference might be inadequate. Thus, chronological information such as a time interval between the first and last occurrences of a term (see column 3, Table 1) or the total number of years for which a term is used in a reference corpus (see column 4, Table 1) might also constitute informative criteria of a term usage.

Taking into account time dimension alone is also insufficient, particularly if term occurrence is sparse. Besides, analyzing frequency and time data separately creates a biased view of term preference. Consider, for instance, synonyms of the concept *COO16627* (Table 1, Figure 1): '*h5n1*' is the most frequent; '*fowl plague*' is the most chronologically prominent.

Synonymous terms	# occurrences	year interval	# years
<i>h5n1</i>	2722	26	20
bird flu virus	20	9	7
avian flu	219	9	8
bird flu	206	13	13
avian influenza	1737	43	40
fowl plague	65	64	42
influenza in birds	15	31	11

Table 1: Analysis of synonyms of the concept *COO16627*.

In this paper we argue that in order to determine the preference of a term among its synonyms, time and frequency criteria should be used in combination. The simplest model that considers both dimensions is a linear regression.

4. Method

We model the series of data of the occurrence of a term over time as a simple linear regression, where α and β are unknown parameters, and ε corresponds to noise:

$$\alpha + x_i * \beta + \varepsilon_i \quad (1)$$

The fitted line is equal to the correlation between term occurrence (y_i) and time (x_i) corrected by the ratio of standard deviations of these variables. The unknown parameter β corresponds to the steepness of the slope. We use an ordinary least squares method for estimating unknown parameters α and β .

$$\hat{\beta} = \frac{Cov[x, y]}{Var[x]} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} \quad (2)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \quad (3)$$

Chronological data sparseness is a major obstacle if we want to compare all synonyms and estimate their parameters for linear regression. From Figure 1 we see that some terms occur rather consistently throughout the years, e.g. ‘fowl plague’, while other occur very rarely, e.g. ‘influenza in birds’. In order to obtain the same number of data points we included all years when at least one of the synonyms has occurred; also, in cases when a synonym has not occurred though other synonyms from the group have occurred during that year, we set the basic value for a non-occurring synonym to 0.1³.

We use relative frequency of occurrences normalized by the total number of occurrences within the set of synonyms occurring during a specific year.

The final ranking of term preferences is based on parameter β multiplied by two constants: 1) the total number of years that a synonym has occurred divided by the maximum number of years available from the set of synonyms; and 2) the total number of occurrences of a synonym divided by the total number of occurrences

³ Arbitrarily chosen in order to differentiate between situations: a) 0, none of the synonyms of a concept have occurred that year; and b) 0.1, a synonym has not occurred, but other synonyms from the concept have.

within the synset.

The estimated parameter β from the linear regression model based on term occurrence and time enables a ranking of different synonyms of the same concept. For instance, the term preference ranking over time for the concept *C0016627* in the PubMed corpus is:

avian influenza	0.00478
h5n1	0.00345
fowl plague	0.00204
bird flu	0.00079
avian flu	0.00024
avian flu virus	0.00019
influenza in birds	0.00018

This approach can be used for several interpretations of term evolution. The first interpretation of the β parameter is that a negative value shows a tendency toward term extinction. However, such an interpretation is only possible in the context of other synonyms of the term. This is the case because we analyze a domain specific corpus and we want to make sure not to include situations such as a temporary disappearance of a term or phenomenon inside the domain literature (e.g. no publications representing a specific disease have been registered during a certain period of time). Only when other synonyms of the same term continue to occur can we talk about extinction of that specific term. The situation of one term showing a tendency to disappear (negative β value) when its synonyms continue to be used (positive β value) is called term replacement (Grigonyte et al., 2012A, 2012B).

Second, the positive value of the β parameter shows an increase in term occurrences over time. The larger parameter means that the term is used proportionally more than its synonyms and its use is therefore increasing with time.

5. Results

We analyzed the terminology of diseases in the UMLS 2012 release. All terminological entries come under the semantic group of disorders.⁴ The set of disease terminology concepts that contain at least two synonymous terms comprises 17,410 concept entries. Each concept entry in the UMLS database has several synonymous terms. One or more of them is marked as the ‘preferred term’.

For evaluation purposes we chose the annotation of MeSH which has only one ‘preferred term’ for each concept.⁵ The test set was therefore left with 2,966 concepts

⁴ Semantic tags of disorders: T020, T190, T049, T019, T047, T050, T033, T037, T048, T191, T046, T184. For more information see: <http://semanticnetwork.nlm.nih.gov/SemGroups/>

⁵ We chose UMLS term entries that match the MeSH Descriptor record.

that have synonymous terms and one ‘preferred term’ tag.

The evaluation was performed by comparing the highest ranking synonym against the manually assigned ‘preferred term’ tag in the UMLS. We used two methods: a) the highest ranked synonymous term modelled by our approach *linreg*; and b) the most frequently occurring synonym *maxocc* (see Table 2).

# of concepts that have synonym synsets	17410	
# of synsets with MESH ‘preferred term’ tag	2966	
# of cases of ‘preferred term’ match by linreg	1805	60.86%
# of cases when a different ‘preferred term’ is suggested by linreg	1161	39.24%
# of cases of ‘preferred term’ match by maxocc	1852	62.55%
# of cases when a different ‘preferred term’ is suggested by maxocc	1114	37.45%

Table 2: Results of term preference evaluation.

Both approaches yielded very similar results. The agreement between *linreg* and *maxocc* is 88%. Around 60% of the preferred UMLS terms match with the most preferred terms used in domain corpora. However, for a substantial number of term entries both methods would also suggest other preferred terms. For instance, the concept *C0008029* has four synonyms, of which ‘*fibrous dysplasia of jaw*’ is the manually assigned preferred term. The highest ranking synonym according to *linreg* and to *maxocc* methods is ‘*cherubism*’.

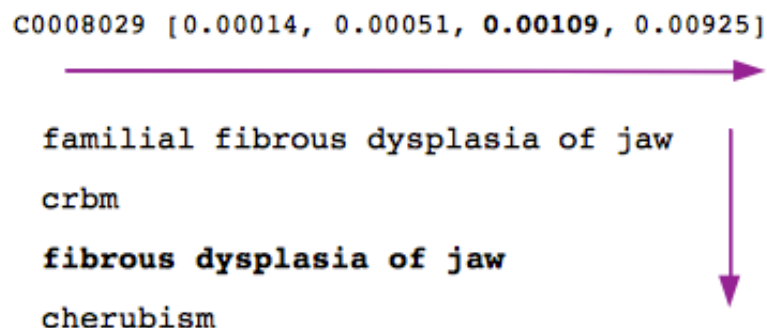


Figure 2: Synonym preference by *linreg* method.

Examples of different suggestions between *linreg* and *maxocc* are:

seasonal allergic rhinitis	hay fever
rheumatic disease	rheumatism

The large proportion of preferred terms not matching the manually assigned ‘preferred terms’ can be explained by at least two contributing factors. First, we performed the ‘hard match’ between the highest ranking term and the UMLS term, which included only exact matching strings, no orthographical deviations were allowed. Second, we only compared one preferred term from the UMLS entry instead of analyzing all preferred terms against the top preferred term suggested by the *linreg* method.

6. Conclusions

We present an approach for term preference detection that relies on term usage in the chronological reference corpus.

The *linreg* method was tested against manually assigned preferred terms. For the task of synonym preference detection the *linreg* method showed similar results to the *maxocc* method which can be partially explained by *linreg* modeling the tendency of a synonym as having increasing usage in the future. However a term preferred by the *linreg* method also indicates that it might not necessarily reflect the most frequently used term.

Lexicographers and terminologists could use the preference ranking of terms for a validation of the contents of existing term bases. As an outlook for employing the *linreg* method, a terminology expert should look at cases where the predictions and the actual preferred term are different. The method described in this paper can be used as a diagnostic tool in terminography, i.e. increases, decreases and temporary absence of term occurrences can assist an interpretation of domain terminology change.

The proposed approach could be implemented in different domains, provided that domain terminologies and large reference corpora spread over many years are available, e.g. legislative and political domains.

7. Acknowledgements

This research was supported by the Conference of Swiss University Rectors organization, Sciex research funding grant 11.002. The authors thank O. Bodenreider and anonymous reviewers for valuable comments.

8. References

- Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating Biomedical terminology. *Nucleic Acids Research*, vol. 32(1), p. 267-270.
- Grigonyte, G., Rinaldi, F., Volk, M. (2012A). Term evolution: use of biomedical terminologies. In *Proceedings of AAAI-2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*, p. 79-80.
- Grigonyte, G., Rinaldi, F., Volk, M. (2012B). Change of biomedical domain terminology over time. In: *Human Language Technologies – The Baltic Perspective*, p. 74-81.
- Pratt, W., Yetisgen-Yildiz, M. (2003) A Study of Biomedical Concept Identification: MetaMap vs. People. *AMIA Annual Symposium Proceedings*, p. 529–533.

Mapping a Traditional Dialectal Dictionary with Linked Open Data

Eveline Wandl-Vogt¹, Thierry Declerck^{1,2}

¹Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Vienna

²German Research Centre for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, Campus D3 2, D-66123 Saarbruecken
E-mail: Eveline.Wandl-Vogt@oeaw.ac.at, thierry.declerck@dfki.de

Abstract

In this paper, we present an approach for turning a traditional dialectal dictionary into a modern digitized and online linked dictionary. We describe steps that have been taken for the transformation of a former paper-based dictionary into machine-readable (semantic) web representation languages. This move raises the possibility of cross-linking dictionary data not only with other types of language resources, but also with many (scientific) domain descriptions that are already available in the Linked Data framework.

Keywords: collaborative lexicography; linked open data; dialectology; Bavarian dialects

1. Introduction

In this paper, we discuss a proposal for turning major dialect-lexicographic enterprises (also known as *Territorialwörterbücher* ‘territorial dictionaries’ or *diatopische Gebietswörterbücher* ‘diatopic area dictionaries’) of the German language (c.f. Moulin 2010: 594), on which some teams have been working for centuries, into modern digitized and online linked dictionaries. We take the Dictionary of Bavarian dialects of Austria (*Wörterbuch der bairischen Mundarten in Österreich*, WBÖ) as an example for showing aspects for this process of transformation.

The WBÖ¹ is believed to be a good example, considering that the institutional infrastructure and conceptualization for the dictionary was set up in the early 20th century. Its systematic data collection continued until the late nineties. In 1998 a rationalization concept (“*Straffungskonzept*”) was issued, with the goals of finalizing the systematic data collection, shortening the dictionary content, fastening the dictionary compilation, and targeting the linking of the dictionary with a (digital) data corpus. Results of the work are being made available in published volumes since 1963 (A- to E-, including P- and T- as well as compounds due to etymological

¹ See examples of entries of WBÖ in the appendix.

lemmatization rules). Digitization of the materials started in 1993 (*Datenbank der bairischen Mundarten in Österreich*, DBÖ, 1993–2007). The *Datenbank der bairischen Mundarten in Österreich electronically mapped* (dbo@ema; since 2007) includes geo-referenced linguistic data as well as lexicographic background data (such as biographies, bibliographies, location hierarchy) and interactive maps. This development enabled the publication of data on the internet immediately after editing in the data base, linked with the digital dictionary itself. With this development, interactive queries by users as well as user-friendly navigation on the basis of cartographic material, is supported (c.f. Wandl-Vogt, 2010; Wandl-Vogt & Nickel, 2011).

We recently launched the subsequent steps consisting of using standardized (semantic) web representation languages, in order to make the data machine-readable and processable. In doing so, the cross-linking of the WBÖ data is supported not only with other types of language resources, but also with many (scientific) domain descriptions that are already available in the linked data framework.² We also address the issue of collaborative approaches to the generation and use of shared dictionary data.³

This may be particularly urgent, considering that there are still projects working on endangered or minority languages with no or little support from modern (language) technologies and which therefore take a long time to produce results, and are extremely costly. Furthermore, this issue is also valid for minority language resources that were worked on before the advent of the Web, focusing here on associated possibilities to store and access collected and analyzed minority or endangered languages resources. At least, one should be able to see such results re-used profitably; i.e. quickly, reaching a larger audience or being integrated into new and different applications.

This way, minority and endangered languages gain the same digital dignity as mainstream languages, even if only a few people are using the language or if only a few documents or resources exist. If we adopt same methods of encoding linguistic descriptions as applied to mainstream languages, data quality can be the same for researchers as in the case of well-resourced languages, in spite of the missing quantity and variety of sources, which is very important for statistical studies and the detection and marking of variants.

To ensure interoperability of our data with other language data, their transformation into a description standard, such as ISO-LMF⁴ or TEI,⁵ is required. Further, it is

² See <http://linkeddata.org/> for more details. Many National Libraries already publish their data within this framework.

³ The most striking example of such a collaborative approach in the dictionary field is Wiktionary: <http://www.wiktionary.org/>

⁴ LMF (Lexical Markup Framework) is a standard for encoding lexical resources, resulting

necessary to encode the language data in a semantic web standard, such as RDF,⁶ SKOS⁷ and SKOS-XL,⁸ to make the data machine-readable and interoperable in web applications.

2. Harmonization of linguistic Information included in WBÖ

Before linking the language data provided by WBÖ – as well as its metadata – to other (linguistic) data, there is a need for a detailed analysis and harmonization of the given dictionary data, in order to ease their cross-linking and make use of the cross-linking potentials (Wandl-Vogt, 2005). The language data to which WBÖ is being linked can consist of entries in (dialect) dictionaries, multilingual semantic networks,⁹ labels and comments in (multilingual) domain thesauri,¹⁰ or language data available in online resources, such as knowledge resources available in the linked data infrastructure¹¹.

It is important to have a clear picture of what linguistic information those language data contain: Does a (dialect) dictionary list as its entries lemmas, base forms or full forms? Do the entries contain synonyms, translations? Are the entries associated with morphological or syntactic information? Are the dictionary entries listing (different) meanings of the words? Metadata describing those fields are important, and we have started to port our dictionary component elements into the TEI standardized representation format for textual documents.

As shown in Table 1 of the Appendix, an entry in WBÖ typically consists of a lemma (*Puss*), morpho-syntactic information (*M., jedoch meist neutr. Dem.*), meanings (*Kuß, Gebäck, PflN*), location (*s-, mbair. m. SI, Egerl, Simmersdf. Igl.*), etymological information (*Schallw. ...*), and references to neighbouring German dialectal dictionaries (*Bayr. Wb. 1,295, Schwäb. Wb. 1,1558*).

If we now consider knowledge organization systems, such as thesauri, taxonomies or

from the cooperation between experts in dictionaries and computational lexicons. See: http://en.wikipedia.org/wiki/Lexical_Markup_Framework

⁵ TEI stands for “Text Encoding Initiative”, see http://en.wikipedia.org/wiki/Text_Encoding_Initiative. See <http://www.tei-c.org/index.xml>

⁶ See <http://www.w3.org/RDF>

⁷ See <http://www.w3.org/TR/skos-primer/>

⁸ See <http://www.w3.org/TR/skos-reference/skos-xl.html>

⁹ The Multilingual extension of WordNet is such an example (<http://www.globalwordnet.org/>).

¹⁰ A good example of a thesaurus available with labels in more than 30 languages is GEMET. (<http://www.eionet.europa.eu/gemet/>).

¹¹ A prominent source of such data in the Linked Data framework (<http://linkeddata.org/>) is DBpedia (<http://dbpedia.org/About>), with a lot of multilingual labels associated to both very generic and specific concepts.

ontologies, we can see that some of those systems contain labels, comments and/or definitions as annotation properties. Such annotation properties are making use of natural language expressions for naming and describing the elements of the knowledge organization system. Do such annotation properties contain precise terms? Do they include linguistic information? If not, they should be lexicalized by applying NLP tools, adding lemma, morphological and syntactic information to the annotation. The output of the lexicalization process should be compatible with lexical and linguistic information available in the (dialect) dictionaries. This way, lexicalization supports the disambiguated mapping of words used in a label or in a definition of a knowledge source to an entry of a (multilingual) semantic network or of a (dialect) dictionary.

The model we adopt for the representation of the results of lexicalized labels is the one described by *lemon*,¹² developed in the context of the Monnet project.¹³ *Lemon* is also available as an ontology.¹⁴

A special focus in our work lies thus in achieving harmonization of all language data included in the various types of sources we are dealing with. We propose to use the ISO LMF standard for encoding all information about the organization of the lexicon or dictionary, whereas for detailed linguistic information, such as the morphology of dictionary entries, we point to the ISO data categories¹⁵ for ensuring interoperability of linguistically relevant tag sets.

In doing so, we obtain a clear view of the commonality between the language data we are working with, so that they can be optimally used in the tasks combining (dialect) language resources with other language resources or with knowledge objects.

3. Transformation of WBÖ into a machine-readable Format

While encoding in LMF and the use of Data Categories are important for getting information about the content of dictionaries and other sources, this does not say anything about the formal representation of such data. LMF is represented as UML diagram, and the correspondingly marked language data can be serialized in XML or RDF. We need to know more about the adequate formal representation of language data if we want to achieve our goal, which is to publish the dictionaries in the Linked (Open) Data framework. We need to make our language data machine-readable and interoperable in a web context. And for this, there is a need to adopt a representation format that can model the interoperability between the information we can find in

¹² *lemon* stands for “Lexicon Model for Ontologies”. See <http://lemon-model.net/> and McCrae et al. (2012)

¹³ See www.monnet-project.eu

¹⁴ See <http://www.monnet-project.eu/lemon>

¹⁵ See www.isocat.org for more details.

very different objects: lexicon entries, taxonomy labels, knowledge objects, etc.

This format should support easy publishing and access on the Web. Therefore we opt for a combination of RDF(s) and SKOS. A first experiment of porting entries of WBÖ to SKOS and *lemon* has been performed and examples of the actual state are presented below.

An additional advantage of using RDF(s) and SKOS is the fact that it allows us to access details of the modelling of the language data, using for this the SKOS-XL extension and the *lemon* model for lexicons in ontologies. This also enables the representation of information about morphological composition, variants, collecting circumstances or methods, etc., which we can conclude from the LMF encoding of the lexical sources.

Our actual realization of WBÖ in SKOS and *lemon*¹⁶ consist in creating a SKOS concept scheme in which each entry of WBÖ is encoded as a concept belonging to it:¹⁷

```
@prefix wboe: <http://www.oeaw.ac.at/wboe/#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .
...
@base <http://www.oeaw.ac.at/wboe/> .

wboe: rdf:type owl:Ontology ;

owl:imports
<http://www.lemon-model.net/lemon> ,
<http://www.w3.org/2004/02/skos/core> ,
<http://www.w3.org/2008/05/skos-xl> .

wboe:ConceptScheme
  rdf:type skos:ConceptScheme .
wboe:Descriptor
  rdf:type owl:Class ;
  rdfs:isDefinedBy wboe:wboe_defs.rdf> ;
  rdfs:label "Descriptor"@en ;
  rdfs:subClassOf skos:Concept , owl:Thing ;
  skos:definition "Descriptors of the WBÖ dictionary"@en ;
  skos:inScheme wboe:ConceptScheme .
```

Above, the reader can see the declaration part of the knowledge organization system

¹⁶ For modeling, we use the Protégé ontology editor, version 4.3. See <http://protege.stanford.edu/> for more details. The examples we show in the following are in the turtle syntax (<http://www.w3.org/TeamSubmission/turtle/>), which is an export format supported by Protégé.

¹⁷ Only the base URI <http://www.oeaw.ac.at/wboe> is for now accessible from outside the institute. Expansions of this URL given below are not yet accessible.

we created for representing the WBÖ in SKOS and RDF(s). Each entry of the original WBÖ is represented as a “concept” belonging to the “wboe” concept scheme, as can be seen in the following:

```
<http://www.oeaw.ac.at/wboe/concept/59600>
  rdf:type wboe:Descriptor ,
  owl:NamedIndividual ;
  rdfs:label „Pusselein"@bar ;
  skos:inScheme wboe:ConceptScheme .
```

The SKOS code above states that there is a concept called “59600”, which is the ID of the entry in the online version of WBÖ, as well as the ID in the dbo@ema (see <http://hw.oeaw.ac.at/wboe/59600.xml?frames=yes> and <http://wboe.oeaw.ac.at/dboe/lemma/59600>). This concept in our SKOS scheme points to a “term” object:

```
<http://www.oeaw.ac.at/wboe/59600-bar>
  rdf:type wboe:Term ,
  owl:NamedIndividual ;
```

This term object represents the concrete entry in WBÖ, which is specified as being a “Bavarian” term (with the ISO code “bar”). It is this term object that carries the preferred label and the list of alternative labels. The preferred label is encoded this way:

```
skosxl:prefLabel
  <http://www.oeaw.ac.at/wboe/59600.1-bar> ;
```

It is important to note that the range of the “prefLabel” is an object in the knowledge system and not just to a string.

This object is encoded in the following way:

```
<http://www.oeaw.ac.at/wboe/59600.1-bar>
  rdf:type wboe:prefLabel,
  owl:NamedIndividual ;
  skosxl:literalForm "Pusselein"@bar ;
  skos:inScheme wboe:ConceptScheme .
```

With this we make it clear that the dictionary entry represented by the “prefLabel” is a complex entity, and not just a string, which is introduced in the SKOS modelling by the property “literalForm”. The term object can also bear a list of related alternative labels, encoded as “altLabel”:

```
<http://www.oeaw.ac.at/wboe/term/59600.1-de>
  rdf:type wboe:altLabel ;
  skos:inScheme wboe:ConceptScheme ;
  skosxl:literalForm „Kuss"@de .
```

Here the reader can see that the alternative label is directly associated with a (German) string. This is reflecting the fact that alternative labels do not point to entries in the WBÖ, and therefore not encoded as complex term objects, contrary to

the preferred labels. Below, we provide a simplified form of the list of alternative labels that can be derived from the WBÖ entry for the word “*Pusselein*”:

```
skosxl:altLabel
  skosxl:literalForm „Busserl"@de-at ;
skosxl:altLabel
  skosxl:literalForm „Kuss"@de ;
skosxl:altLabel
  skosxl:literalForm „süßes Gebäck"@de ;
skosxl:altLabel
  skosxl:literalForm „Gewöhnliches Gänseblümchen"@de ;
skosxl:altLabel
  skosxl:literalForm „Kriech-Hahnenfuß"@de-at ;
skosxl:altLabel
  skosxl:literalForm „Gartenranunkel"@de-at ;
skosxl:altLabel
  skosxl:literalForm „Bellis perennis"@la ;
skosxl:altLabel
  skosxl:literalForm "Ranunculus repens"@la ;
skosxl:altLabel
  skosxl:literalForm "Ranunculus asiaticus"@la ;
...
```

The reader can observe that, for the time being, we associate, to the alternative label(s) of a concept, the modern German or Latin equivalent(s) of the preferred labels (reserved for the Bavarian entries of WBÖ).

In summary, in this simplified view of an entry in the SKOS representation of the WBÖ dictionary, the reader can see that each entry of the dictionary is encoded as a concept belonging to the whole concept scheme. The number associated with each concept is the ID given to the entries in WBÖ and DBÖ. The concept itself points to term objects that bear either preferred or alternative labels in various languages.

4. Mapping WBÖ to Open Linked Data

As they appeared in the example in section 3 above, alternative labels for the concepts (entries) of the “wboe” concept scheme have just strings as values. This is due to the fact that those words, modern German or Latin equivalents of the Bavarian entries, are themselves not part of the dictionary. One should expect this: WBÖ contains only Bavarian lexical material as entries. Due to this, and to the sophisticated lemmatization rules (in the example used: *Puss* with the variant *Pusselein* for the Austrian German word *Busserl*), it would be helpful for the user if some linguistic and semantic information about the words in other languages that are associated to the Bavarian entries were provided. For this purpose, we investigate the mapping of the content of the range of altLabel properties to existing lexical and linguistic information available on the Web, and more precisely in the Linked Open Data cloud.

A first experiment has been made with the actual DBpedia instantiation of Wiktionary (Wiktionary RDF extraction 2013).¹⁸ Since in WBÖ we have the linking of

¹⁸ There, *lemon* is also used for the description of certain lexical properties.

the Bavarian word “*Pusselein*” (see the example in section 3 above) to a number of German standard words, one can link the altLabel attribute for the Bavarian word directly to the entry in the DBpedia instantiation of Wiktionary. We discuss three cases here:

1. Corresponding with the value of altLabel `Kuss`, we have the entry of DBpedia/Wiktionary:
<http://wiktionary.dbpedia.org/page/Kuss-German-Noun-1de>
 47 translations, e.g.
 en = "kiss", et = "suudlus"
2. Corresponding with the value of altLabel "Bellis perennis" (Germ "Gänseblümchen"):
<http://wiktionary.dbpedia.org/page/G%C3%A4nsebl%C3%BCmchen-German-Noun-1de>
 39 translations, e.g. en = "daisy"
3. Corresponding with the value of altLabel `süßes Gebäck`, we have two entries in DBpedia/Wiktionary:
<http://wiktionary.dbpedia.org/page/s%C3%BC%C3%9F-German-Adjective-1de>
 21 translations, e.g. en = "sweet"
<http://wiktionary.dbpedia.org/page/Geb%C3%A4ck-German-Noun-1de>
 11 translations, e.g. en = "pastry"

In these three examples, we notice a number of things. First, the links contain information about the language, the Part-of-Speech and a specific meaning (the integer number indicates one of the possible meanings). Within the page accessed by the link, this information is made explicit and can be linked to.

In the second example, the reader can see that for the Latin word “*Bellis perennis*”, we refer to a German entry in DBpedia/Wiktionary. The fact is that this expression is used commonly in German. Since there is an entry for this compound term, we do not perform decomposition. But this can be performed additionally, and we could have a link to each of the Latin words “*bellus*” and “*perennis*”,¹⁹ similar to the third example discussed below.

In the third example (“*süßes Gebäck*”), the advantage of providing a lexicalization of the labels is clear: we find no link in DBpedia/Wiktionary with the URL ending in “*süßes Gebäck*.” Lexicalization is helpful, since it informs us that we have two tokens in the label, and provides the lemmas of each token. We can thus point to the two URLs in DBpedia/Wiktionary. In our SKOS modelling we use for this purpose the

¹⁹ Both entries are included in the English Wiktionary (<http://en.wiktionary.org/wiki/bellus> <http://en.wiktionary.org/wiki/perennis>)

lemon property “decomposition”:

```

<http://www.oeaw.ac.at/wboe/59600.2-de>
  rdf:type
    wboe:altLabel ,
    owl:NamedIndividual ;
  <http://www.lemon-model.net/lemon#decomposition>
    <http://wiktionary.dbpedia.org/page/s%C3%BC%C3%9F-German-Adjecti
ve-1de> ,
    <http://wiktionary.dbpedia.org/page/Geb%C3%A4ck-German-Noun-1de>
    ;
  skos:inScheme wboe:ConceptScheme .

```

In this simplified view, the reader can see how the decomposition of the content of the label can be explicitly represented, and how each component can be linked to a lexical entry (with the corresponding meaning) in the Linked Open Data cloud (in this case, the DBpedia instantiation of Wiktionary).

For each of the cases above, we have been adding a number of available translations. In the DBpedia/Wiktionary entries, very often the property “hasTranslation” is added to an entry, with a varying number of translations for different entries. By transitivity, we can add to the Bavarian entries all those translations available for the German alternative labels.

We see an advantage for the lexicographers in using such an approach by the fact that they can concentrate on the lexical entries in one language and are not required to encode related information in their own dictionary or lexicon, but can link to existing resources.

5. Conclusion

In conclusion, we can assume that the discussed proposal can aid in complex lexicographic processes of encyclopedic dictionaries on the Web. The lexicographers can concentrate on the specific data on which they are working, and link to resources in the LOD for additional information. Linking to Wiktionary-like resources is not the only way to go. In a next step, we will link to language data available in domain descriptions available in the LOD, thus mapping indirectly to expert knowledge in fields other than lexicon and linguistics. We intend to test the approach in the field of botany.

We encourage lexicographers to work together to store their data in appropriate formats in order to allow cross-linking and merging of data. This can also contribute to maintaining the availability and accessibility of these precious sources for future generations.

6. References

- Declerck, T., Lendvai, P., Mörth.K. (2013) Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data. In Francopoulo, G. (ed) *LMF Lexical Markup Framework*. Wiley 2013.
- Francopoulo, G. (2013) *LMF -- Lexical Markup Framework*. Wiley.
- Gennari, J., Ferguson, R., Grosso, W. E., Crubezy, M., Eriksson, H. , Noy, N. F. , Tu, S. W., Musen, M. A. (2002) *The Evolution of Protégé: An Environment for Knowledge-Based Systems Development*
- McCrae, J., Aguado-de-Cea, G., Buitelaar P., Cimiano P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2012) Interchanging lexical resources on the Semantic Web. In: *Language Resources and Evaluation*. Vol. 46, Issue 4, Springer:701-719.
- Miles, A., Matthews, B., Wilson, M. D., Brickley, D. (2005) SKOS Core: Simple Knowledge Organisation for the Web. In Proc. International Conference on Dublin Core and Metadata Applications, Madrid, Spain.
- Moulin, C. (2010) Dialect dictionaries - traditional and modern. In: Auer, P., Schmidt, J.E. (2010) (eds) *Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods*. Berlin / New York. pp: 592-612. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1).
- Romary, L. (2009) Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*. Mentis Verlag,
- Schreibman, S. (2009) The Text Encoding Initiative: An Interchange Format Once Again. *Jahrbuch für Computerphilologie 10*. Mentis Verlag.
- Straffungskonzept für das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1998). Accessed at http://www.oeaw.ac.at/dinamlex/Straffungskonzept_1998.pdf (25.5.2013).
- 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). In: *Complex 2005. Papers in computational lexicography*. Budapest: 243-254.
- Wandl-Vogt, E. (ed.) (2010) Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Teil 1: Pilze [Part 1: Mushrooms]. Vienna. Accessed at <http://wboe.oeaw.ac.at>. (25.5.2013).
- Wandl-Vogt, E. (ed.) (2010) Interaktive, georeferenzierte Bibliographie, Biographien und Belegbiethierarchien zum Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Vienna. Accessed at <http://wboe.oeaw.ac.at>. (25.5.2013).

Wandl-Vogt, E., Nickel, J. (2011) dbo@ema. Die Datenbank der bairischen Mundarten in Österreich (DBÖ) auf dem Weg ins Internet. In: Klagenfurter Beiträge zur Sprachwissenschaft: 458-471.

Wiktionary RDF extraction. Accessed at <http://dbpedia.org/Wiktionary> (25.8.2013)

Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-). Wien.
Accessed at <http://hw.oeaw.ac.at/wboe/31205.xml?frames=yes> (25.5.2013).

7. Appendix

In this appendix, we display some screen shots that display some relevant content of the WBÖ for our work. In Table 1 we display a WBÖ entry. The prefLabel in our SKOS model would be “Puss” (“Puss(e)lein” will be marked in the future as a related variant). Tables 2–4 display different meanings associated with the Bavarian entry.

Puss, Puss(e)lein

M. (jedoch meist neutr.Dem.), Kuß („Busserl“), Gebäck, PflN s-,mbair. m. SI, Egerl. nur als → (*Zwick[er]*)-, Simmersdf. Igl.; Schallw., vgl. KLUGE²⁰ 114; frühhd. *buß* M. Kuß GÖTZE Frühhd.Gl. 44; s.a. KRANZMAYER Kennw. 10; entl. ins Magy. als *puszi* Kuß u. *puszedli* Gebäck KOBILAROV-GÖTZE 355f., ins Slow. als *púšek* Kuß PLETERŠNIK 2,366 u. ins Kä.Slow. als *pushei* Kuß GUTSMANN Dt.-Wind.Wb. 261. — Bayer.Wb. 1,295, Schwäb. Wb. 1,1558.

Table 1: WBÖ 3,1515: Entry – Overview.

Bed.: 1. Kuß im gesamten Verbr.Geb. (meist als 1. od. 2. Dem.), Syn. → (*Fotz*)/*pemperer*,

Table 2: WBÖ 3,1516: Meaning 1: “kiss”

2. Kl. süßes Gebäck

m. flacher kreisförmiger Unterseite u. gewölbter Oberseite ugs. (meist 2., seltener 1. Dem.), s.a. EBNER² 51; rundes Nußgebäck auf Kirchtagen Gott.Wb. 1,91 (2.Dem.);

Table 3: WBÖ 3,1516: Meaning 2: “sweet pastry”

3. PflN:

a) f.d. Garten veredelte Art v. Gew. Gänseblümchen (→ *Bellis perennis*) BöW MARZELL PflN 1,555 (2.Dem.); *rote, weiße Busserl* dass. BöW SCHREIBER Bö.(1910) 134; — b) f.d. Garten veredelte Art v. Kriech-Hahnenfuß (→ *Ranunculus repens*): *gelbe Busserl* BöW SCHREIBER ebd. 156; s.a. MARZELL PflN 3,1278; — c) Gartenranunkel (→ *Ranunculus asiaticus*) OÖ (1893, 2.Dem.).

Table 4: WBÖ 3, 1516: Meaning 3: “plant”, e.g. “daisy”

Writing assistants and automatic lexical error correction: word combinatorics

Leo Wanner¹, Serge Verlinde², Margarita Alonso Ramos³

¹ICREA and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona (Spain)

²Leuven Language Institute, KU Leuven, Dekenstraat 6, B-3000 Leuven (Belgium)

³Faculty of Philology, Universidade da Coruña, Campus da Zapateira sn, 15071 A Coruña (Spain)

E-mail: leo.wanner@upf.edu, serge.verlinde@ilt.kuleuven.be, lxalonso@udc.es

Abstract

Genuine lexical writing assistants that attempt to detect lexical errors such as miscollocations are traditionally less common in Computer Assisted Language Learning than spell and grammar checkers. However, there is empirical evidence of the importance of capturing and correcting miscollocations in the writings of language learners, and therefore an increasing number of proposals deals with the detection of errors in collocations and the delivery of lists of correction suggestions. However, very few of these proposals take into account the varying ease with which learners can master different collocation and miscollocation types, or the fact that some collocation type errors might be more common than others, given that a writing assistant should be capable of handling at least the most common types of miscollocation. Furthermore, existing proposals explore collocation error-specific strategies, implicitly assuming that with one universal strategy all types of miscollocations can be detected and corrected. Our preliminary study, conducted on Spanish and French material, highlights one type of collocation in which learners err the most: support verb constructions (SVCs). To account for this, we explore a SVC-specific collocation error detection and correction strategy.

Keywords: CALL, collocation (error) typology, collocation identification, collocation error detection, collocation error correction

1. Introduction

Electronic lexicography supporters argue that e-lexicography needs to design new applications that take advantage of the potential offered by the electronic medium, while still drawing upon data from traditional lexicography (Gouws, 2011). From this perspective, lexical writing assistants seem to be an ideal solution. On the one hand, they imply the use of features offered only by the electronic medium (real time interaction with the user, on-the-fly error detection, correction suggestions, etc.), whereas on the other hand, they require the use of “traditional” data, i.e., lexical resources.

Genuine lexical writing assistants are much less common than spell and grammar checkers¹ (although they are increasing; see below) and are not as mature: performance tests show differences in quality between writing assistants that focus on lexical errors; many of them achieve only a limited rates of successful error recognition and/or correction. However, this is not only due to the immaturity of the technologies. In addition, lexical errors are very heterogeneous (including, e.g., preposition use, choice of synonyms, word combinatorics etc.) are thus more difficult to capture and, at the same time, very frequent in foreign language learners' text production (e.g., Granger, 2003 for French, Alonso Ramos et al., 2010b and Agustin Llach, 2011 for Spanish).

In this paper, we address the problem of writing assistants for collocations, where one of the elements of a combination (the *base*, *B*) is chosen freely and the other (the *collocate*, *C*) is chosen idiosyncratically depending on *B*; cf., e.g., *ask a question*, *poser une question*, *hacer una pregunta*, *eine Frage stellen*. Several studies show that collocations pose major problems to language learners (see, among others, Granger, 1998; Lewis, 2000; Nesselhauf, 2004, 2005; Lesniewska, 2006). Therefore, it is not surprising that over the last decade, proposals for collocation writing assistants have been put forward. Some of them allow for a verification of the correctness of a combination introduced via an interactive interface as a collocation and suggest, in the case of a presumed erroneous collocate, a list of possible corrections. Others provide (usually lists of) suggestions for the correction of detected collocation errors in the writing of learners (see, among others, Chang et al., 2008; Park et al., 2008; Yin et al., 2008; Liu et al., 2009; Wu et al., 2010; Ferraro et al., 2011, for a variety of different proposals).

Accuracy varies between proposals with respect to both recognition of collocation errors and provision of correction suggestions. However, most proposals attempt to cover all types of collocations (at least those with the same morpho-syntactic pattern: usually, V+N or N+V collocations), applying the same error detection and correction strategy. In light of the great variety of collocations, ranging from prototypical support verb constructions such as *take [a] walk* to combinations with semantically full verbs such as *fulfill [a] condition*, this might not be the best approach. Thus, on the one hand, learners might have more problems with one specific type of collocation than with others, whereas on the other hand, collocation type-specific detection and correction strategies might be more efficient than a universal strategy.

To address these questions, we conducted research on French and Spanish texts, with the goals of (i) investigating whether language learners show any preference with respect to the use of a specific type of collocation and whether any peculiarities can be observed with respect to the distribution of miscollocations in learners' writings; and

¹ Some of the spell and grammar checkers also try to detect and correct lexical errors. However, we focus here on pure lexical "checkers".

(ii) assessing to what extent a universal strategy for the automatic detection and correction of miscollocations is feasible or whether collocation type-specific strategies are more promising. The outcome of our study was that learners make more errors on support verb constructions (SVCs), which they use relatively often, than on other V+N collocations, and that a collocation error detection and correction strategy that specifically targets SVC errors is required. We propose such a strategy, which we ultimately plan to integrate into the online application *Interactive Language Toolbox* (<https://ilt.kuleuven.be/inlato/>; Verlinde and Peeters, 2012) for French, and HaRenEs (<http://patexpert-engine.upf.edu/HARenEs-devel/index.php>)² for Spanish.

In the next section, we evaluate the collocation error type distribution in a fragment of a Spanish learner corpus (showing that SVCs play an extraordinary role in learners' writings) and explore the variety of strategies that can be applied to automatic collocation error detection and correction, assessing whether all of them are equally well-suited for all types of collocation. Section 3 elaborates on a possible strategy for the detection of SVC errors and their correction. In Section 4, we show how the collocation error detection/correction functionality can be integrated into a writing assistant environment. In section 5 we draw some conclusions from our presentation and outline the directions of our future work in the area of automatic collocation error detection and correction.

2. What should a collocation-oriented writing assistant focus on?

As aforementioned, collocations are of different types and levels of complexity. They are not all likely to be of the same relevance to learners (in the sense that some may be less common than others) or to pose equal difficulties to the learners. Their varying complexity has additional consequences for the prospect of successful automatic recognition and correction in the case of erroneous use or composition: some will be easier to recognize and more accurately corrected by the given state-of-the-art techniques than others; and some will require additional techniques.

Surprisingly, very few studies in Computer Assisted Language Learning (CALL) that deal with collocations address these questions: neither from the didactic nor the computational perspective. In cases where a proposal focuses on a distinct collocation type, this is nearly always performed *ad hoc*, with no theoretical justification, while strategies for automatic detection and correction of errors of collocations do not usually cope better with any particular type. In what follows, we attempt to shed some light on these questions.

² HaRenEs stands for “Herramienta de Ayuda a la Redacción en Español: Procesamiento de Colocaciones”.

2.1 A closer look at the use of collocations

Collocations can be distinguished with respect to their syntactic patterns and semantic features. In the past, different types of typologies have been suggested; see, e.g., Hausmann (1985) and Heid (1996), who propose distinction between V+N, Adj+N, N+N, Adv+V, and Adv+Adj combinations; and Benson et al. (1997), who distinguishes between eight types of *grammatical collocations* and seven types of lexical collocations, some of them based on syntactic and some on semantic grounds. The most detailed and homogeneous typology is provided by *lexical functions* (LFs) as introduced in the Explanatory Combinatorial Lexicology (Mel'cuk, 1995). Each LF is characterized by a specific syntactic pattern and a semantic interpretation. The team led by M. Alonso Ramos of the University of La Coruña analyzed the use of collocations in a fragment of the Spanish learner corpus CEDEL2³, with the LFs as reference typology. In total, 1948 LF instances have been identified;

Of them, 1491 were correct and 457 erroneous (i.e., about 23.5% of all used collocations were wrong). Of the correct collocations, 532 (35.7%) were LFs that capture SVCs.⁴ The share of the other LFs was considerably lower: e.g., the LF 'intensity' (Magn) was used 97 times (6.55%) and the 'causation' (CausFunc) 87 times. From the 457 erroneous collocation instances, 110 (24%) were instances of the SVC-LFs; 83 were instances of the LFs 'realize' or 'fulfill' (Real); and 26 (5.69%) LFs CausFunc. The frequency of erroneous use of the other LFs oscillated between 1 and 9. It seems therefore clear that SVCs are both more used and more erred in by learners.⁵ Previous studies also show that SVCs constitute a major challenge for learners. This is plausible because SVCs tend to be idiosyncratic, i.e., language-specific and unpredictable.⁶ Thus, Nesselhauf (2004) reports an error rate of 32% of SVCs with the verb 'to make' produced by advanced learners of English with German as L1. Most mistakes were due to the inappropriate use of the verb. Bolly (2010:188) comes to a similar conclusion in a study on learners of French with Dutch and English as L1 with respect to the verb *faire* 'to make'.

³ CEDEL2 is an L1 English-L2 Spanish learner corpus under construction by Cristóbal Lozano in the framework of a bigger corpus-oriented project directed by Amaya Medikoetxea at the Universidad Autónoma de Madrid. Currently, CEDEL2 contains about 730,000 words of essays in Spanish on a predefined range of topics by native speakers of English and (to a smaller extent, for contrastive studies) by native speakers of Spanish. The level of Spanish of the authors of the essays varies from "elementary", "lower intermediate", "intermediate", and "advanced" to "very advanced". For further information on CEDEL2 see <http://www.uam.es/proyectoinv/woslac/cedel2.htm>; cf. also Lozano (2009).

⁴ For readers familiar with LFs: the LFs in question were Oper1/2/3; A detailed presentation of the LFs can be found in (Mel'cuk, 1996).

⁵ Note, however, that this does not mean that the share of erroneous SVCs in all used SVCs is bigger than in the case of other collocation types. For instance, in our study, the share of erroneous SVCs oscillated around 17%, while the share of erroneous 'fulfill' collocations (Real) in the total of the used 'fulfill' collocations was about 25%.

⁶ As shown in (Alonso Ramos et al., 2011), in a significant number of SVC errors, learners either literally translate L1 collocates into L2, or, on the contrary, attempt to avoid collocates that they perceive as a "too similar to" the L1 collocates.

The consequence we can draw from this abundance of SVCs is that the detection of SVC errors and their correction is a high priority task of collocation-oriented writing assistants across different L2s. The prominence of other types of miscollocations may depend more on L2; additional studies, such as that conducted by Alonso Ramos et al. (2010b), are needed to obtain a clearer picture in this respect.

2.2 A closer look at collocation error detection

Collocation error detection passes through collocation identification. In Computational Linguistics, collocation detection in corpora has been discussed and studied since the late eighties (cf. e.g., Choueka, 1988; Church and Hanks, 1989; Smadja, 1993). Mostly, word co-occurrence frequency-oriented metrics are used; see Pecina (2008) for an extensive list of such metrics. Wanner (2004) and Wanner et al. (2005) are among the few reports of semantic co-occurrence instead of word co-occurrence. In CALL, where the interest in collocations is considerably more recent, word co-occurrence metrics for the identification of miscollocations and collocations in a reference corpus are equally prominent (see, e.g., Yin et al., 2008; Chang et al., 2008; Liu et al., 2009; Dahlmeier and Ng, 2011; Ferraro et al., 2011). A co-occurrence (most often V+N co-occurrences) is considered a miscollocation if its frequency in a reference corpus is below a given threshold. Using this technique, Chang et al. (2008) report a precision of 97.5% for the recognition of English collocations and 90.7% for the recognition of English miscollocations from learners with Chinese as L1; Ferraro et al. (2011) report an accuracy of 90% for the recognition of Spanish miscollocations from learners with English as L1.

Obviously, this implies that correct rare (e.g., literary) collocations will be qualified as miscollocations. However, the results of collocation classification experiments suggest that this risk is likely to vary between collocation types. Thus, Moreno et al. (2013) report on a higher accuracy of the recognition of genuine SVCs (Oper1-LFs) than of other types of collocation by a Support Vector Machine (SVM) classifier when Vs are used as classification features. Our study in Section 2.1 also suggests that SVCs are common and thus the risk of interpretation of a rare SVC as a miscollocation by a frequency-based metric is reduced.⁷ Furthermore, SVCs are a type of lexical co-occurrence that tends to be included in general purpose dictionaries, such that lists of SVCs to match with (as, e.g. Shei and Pain, 2000) during the collocation error detection procedure are more likely to be retrieved for SVCs.

2.3 A closer look at collocation error correction

State-of-the-art collocation error correction strategies are more diverse than (mis)collocation recognition strategies. Some focus on L1 interference in learners (see, e.g., Chang et al., 2008 and Dahlmeier and Ng, 2011). Chang et al. (2008) first extract

⁷ This assumption requires further verification by broader empirical studies.

V+N co-occurrences from a given written text. Then, they check the extracted co-occurrences against a collocation list previously obtained from a reference corpus. Co-occurrences not found in the collocation list are variegated in that their verbal elements are substituted by all English translations of their L1 counterpart (Chinese, in this case) in an electronic dictionary. The variants are again matched against the collocation list. The resulting matching co-occurrences containing the noun of a non-matching co-occurrence are offered as correction suggestions. The Mutual Reciprocal Rank (MRR) of the correction list is reported to reach 0.66.

Dahlmeier and Ng (2011) work with *confusion sets* of semantically similar words. Given an input text in L2, they generate L1 paraphrases, which are then looked up in a large parallel corpus to obtain the most likely L2 co-occurrences. For this strategy, they report a precision of 38%.

Futagi et al. (2008) target the detection of miscollocations in learner writings, without considering the correction. Unlike the above proposals, they are not restricted to V+N co-occurrences. But similarly to Chang et al. (2008), they extract the co-occurrences from a learner text, variegated them and subsequently look up the original co-occurrence and its variants in a reference list to decide on its status. To obtain the variants, they apply spell checking, vary articles and inflections and use WordNet to retrieve synonyms of the collocate.

Wu et al. (2010) use a classifier to provide a number of collocate corrections. The classifier takes the learner sentence as lexical context. The probability predicted by the classifier for each suggestion is used to rank the suggestions. According to the evaluation included in Wu et al. (2010), an MRR of 0.518 for the first five correction suggestions has been achieved.

Liu et al. (2009) retrieve miscollocation correction suggestions from a reference corpus using three metrics: (i) mutual information (Church and Hanks, 1989), (ii) semantic similarity of an incorrect collocate to other potential collocates based on their distance in WordNet, and (iii) the membership of the incorrect collocate with a potential correct collocate in the same “collocation cluster”.⁸ A combination of (ii)+(iii) leads to the best precision achieved for the suggestion of a correction: 55.95%. A combination of (i)+(ii)+(iii) leads to the best precision, 85.71%, when a list of five possible corrections is returned.

Ferraro et al. (2011) suggest a two-stage strategy for correction of miscollocations in Spanish. The first stage is rather similar to the one proposed by Futagi et al. (2008): it retrieves the synonyms of the collocate in the miscollocation in question from a number of auxiliary resources (including thesauri, bilingual L1-L2 dictionaries, etc.) and combines them with the base of the miscollocation to candidate corrections. The

⁸ Roughly speaking, members of the same “collocation cluster” are values of the same LF.

candidate corrections that are valid collocations of Spanish are returned as correction suggestions. In the case that none are, the second stage applies a metric to retrieve correction suggestions. Three metrics have been investigated: affinity metric, lexical context metric and context feature metric. The context feature metric, which uses the contextual features of the miscollocation (tokens, PoS tags, punctuation, grammatical functions, etc.), performed best in that it achieved an MRR of the top five suggestions of 0.72.

Again, we can observe that all proposed miscollocation correction strategies are assumed to be equally valid for any type of miscollocation. This can be considered a valid assumption if we dispose of (i) a universal technique to identify the meaning intended by the learner when using the miscollocation (or, in other words, to automatically classify miscollocations in terms of a (semantically-motivated) collocation error typology as proposed by, e.g., Alonso Ramos et al., 2010b); and (ii) a universal technique to identify collocations of a specific type (LF) in a reference corpus. Since, to the best of our knowledge, no collocation error classification techniques are as yet available and, as we have seen in Section 2.2, state-of-the-art techniques cannot be used to retrieve collocations of a given type (at least not with an equal accuracy), collocation type-specific miscollocation correction techniques seem more promising. In the light of the characteristics of SVCs (see above), it is especially promising to single out SVC error correction.

3. Towards SVC error correction

In this section, we present an experimental set up of SVC error detection and correction. The setup involves the following stages:

1. Detection of binary word co-occurrences that are potential SVCs.
2. Assessment of their correctness.
3. In case of being judged incorrect, suggestion of a ranked list of corrections.

Each of these stages shall now be discussed in turn.

3.1 Detection of SVC candidates

Since SVCs are verb + object co-occurrences, the most reliable way to obtain candidate SVCs is dependency parsing. However, off-the-shelf parsers tend not to perform well on non-native texts; see, e.g. Heift and Schulze (2007); Krivanek and Meurers (2011). Therefore, many authors use simpler and more reliable (although more approximate) approaches. For instance, Wanner et al. (2005) use a chunker, while Yin et al. (2008), Chang et al. (2008), Ferraro et al. (2011) and others extract N+V co-occurrences identified within a sequence of words of a specific length, i.e., PoS tags. In our preliminary experiments, we also use only PoS. Obviously, this

low-tech practice can (and should) be improved to obtain optimal candidates as it collects both collocations and free word combinations. Without any analysis, the subject/object relation between noun and verb also remains unclear. However, this is a fast and quite robust approach, the quality of which is sufficient for our first round of experiments.

3.2 Assessment of the correctness of a candidate

As discussed in subsection 2.2, assessment of the collocation status of an extracted word combination and examination of the correctness of a collocation candidate can be done in one stage, using the same technique. For SVCs, two techniques seem most straightforward. The first is to match a given candidate co-occurrence with collocation lists compiled from existing (collocation) dictionaries (see the Introduction). Thus, for French, data from Fontenelle (1997) and *Dafles* (Selva, Verlinde and Binon, 2002) can be exploited; for our experiment, we compiled a matrix of a non-exhaustive list of 233 support verbs, combined with 673 different nominal bases. For Spanish, DICE (Alonso Ramos, 2004; Alonso Ramos et al., 2010a) currently contains 21,324 collocations, a significant part of which are SVCs. With extensive collocation lists at hand, a very high accuracy of collocation error recognition can be achieved.

The second technique is to draw on the distribution of SVCs in corpora. Thus, since we can assume that SVCs are used considerably more often than other types of collocations (see also subsection 2.1), a simple frequency-based technique is likely to suffice: a V+N co-occurrence whose context of use shows significant similarity with the average context of an SVC, but whose frequency is significantly below the average frequency of known SVCs, can be assumed to be an SVC miscollocation.

3.3 Correction of collocation errors

In order to find the most relevant suggestions for incorrect collocates, possible candidates have to be selected and ordered according to specific criteria, before they are presented to the user. Subsequently, either a list of possible corrections or the most relevant (or plausible) correction can be offered. As mentioned above, the limited accuracy of the state-of-the-art collocation correctors suggests caution and provides (ranked) correction candidate lists from which the user can choose.

Due to the observed distribution of SVCs, we can assume that a given noun is more likely to co-occur in a reference corpus with its support verbs (forming SVCs) than with any other verbs. As a consequence, we can retrieve the most likely (or most prominent) verbal co-occurrences as correction suggestions for the noun in question. In the context of our experiments on French, we explored some standard likeliness measures: frequency, an association measure (Z-score) and the product of both

metrics.⁹ According to the MRR of the top five suggestions for the 673 nominal bases we analyzed, Z-score and the product of this association measure with frequency lead to the best results: 0.87 and 0.88 of MRR. Both measures are superior to simple frequency, which seems to be used, e.g., by the MUST collocation checker¹⁰ for ranking their correction suggestions, because they give less weight to very frequent verbs (*avoir, être, faire*). They are comparable to the performance of the ranking metrics used in the *Just The Word* (jtw) collocation checker.¹¹

Once the list of possible corrections of a miscollocation has been determined and ranked, we need to decide how many candidates should be proposed to the user, i.e., from which rank do we believe the uncertainty of the proposed correction to still be appropriate; and indeed an SVC is too high. The graph in Figure 1 provides some evidence on this.

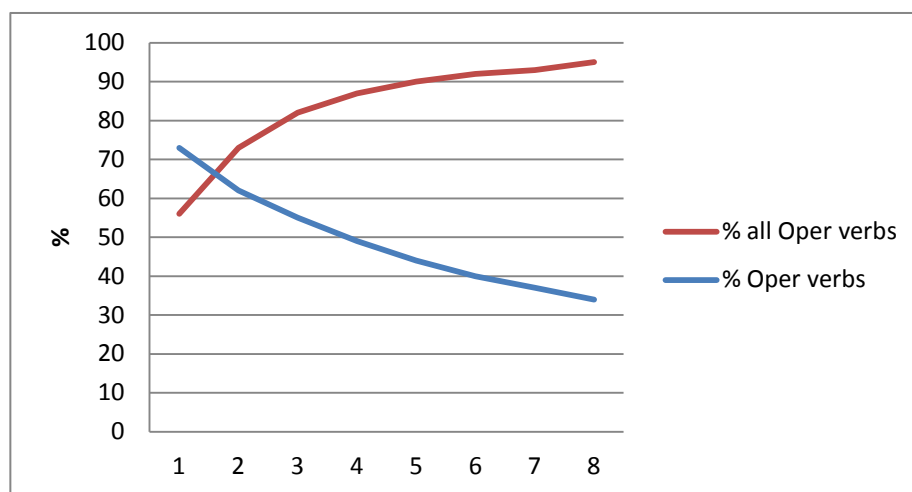


Figure 1: The quality of the SV correction suggestions, depending on the number of presented suggestions

The red line indicates the percentage of support verbs (SVs) that are available for a given base (noun) in the set of suggestions offered to the user, depending on the size of the set. It shows that nearly all SVs are contained in the first eight correction suggestions. The blue line indicates the percentage of SVs in the set of correction suggestions, again depending on the size of the set. It shows that in our experiment on French the first suggested collocate was indeed a support verb for 73% of the bases considered.

⁹ As argued previously, our experience is that purely frequency-based measures tend to perform well for SVCs since SVCs are rather common.

¹⁰ <http://miscollocation.appspot.com/>; <http://candle.fl.nthu.edu.tw:9000/>

¹¹ <http://www.just-the-word.com/>. Consider, for illustration, the ranking of the correction suggestions provided by *jtw* for *make a walk*:
<http://www.just-the-word.com/main.pl?word=make+a+walk&alternatives=alternatives&db=thesaurus>

The more suggestions, however, the greater the number of non-relevant verbs: with eight suggestions, only 35% are SVs. At the same time, more different support verbs are displayed if the number of suggestions increases, leading to a better coverage of all uses of SVs (or Oper-LFs).

In general, each application will need to decide on where to draw the line and how many correction suggestions to show. What is important is that the decision be informed.

4. Integrated online writing assistants

The programs for collocation error detection and correction (be they collocation type-specific or generic) can be used either as collocation checker demons, integrated into an editor and switched on or off by the user as deemed appropriate, or integrated into an online writing assistant; see, e.g., StringNet,¹² MUST or Just The Word. It is the latter option that we have chosen for both French and Spanish. For French, the automatic correction of collocation errors, limited to N+V and V+N SVC combinations is due to be integrated into the *Interactive Language Toolbox* website. For Spanish, the corresponding module is integrated into the HaRenEs writing assistant environment.

In what follows, we briefly present each of these environments.

4.1 Interactive Language Toolbox

This online application offers access to the most relevant online lexicographical resources available for Dutch, English and French (*predictive writing aid*) and a spell, grammar and lexical checker for French (*corrective writing aid*)¹³; see Ziyuan (2012).

As shown in Figure 2, the application will not only display a list of alternatives for incorrect collocate selection, but will also give more information on the real use of word combinations with information on determiner use, for instance, (Figure 3) or authentic examples taken from a corpus or found on the web. Figure 3 shows that in almost 91% of corpus occurrences, the determiner *des* is used to combine *forces* (base) with *reprendre* (collocate).

¹² <http://www.lexchecker.org/>

¹³ Similar checkers for Academic Dutch and Dutch as a foreign language are in development and we plan to conceive a similar tool for Academic English.

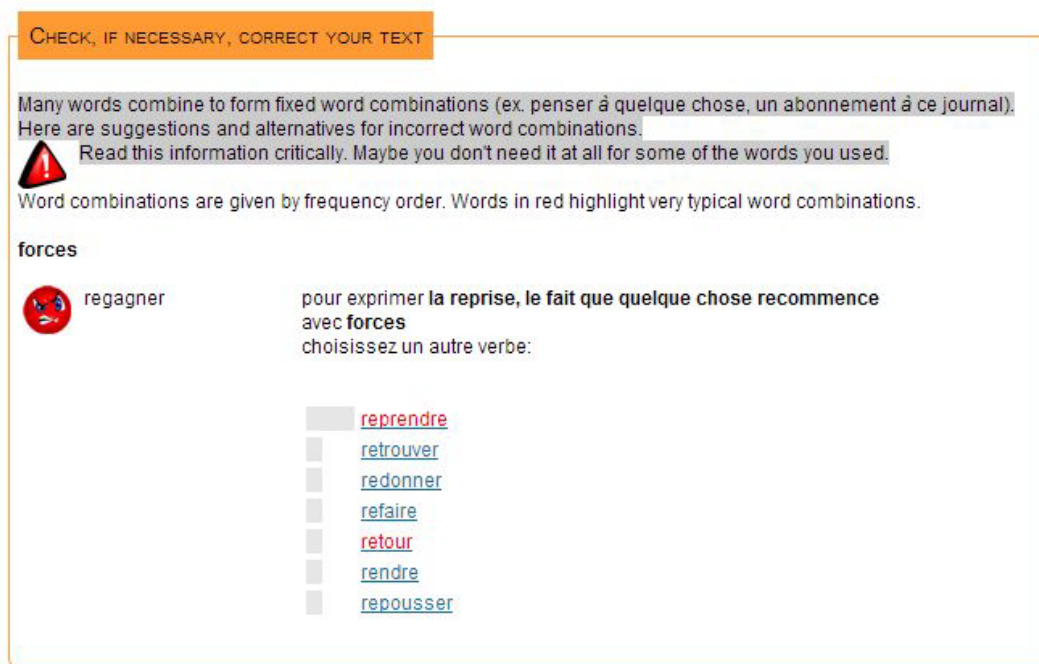


Figure 2: Interactive Language Toolbox: automatic collocation error detection and correction

		pourcentage des cas enregistrés
reprendre	des forces	90.95
reprendre	ses forces	3.41
reprendre	quelques forces	2.89
reprendre	les forces	1.03
reprendre	mes forces	0.77
reprendre	leurs forces	0.63
reprendre	vos forces	0.31

Figure 3: Interactive Language Toolbox: usage notes

In the advanced version, contextual information will aim to be even more extensive, similar to *StringNet* (Wible and Tsao, 2012).

4.2 HaRenEs Writing Assistant

The HaRenEs Writing Assistant is currently being developed in a common project by the University of La Coruña and Pompeu Fabra University. It allows the learner to verify the correctness of a specific Spanish collocation and, in the case of incorrectness, solicit correction suggestions, solicit examples of the use of a given collocation in context (in the reference corpus), and solicit the correction of collocations in a writing, etc. Figure 4 shows a snapshot of the user interface of the prototypical implementation of HaRenEs.

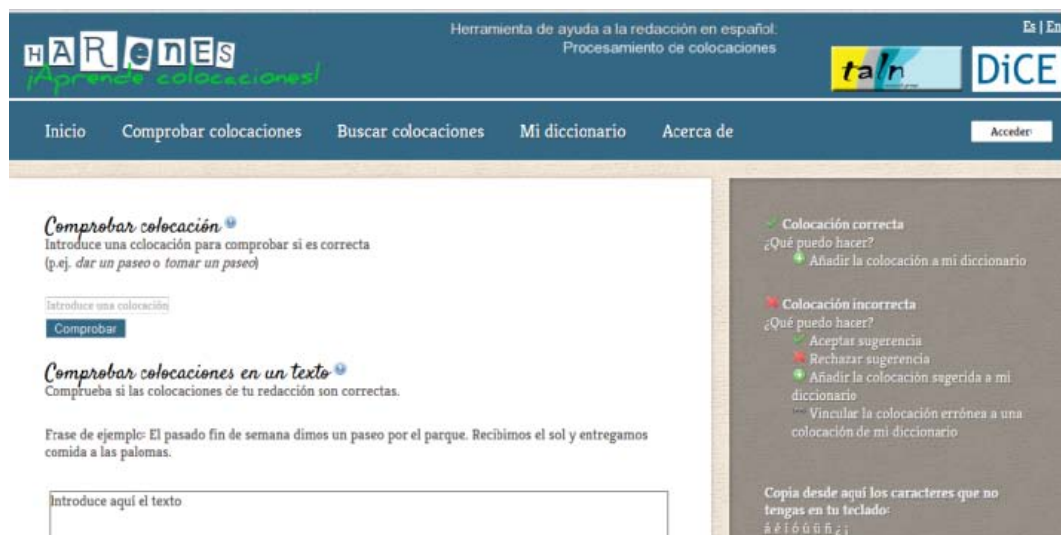


Figure 4: User interface of HaRenEs

Figure 5 shows the correction suggestions provided for the erroneous collocation *tomar [un] paseo*, lit. ‘take [a] walk’.

In an advanced version of the HaRenEs environment, users will be able to configure strategies for collocation error recognition and correction, choosing to either focus on selected types of collocations or to capture all collocations, but apply collocation type-specific error detection and correction strategies, to the extent available.

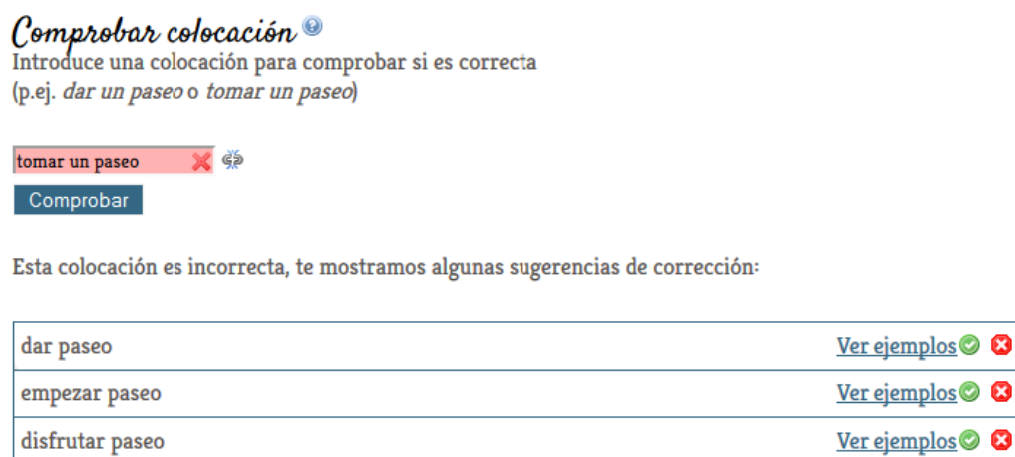


Figure 5: Correction suggestions provided by HaRenEs for *tomar [un] paseo*, lit. ‘take [a] walk’

5. Conclusions

As already demonstrated by previous works (see, e.g., Nesselhauf, 2004, Bolly, 2010) and as further supported by the studies presented in section 2 above, not all types of collocations are equally used by language learners and not all types pose the same

difficulty to learners. SVCs are the most problematic collocations for learners (at least for English learners of Spanish), such that the detection of their erroneous use and correction is of high priority.

To account for this need, we argue for collocation type specific error detection and correction strategies. For SVC verification and correction (collocation) dictionaries can play an important role. Thus, parallel corpora available on the Internet may fill existing gaps in traditional (translation) dictionaries, but non-native users will feel equally uncomfortable facing the amount of data provided by these applications. The need for a translation dictionary which offers translations in a more systematic way, for instance according to Mel'čuk's lexical functions as suggested by Kjaersgaard (2006), has been expressed for some time (see, e.g. Atkins, 1996; Danlos and Samvelian, 1992), but remains urgent.

On the other hand, corpus-based metrics that draw upon the insight that SVCs are the most common collocations are of relevance. A combination of lexicographic data, corpus analysis tools, and results and statistics combined with NLP-derived data thus provides new opportunities for (e-)lexicography.

In general, collocation error correction programs are a crucial writing aid for any L2 speaker. Such programs can be either used in a stand-alone sense (in the way the Interactive Language Toolbox and HaRenEs are currently conceived) to be consulted during writing, or be integrated into editor environments, such that an erroneous collocation is automatically highlighted and correction suggestions are offered upon request of the writer.

6. Acknowledgements

The work by M. Alonso Ramos, L. Wanner and their teams presented in 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-C02-01/02.

7. References

- Alonso Ramos, M. (2004). *Diccionario de colocaciones del español*
<http://www.dicesp.com>.
- Alonso Ramos, M., Nishikawa, A; Vinczthe, O. (2010a): "DiCE in the web: An online Spanish collocation dictionary", S. Granger, M. Paquot (eds.), *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009, Cahiers du Cental 7*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 367-368.
- Alonso Ramos, M. L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, S. Prieto, (2010b): "Towards a Motivated Annotation Schema of Collocation

- Errors in Learner Corpora”, *7th International Conference on Language Resources and Evaluation (LREC)*, La Valetta, Malta, pp. 3209-3214.
- Agustin Llach, M.P. (2011). *Lexical Errors and Accuracy in Foreign Language Writing*. Bristol, Buffalo, Toronto, Multilingual Matters. (Second Language Acquisition, 58).
- Atkins, B.T.S. (1996). Bilingual dictionaries. Past, present and future. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Rösler, P. Papi (eds). *Euralex '96 Proceedings. Papers submitted to the seventh EURALEX international congress on lexicography in Göteborg, Sweden*, Part II, pp. 515-546.
- Benson, M., Benson, E. and Ilson, R. (1997). *The BBI Dictionary of English Word Combinations*. Benjamins Academic Publishers, Amsterdam.
- Bolly, C. (2010). *Phraséologie et collocations. Approche sur corpus en français L1 et L2*. Brussels: P.I.E. Peter Lang.
- Chang, Y.C., J.S. Chang, H.J. Chen, and H.C. Liou. (2008). An Automatic Collocation Writing Assistant for Taiwanese EFL Learners. A case of Corpus Based NLP technology. *Computer Assisted Language Learning*, 21(3):283-299.
- Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pp. 34–38.
- Church, K.W. & P. Hanks. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pp. 76–83.
- Dafles* – Dictionnaire d'apprentissage du français langue étrangère ou seconde (<http://ilt.kuleuven.be/inlato>).
- Dahlmeier, D. and H.T. Ng. (2011). Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 107-117, Edinburgh, Scotland.
- Danlos, L., Samvelian, P. (1992). Translation of the predicative element of a sentence: category switching, aspect and diathesis. In *TMIMT-92, Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montréal, CWARC, pp. 21-34.
- Ferraro, G., Nazar, R., Wanner, L. (2011). Collocations: A Challenge in Computer-Assisted Language Learning. In I. Boguslavsky, L. Wanner (eds). *Proceedings of the 5th International Conference on Meaning-Text Theory (Barcelona, September 8-9, 2011)*, pp. 69-79.
- Fontenelle, Th. (1997). *Turning a bilingual dictionary into a lexical-semantic database*. Tübingen, Niemeyer. (Lexicographica, Series maior, 79).
- Futagi, Y., P. Deane, M. Chodorow, and J. Tetreault. (2008). A computational

- approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(1):353-367.
- Gouws, R. (2011). Learning, unlearning and innovation in the planning of electronic dictionaries. In P.A. Fuertes-Olivera, H. Bergenholtz. *e-Lexicography. The internet, digital initiatives and lexicography*. London, New York, Continuum, pp. 17-29.
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*. Oxford University Press, Oxford, pp. 145-160.
- Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. *Calico Journal*, 20(3), pp. 465-480.
- Hausmann, F.-J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(4):395-406.
- Heid, U. (1996). "Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora". In L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/ Philadelphia: Benjamins.
- Heift, T., Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning. Parsers and pedagogues*. New York, Routledge.
- Kjaersgaard, P. J. (2006). Esquisse d'un dictionnaire bilingue idéalisé. In Th. Szende (ed). *Le français dans les dictionnaires bilingues*. Paris, Champion, pp. 269-282.
- Krivanek, J. and D. Meurers (2011). "Comparing Rule-Based and Data-Driven Dependency Parsing of Learner Language." In *Proceedings of the Int. Conference on Dependency Linguistics (Depling 2011)*, Barcelona.
- Lesniewska, J. (2006). *Collocations and second language use. Studia Lingüística Universitatis Iagellonicae Cracoviensis*, 123:95-105.
- Lewis, M. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.
- Liu, A. Li-E., D. Wible, and N.-L. Tsao. 2009. Automated suggestions for miscollocations. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 47-50, Boulder, CO.
- Lozano, C. 2009. CEDEL2: Corpus escrito del español L2. In C.M. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind*. Universidad de Almería, Almería, pp. 197-212.
- Mel'čuk, I., Clas, A., Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve, AUPELF-UREF/Duculot.
- Moreno, P., G. Ferraro and L. Wanner. (2013). "Can we determine the semantics of

- collocations without semantics?”. In *Proceedings of eLex 2013*, Tallinn.
- Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, D. Stewart (eds). *Corpora and language learners*. Amsterdam, Philadelphia, Benjamins, pp. 109-124.
- Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: Benjamins.
- Park, T., E. Lank, P. Poupart, and M. Terry. (2008). Is the sky pure today? AwkChecker: An assistive tool for detecting and correcting errors. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08)*, New York.
- Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 54–57. Marrakech.
- Selva, Th., Verlinde, S., Binon, J. (2002). Le Dafles, un nouveau dictionnaire électronique pour apprenants du français. In A. Braasch, C. Povlsen (eds). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, August 13-17, 2002*. Copenhagen, CST, vol. I, pp. 199-208.
- Smadja, F. (1993). Retrieving Collocations from Text: X-Tract. *Computational Linguistics*.19.1:143–177.
- Verlinde, S., Peeters, G. (2012). Data access revisited : the Interactive Language Toolbox. In S. Granger, M. Paquot (eds). *Electronic lexicography*. Oxford, Oxford University Press, pp. 147-162.
- Wanner, L. (2004). Towards Automatic Fine- Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*. 10.2:95–143.
- Wanner, L., Alonso Ramos, M., Vincze, O., Nazar, R., Ferraro, G., Mosqueira, E., Prieto, S. (2011). *Annotation of Collocations in a Learner Corpus for Building a Learning Environment*. Paper presented at Learner Corpus Research conference, 2011.
- Wanner, L., B. Bohnet, M. Giereth and V. Vidal. (2005). ‘The first steps towards the automatic compilation of specialized collocation dictionaries’. *Terminology*, 11(1):137-174, 2005.
- Wible, D., Nai-Lung, T. (2012). Towards a new generation of corpus-derived lexical resources for language learning. In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds). *A taste for corpora. In honour of Sylviane Granger*. Amsterdam, Philadelphia, Benjamins , pp. 237-254. (Studies in corpus linguistics, 45).
- Ziyuan, Y. (2012). *Breaking the language barrier: a game-changing approach*. (<https://sites.google.com/site/yaoziyuan/publications/books/breaking-the-language-barrier-a-game-changing-approach>)

Advanced graph-based searches in an Internet dictionary portal

Peter Meyer

Institut für Deutsche Sprache, Mannheim

E-mail: meyer@ids-mannheim.de

Abstract

The web portal *Lehnwortportal Deutsch* (lwp.ids-mannheim.de), developed at the Institute for the German Language (IDS), aims to provide unified access to existing and possibly new dictionaries of German loanwords in other languages. Internally, the lexicographical information is represented as a directed acyclic graph of relations between words. The graph abstracts from the idiosyncrasies of the individual component dictionaries. This paper explores two different strategies to make complex graph-based cross-dictionary queries in such a portal more accessible to users. The first strategy effectively hides the underlying graph structure, but allows users to assign *scopes* (internally defined in terms of the graph structure) to search criteria. A second type of search strategy directly formulates queries in terms of the relational graph structure. In this case, search results are not entries but n-tuples of words (metalemmata, loanwords, etyma); a query consists of specifying properties of these words and relations between them. A working prototype of an easy-to-use human-readable declarative query language is presented and ways to interactively construct queries are discussed.

Keywords: graph database; loanword lexicography; search technology

1. Introduction

The *Lehnwortportal Deutsch* (lwp.ids-mannheim.de) is a freely accessible online lexical information system, developed at the Institute for German Language (IDS), that provides unified access to dictionaries of German loanwords in other languages. As well as conventional access to the individual dictionaries, the portal offers complex cross-dictionary search functionality; in particular, it can be used as an “inverted loanword dictionary” to trace the way of German words into different recipient languages. The portal web software operates on a database that represents pertinent lexicographical information as a cross-dictionary network of relations – more technically, a directed acyclic graph (DAG; cf. Bang-Jensen & Gutin, 2012) – between word forms of all included dictionaries.

This paper focuses on the problem of making complex graph-based cross-dictionary searches in the portal accessible to a wide range of users. In section 2, the general architecture of the *Lehnwortportal Deutsch* is described from a user’s point of view. The graph-based structure of the underlying unified data representation used for cross-dictionary searches is discussed in section 3. Section 4 shows how the web portal currently integrates some graph-related concepts in a unobtrusive way into fairly conventional HTML search forms suitable for average users. Section 5

concludes the discussion by outlining an alternative type of search strategy that provides advanced users with the opportunity to directly search the relational graph structure through an easy-to-learn, human-readable query language.

2. Basic access structure of the *Lehnwortportal Deutsch*

2.1 General information on the web portal

In its initial version, released in November 2012,¹ the web portal comprises three dictionaries on German loanwords in Standard Polish (de Vincenz & Hentschel, 2010), in the dialect of Polish spoken around the town of Cieszyn (Menzel & Hentschel, 2005), and in Slovene (Striedter-Temps, 1963). The two Polish dictionaries have previously been published electronically, whereas the Slovene dictionary was integrated through a combination of image digitization and manual extraction of relevant lexicographical information. The system is under active and continuous development and has a modular architecture that allows easy addition of new digital or digitized resources in XML format. In particular, a project is underway to integrate a newly-compiled dictionary of German loanwords in East Slavic languages that were mediated through Polish. There are long-term plans to incorporate a large number of further lexicographical resources on German loanwords in a wide range of other languages of the world.²

2.2 Accessing and navigating individual loanword dictionaries

The portal provides uniform access to the entries of all integrated loanword dictionaries. As a first step, a dictionary must be chosen from a menu on the right bar of the web page. In order to look up an entry in the dictionary, users may either type the beginning of a headword into an autocomplete text box or scroll through the alphabetical lemma list after selecting the initial letter in an alphabet bar (see Figure 1).

The microstructure of entries is entirely specific to the individual dictionaries. Due to considerable differences regarding intent, coverage and granularity, no attempt has been made to define a uniform one-size-fits-all entry structure (Meyer & Engelberg, 2010). There is, for each dictionary, a dedicated XML schema for its entry documents and, with the exception of those dictionaries where digitized images of print articles are shown, an accompanying XSLT stylesheet that transforms the XML source of its entries into HTML fragments.

¹ The web portal in its present form has been developed in a project funded by the Federal Government Commissioner for Culture and the Media upon a Decision of the German Bundestag.

² So far, there is little web traffic on the portal, possibly due to the limited number of available resources and the highly specialized targeted audience. On average, the number of page visits per day is still well below 100 and the advanced graph-based search options discussed in this paper are consulted less than twice a day.

Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache
(de Vincenz/Hentschel 2010)
A B C Ć D E F G H I J K L Ł M N O P R S Ś T U W Z Ź

→Zu diesem Artikel gehörige

dru
drukować
drumla
drut
druza
drejbogien
drejer
drejkienig
drelich
drelink
dreszajba
drezlować
dreznar
druk
drukarcz
drukować
drumla
drut
druza

drut

subst. m., ab 1494; auch *drot*, *drót*.
Zu: frühnhd. *droht*, *drot*, nhd. *Draht*.

1. Metallfaden, langer, feiner Stab aus Metall – metalowa nić lub długi ci
1528 Mymer¹ [31²], Sp^{xvi}
1562 WyprKr 8v, Sp^{xvi} *DwanaŃti AlŃpant, w nim dziefecz Ballas rul
ofmnaŃcie miedzi niemi na drotach zlotich.*
1593 KołakSzczęśl Dv, Sp^{xvi} *Tákże z Autorow text tu położony / Iá
drotow we wzor fznur plećionj.*
(†1611) 1613 SyrZiel 387, Sp¹⁷ *Drzeń chędogo drotom z niego wyi
korzenia.*
1749 BeimJelMed 263, Sp¹⁷ *Weźmi żelazny rozpalony drot [...] y.
oftrožnością przypal.*
1801–1805 N.Pam. 15 352, L *Żelazo ciągnione na drót.*
vor 1861 Swil *Drut złoty, srebrny, mosiężny.*
1948 Duch.Chem. 7, DoR *Pewne metale dadzą się wyciągać w dru*

Figure 1: Navigational elements in a sample article

2.3 Etymological metalemmata and the inverted loanword dictionary

The *Lehnwortportal* features an ‘inverted’ loanword dictionary (Engelberg, 2010) that lemmatizes all words of the donor language, German, that have been borrowed into the recipient languages represented by the different loanword dictionaries included in the portal. The concept of an inverted loanword dictionary was proposed more than forty years ago by Karaulov (1979), but dictionaries of this type are virtually non-existent to this day, with the notable exception of van der Sijs (2010) for Dutch loanwords in the world’s languages.

Setting up the inverted loanword dictionary for the *Lehnwortportal* is not a trivial task and cannot be performed automatically since any German etymon may appear in a variety of orthographical, diachronic, dialectal and other forms (henceforth referred to as ‘variants’ of the etymon) in different entries within and across loanword dictionaries. As an example, Standard Polish *lichtarz* is linked to a Middle High German etymon *liuhtaere* in de Vincenz & Hentschel (2010), whereas Slovene *lajhter* is related to New High German *Leuchter* and Middle High German *liuhtære* in Striedter-Temps (1963). Looking up the contemporary German word *Leuchter* ‘candlestick’ in the inverted loanword dictionary, the average user may reasonably expect to also be directed to entries that only list the corresponding Middle High German form of *Leuchter* in one of its orthographical variants *liuhtaere* or *liuhtære*. As a solution to this requirement, all German etymon word forms as they appear in the entries of the portal dictionaries were mapped to etymologically corresponding ‘normalized’ word forms, and wherever possible contemporary Standard German words. These normalized entries, henceforth *metalemmata*, are used as headwords

of the inverted loanword dictionary, whose entries, for the time being, mainly consist of hyperlinks to all loanword dictionary entries that list the metalemma or any of its diachronic, dialectal or other variants as an etymon. For each link, the corresponding German words in the target entry are given together with their definitions, if present.

Defining and mapping metalemmata involves many subtle philological and lexicographical problems and requires linguistically informed manual work. As the list of metalemmata grows rapidly with each newly included dictionary, and may require complex editing and correcting, using an administrative software tool for these tasks is indispensable. For the purposes of the initial version of the *Lehnwortportal*, a Java desktop application was developed that simply stores all information on metalemmata together with references to the exact places of corresponding etyma in the XML source documents in a separate file (henceforth ‘metalemma file’). The metalemma administration tool is also used to edit the cross-references within the metalemma list; thus, it is possible to mark a metalemma as a morphological derivative or constituent of another metalemma. This kind of internal cross-referencing is a prerequisite for finding loanwords borrowed from compounds or derivatives of a given German word. In a more advanced multi-user setting, however, a database solution would be more appropriate than locally editing a file.

The presentation of each loanword dictionary entry in the portal is complemented by links to all German metalemmata that correspond to etyma appearing in the entry. This information is dynamically constructed from the information contained in the inverted loanword dictionary. There may be references to multiple metalemmata for a given entry in case the entry discusses borrowings from several different, possibly morphologically related, etyma.

3. Using a directed acyclic graph (DAG) for unified data representation across heterogeneous resources

One of the distinctive features of the *Lehnwortportal* is the possibility of powerful cross-dictionary searches. Apart from obvious performance considerations, there are two lexicographical obstacles to using the unaltered XML source documents of the various component portal dictionaries for portal-wide search processes (cf. Meyer, 2013 for details):

(i) As mentioned, the individual dictionaries differ widely with respect to the microstructure of their respective entries (as reflected in the dictionary-specific XML schemata). Put simply, information of a certain kind can usually not be found “at the same place” in XML documents belonging to different dictionaries.

(ii) The terminology, concepts and data formats for specifying, e.g., the time of borrowing, grammatical features, or dialect appurtenance may vary considerably between dictionaries.

As a consequence of (i), an additional layer of lexicographical data is needed that represents relevant information of all component dictionaries in a unified structural format amenable to fast and efficient database queries. The solution opted for in the *Lehnwortportal* is to represent this lexicographical information as a network of relations (such as ‘is borrowed from’ or ‘is a derivative of’) between word forms (metalemmata, etyma and loanwords as well as their respective variants, derivatives etc.). To overcome the problem stated in (ii), the words that form the vertices of this network are annotated with grammatical, diasystemic and other information that is extracted from the original lexicographic resource and translated into a uniform data format.

More formally, advanced searches in the portal operate on a directed acyclic graph (DAG) whose vertices are word forms and whose edges are relations between word forms.³ At present, the following types of relations between two word forms *x* and *y* are used in the DAG:

- etymon *x* is mapped to metalemma *y*;
- loanword *x* is borrowed from etymon *y*;
- etymon or loanword *x* is an (orthographical, phonological, ...) variant of etymon/loanword *y*;
- *x* is a derivative of *y*;
- *x* is a compound of which *y* is a constituent;
- *x* is an etymologically related lexical parallel to *y* in another language (relevant for entries in Menzel & Hentschel, 2005).

In what follows, we will call *x* the ‘child’ and *y* the ‘parent’ of the relations enumerated above; in obvious graph-theoretical fashion, we will call the transitive generalizations of these terms ‘descendant’ and ‘ancestor’, respectively.

The DAG completely abstracts from the micro- and macrostructural idiosyncrasies of the individual component dictionaries; instead, it is generated in a fully automated process from parsing the underlying dictionary data and the metalemma file mentioned above. From the XML source of each dictionary entry in the portal (at least) one subgraph of the DAG – containing a loanword and its German etymon together with variants, derivatives etc., of either – is constructed in a dictionary-specific way. Roughly speaking, relations between word forms (edges in the DAG) are deduced from dictionary-specific structural relations between the corresponding XML elements or attributes.

³ A DAG has also been employed in the construction of the *Wörterbuchnetz* (<http://woerterbuchnetz.de/>) by the Trier Center for Digital Humanities, but its vertices correspond to dictionary entries, not individual words within entries (cf. Burch & Rapp, 2007).

Information from the metalemma file is used to connect etymologically-related subgraphs extracted from different entries and/or dictionaries – whose sources (vertices with in degree 0) are German etyma – in order to create larger, possibly cross-dictionary subgraphs whose sources are metalemmata. The web portal offers interactive visualizations of these larger subgraphs on the entry pages for the respective metalemmata, thus making it possible to get a visual impression of borrowings from a German word (cf. Figure 2).

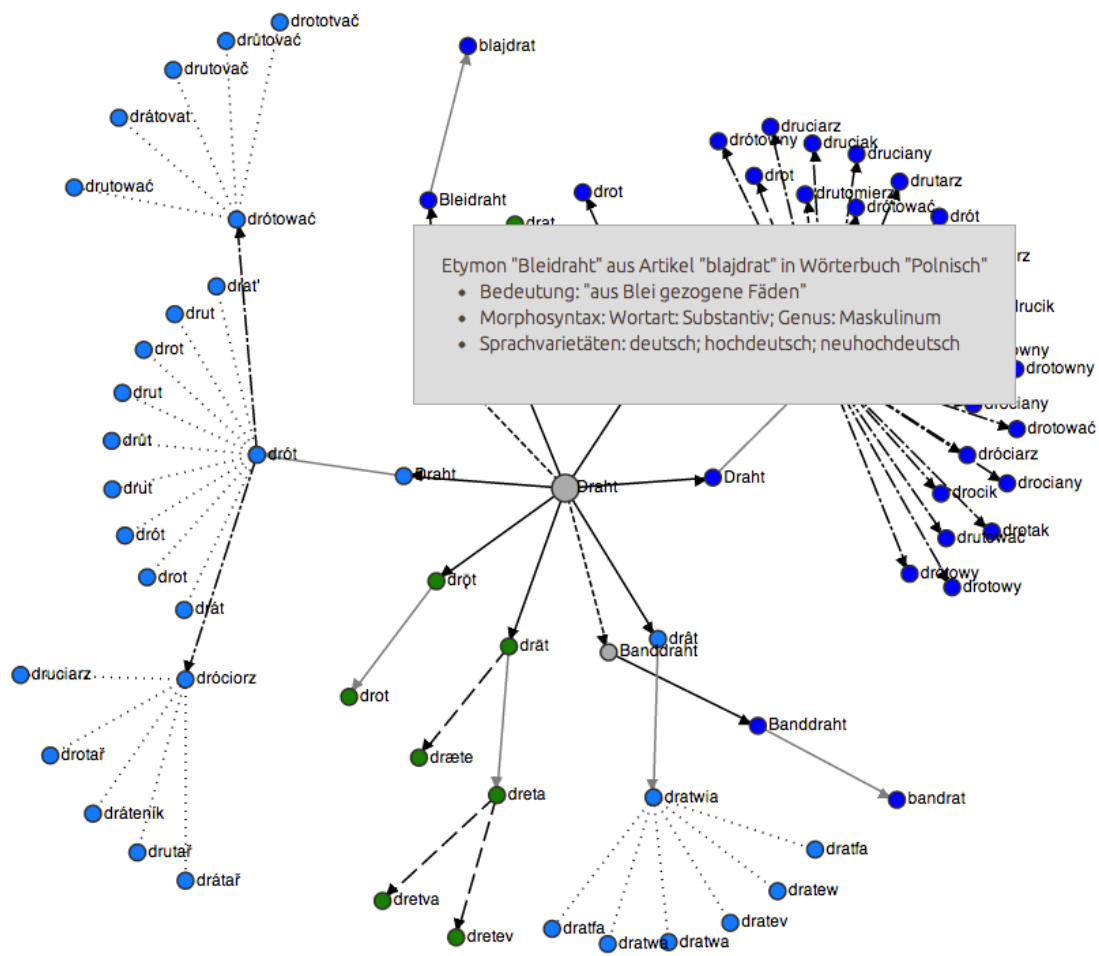


Figure 2: Interactive graphical representation of the subgraph related to the German metalemma *Draht* ‘wire’

As stated above, all vertices (word forms) of the DAG are annotated with morphosyntactic, diasystemic and meaning information in a standardized cross-dictionary format. This implies that for each dictionary an automated procedure has to be defined that translates lexicographical specifications from the dictionary-specific format into the standardized one. The intricacies involved in this task will not be discussed here; just one example: the German language variety

(dialect, historical stage) of an etymon may be used as a search criterion in portal-wide queries; therefore, a unified taxonomy of such varieties was defined for the *Lehnwortportal*, with each dictionary-specific language/variety name (e.g., ‘Silesian’) being mapped onto a set of ever-narrower language categories that can be used in searches (e.g., *High German, Central German, East Central German, Silesian German*). As soon as ‘fuzzy’ categories such as date of borrowing come into play, the picture gets considerably more complicated: if the etymon of a loanword *w* is just tagged as ‘Middle High German’ in the original entry, a query for words borrowed from German between 1300 and 1700 should return *w*, if possible with a low rating or weight. One possible way to account for such cases would be the use of a fuzzy ontology (Sanchez & Yamanoi, 2006).

A major advantage of using a DAG in the context of loanword dictionaries is the ability to adequately handle chains of borrowings in forthcoming extensions of the web portal. Thus, the Polish loanword *drukarz* ‘printer (profession)’ was borrowed from German *Drucker* and served in turn as the etymon for Ukrainian *drukar*. The indirect borrowing relationship between the Ukrainian and the German word is neatly expressed by a path in the DAG: *Drucker* (German metalemma) → *Drucker* (German etymon) → *drukarz* (Polish loanword) → *drukarz* (Polish etymon) → *drukar* (Ukrainian loanword). Note how the Polish intermediate appears twice in this graph on account of its dual role: as a German loanword it is a headword in a Polish loanword dictionary, and as the etymon for a Ukrainian loanword it appears in an entry of a Ukrainian loanword dictionary. It is even possible that these two dictionary entries contain contradictory information on the lexeme in question. Identifying the two words through a relationship ‘etymon *x* corresponds to loanword *y* in a borrowing chain’ is therefore additional information that has to be added to the lexicographical database by an expert lexicographer.

Note that the DAG is not a standalone database resource; it has to be recreated each time one of the underlying resources (including the correspondence information just mentioned) is altered or a new resource is added to the portal.

At present, the DAG is stored in a standard relational database, basically using two tables, one for the vertices and their properties, and one for representing the directed edges (relations between words) as ordered pairs of vertex IDs. The database does not only store all direct relations (edges) between words as enumerated above, but also their transitive closure, i.e., all indirect ancestor-descendant relations are also stored, which improves lookup times for complex queries. There are plans to migrate to a dedicated graph database such as *Neo4j* in the near future.

The overall architecture of the portal as outlined above, with its combination of heterogeneous XML-based resources and a uniform cross-resource DAG representation of both micro- and mediostructural information, is obviously applicable to other projects where unified access and search structures for interlinked

heterogeneous lexicographical resources are required. From a technical point of view, however, creating a programmatic abstraction layer that separates the backend, database-related core technology from specific issues of the *Lehnwortportal*, such as the specific lexicographical toolchain and the particular web application framework used for the portal, is not a trivial task and has not been accomplished so far. Publishing such an abstraction layer as an open source Java library is a long-term goal of the *Lehnwortportal* project.

4. Graph-based searches for the layman: Hiding the complexity

Adding a DAG-based homogenized data layer to the *Lehnwortportal* opens up a range of new possibilities for advanced cross-dictionary queries, but also increases the complexity for the average user who might not wish for graph-based data modeling just for moderately complex searches. So the question naturally arises as to how to reconcile usability requirements with the inherent complexity of data representation. In this section, we discuss the strategy that is pursued in the present version of the portal, i.e. using a fairly standard form-based search interface that effectively hides the underlying graph structure from the user. The HTML form for advanced portal-wide searches (<http://lwp.ids-mannheim.de/search/meta>) is split into three sections. In the initial default view, the topmost section offers users four search options for German etyma, viz. (a) an input field for specifying the etymon word form or its initial, final or middle part; (b) an input field for specifying a search string within the definition of the etymon; (c) a drop-down list of German varieties (mostly dialects and language stages) the etymon might belong to; and (d) a drop-down list of possible grammatical and morphosyntactical characteristics (such as POS, gender) of the etymon. The middle section offers analogous search criteria for loanwords. The bottom section permits a choice between two different modes of presentation for search results: per default, all matching entries in all loanword dictionaries are shown in alphabetical order of their respective headwords; alternatively, the set of matching metalemmata from the inverted dictionary can be displayed.

A loanword dictionary entry is considered matching if and only if it contains both an etymon (including variants etc.) and an *associated* loanword (again including variants, derivatives etc.) that both match their respective search criteria. A loanword *L* is considered associated with an etymon *E* if and only if *E* and *L* have a German metalemma *M* as a common ancestor in the DAG. *M* is called a matching metalemma for the search. The requirement that *L* must be associated with *E* is not trivial since a dictionary entry might discuss several etymologically different loanwords with their respective etyma. The condition for being associated is certainly not the most obvious one (which would be to have *E* as an ancestor to *L* in the DAG) but has the advantage of being less sensitive to the exact structure of the DAG: if, for instance, *L*'s etymon is represented as a variant of *E* in the DAG, this does not necessarily imply that *E* itself

cannot also be called an etymon for *L*; a lot depends on the lexicographical practice and granularity of each individual loanword dictionary.

Internally, each query returns all matching etymon-loanword pairs together with their respective matching metalemmata. Depending on the selected presentation mode, either the entries corresponding to the etymon-loanword pairs or the metalemmata are shown. In the metalemma search mode, all matching etymon-loanword pairs, sorted by dictionary entry, can be displayed. Thus, the underlying search is formulated and executed in graph-related terms: the etyma-loanword-metalemma triples correspond to *subgraphs* of the DAG. From the user's point of view, however, only a simple conjunction of search criteria concerning etyma and/or related loanwords is specified as a query, the search result being a straightforward list of dictionary entries. As an example, Figure 3 shows a simple query for dictionary entries containing both a German etymon whose definition contains the word *Metall* 'metal' and an associated loanword that is a Polish noun. Neither the search form nor the search result (a list of links to dictionary entries) refers explicitly to graph-theoretical concepts, although they are implicit in the requirement that matching loanwords must somehow 'belong to' matching etyma.

For even more advanced queries, all eight search fields in the HTML form can be expanded to yield a conjunction of at most 16 search criteria altogether. Each criterion in turn can be a conjunction or a disjunction of two similar criteria (e.g., 'is a noun OR is a verb') and, more importantly, can be assigned what will be hereafter referred to as a *scope*. Apart from default scope (meaning that the criterion applies to the etymon or loanword in question) a user can assign *entry scope* or *portal scope* to any criterion. In this way, it is possible to additionally specify properties of *other* loanwords or etyma that are *associated* with the etymon-loanword pair in question and that appear either elsewhere within the entry (entry scope) or in any arbitrary dictionary entry of the portal (portal scope). Again, being associated is defined with respect to the DAG as having a common metalemma ancestor. A typical scenario for using a wider scope might be a search for loanwords that have derivatives or compounds with certain properties. Figure 4 presents a sample extension of the query shown in Figure 3 requiring that matching entries include an etymologically related word ending in *-owy* or *-owny* (both are typical denominal adjective suffixes in Polish). A reasonable example for a criterion with portal scope would be 'language: Slovene' in the loanword section; this amounts to the requirement that there be an etymologically-related loanword in Slovene.

The idea of 'annotating' search criteria could easily be extended to cover the problem of handling borrowing chains: users may wish to specify whether a certain criterion applies to intermediate or to terminal etyma or loanwords in a chain.

Angaben zum deutschen Herkunftswort			
Herkunftswort	ist gleich	<input type="text"/>	+
Bedeutungserläuterung enthält		Metall	+
sprachliche/raumzeitliche Einordnung		(beliebig)	+
grammatisches Merkmal		(beliebig)	+

Angaben zum Lehnwort			
Lehnwort	ist gleich	<input type="text"/>	+
Bedeutungserläuterung enthält		<input type="text"/>	+
Sprache		polnisch	+
grammatisches Merkmal		Wortart: Substantiv	+

Suchen		
Lehnwörter anzeigen	Abfrage starten	Formular leeren

Suchergebnisse

- abszrot
- bankajza
- basethorn
- bestocajg

Figure 3: Example of an advanced cross-dictionary search query in the *Lehnwortportal*

Angaben zum Lehnwort			
Lehnwort	endet auf	<input type="text" value="owy"/>	-
	oder	<input type="text" value="owny"/>	
Kriterium gilt für		ein etymologisch zugehöriges Wort im Artikel	

Figure 4: Assigning a scope to a search criterion

As a downside of this approach, queries might return surprisingly complex semantics. To really understand the results returned, the user has to be aware of the fact that the underlying query is formulated in terms of etymon-loanword pairs. Suppose, for instance, that only one criterion C is specified in the loanword section of the HTML form and that it happens to have entry scope. If at least one relevant loanword L in a dictionary entry complies with C , then the underlying result pairs every etymon E in this entry that matches the etymon-related search criteria, with *all* loanwords in the same entry that are associated with both L and E . This is in contrast to the case of

default scope of *C* where only those loanwords that fulfill *C* can be a component of the etymon-loanword pairs returned. Even more confusing is that the list of dictionary entries presented as the search result to the user is the same in both cases (default vs. entry scope of *C*); this is because in both cases the only loanword-related requirement is that matching entries contain at least one loanword fulfilling *C* and be associated with an etymon matching the other search criteria.⁴

Another restriction is that multiple criteria with extended scope cannot be made to refer to the same words. Thus, if a user assigns entry scope to two loanword-related criteria (such as ‘language: Polish’ and ‘POS: adjective’) this does not equate to the requirement that there be an etymologically-related Polish adjective in the entry; rather, it simply means that among the loanwords in the article there must be both an adjective and a (possibly identical) Polish word. Of course, it would be possible to refine the annotation scheme to cover at least the most useful relations between scoped criteria, but at the cost of reduced usability.

5. Graph-based searches for professionals: Using a declarative domain-specific query language

Under the hood, advanced searches in the *Lehnwortportal* as outlined above are all based on the graph-theoretical notion of a common ancestor of two or more nodes. To unleash the full range of structural search possibilities it is desirable to have the possibility of formulating queries directly in terms of arbitrary graph configurations.

For this kind of search technology to be accessible to interested professionals without IT background, an easy-to-use human-readable query language should be employed that allows the user to describe the properties of the subgraphs s/he is looking for. The language should be *declarative* in that the actual process of finding subgraphs with the desired properties in the DAG need not be defined by the user. The following remarks report on the results of some preliminary research work on a tailor-made query language for the *Lehnwortportal*.

Most currently used generic query languages (cf. Wood, 2012, for an overview) for graph databases are geared towards IT professionals, typically having an SQL-like syntax, like the *Cypher* language for the Neo4j database (see <http://www.neo4j.org/learn/cypher>; cf. Robinson et al., in press). The approach taken for the *Lehnwortportal* was to design a highly domain-specific language whose expressions are actually very close to human language; furthermore, complex queries should be expressible through an unordered list of short ‘sentences’ that can easily be adapted from some sample set. Here is how a query in such a language might appear for the search task that was used as an example above:

⁴ As a convention in the *Lehnwortportal*, at least one criterion in an advanced query must have default scope because otherwise search results can easily get incomprehensible.

```

/* (1) Declare node variables: */
find metalemma metaLemma.
find etymon metalWord.
find loanword polishNoun.
find loanword polishAdj.
find loanword sloveneWord.

/* (2) Define relations between words: */
metaLemma is metalemma for metalWord.
polishNoun is descendant of metaLemma.
polishAdj is derivative of polishNoun.
sloveneWord is descendant of metaLemma.

/* (3) Express constraints on words: */
definition of metalWord contains 'Metall'.
language of polishNoun is Polish.
part of speech of polishNoun is noun.
part of speech of polishAdj is adjective.
polishAdj ends in 'owy'
or polishAdj ends in 'owny'.
language of sloveneWord is Slovene.

/* (4) Define how results are shown: */
show metalWord, polishNoun, polishAdj.

```

This query is obviously both more precise and semantically more perspicuous than its HTML form-based counterpart. Each query expression consists of an unordered list (a conjunction) of *clauses*, each ending with a period, that together specify a ‘graph pattern’ for subgraphs of the DAG. This is close to the syntax of the query language used for the NAGA search engine (Kasneci et al., 2008) with an additional layer of ‘syntactic sugar’ on top. Internally, the period-delimited clauses are just constituents of the query expression as defined in the context-free grammar for the query language. Strings enclosed between ‘/*’ and ‘*/’ are also constituents and are treated as comments. In (1), the nodes in the graph pattern (word forms) are labelled by user-defined node variables and simultaneously classified as metalemmata, etyma or loanwords. In (2), specific relations between these nodes are defined; edges between two vertices are specified by their type (e.g., ‘is derivative of’), while indirect connections through paths of arbitrary length can be given in abstract graph-theoretical terms (‘is descendant of’). Properties of vertices (words) are defined in (3). The clause in (4) controls how the search result is to be displayed. Formally, search results are ordered as n-tuples of words (metalemmata, loanwords, etyma) belonging to the appropriate vertices of matching subgraphs. In our example, all matching combinations of three of the five variables are to be shown, ordered alphabetically first by metalWord, then by polishNoun and finally by polishAdj.

The convoluted process of translating such query expressions into native database queries⁵ creates a useful layer of domain-specific abstraction from implementation details. One advantage is ease of use: for each of the steps (1) to (4) demonstrated above, users can simply choose component clauses of their queries from a limited number of pre-defined clause templates and combine them, where necessary, with Boolean operators. It is straightforward to construct an interactive drag-and-drop user interface – similar to the Scratch programming environment (<http://scratch.mit.edu/>) – that guides users through the process of selecting templates and operators and constantly checks for errors such as misspelled variable names, illegal cycles in graph patterns etc.⁶ As an additional benefit, it becomes almost trivial to create a multilingual version of the query language.

6. Conclusion: Making complex graph-based searches more accessible

The *Lehnwortportal Deutsch* offers an innovative and principled way of making a portal of heterogeneous lexicographical online resources more than the sum of its parts by providing a unified graph-based database representation of all lexicographical data. The benefits of this approach come at a price – not only on the lexicographical side, but also for the user who has to tackle increased complexity of search options. This paper has shown how the present version of the portal manages to shield users from direct exposure to the graph database, which, however, severely restricts and sometimes obscures the semantics of such queries. An alternative strategy has been outlined that tries to make it as easy as possible to use a graph-based query language. It must be emphasized, however, that both strategies address not casual users but experts who wish to use the portal as a research instrument. Integrating a graph database into a semantic-search system (such as Google Knowledge Graph or Wolfram Alpha) that is suitable for use by laypeople is a much more difficult task.

⁵ On a technical note, a parser combinator library is used to construct an Abstract Syntax Tree (AST) from the query expression; the AST is then traversed and processed recursively to generate the underlying database query, at present a SQL query. For each node of the AST, an instance of a certain Java class is created that represents the different parts of the SQL query (select/from/where/order by) as they are partially determined by this node. The object corresponding to the root node of the AST is used to produce the SQL string.

⁶ A further step would be the use of a visual version of the query language, comparable to *qGraph* (cf. Blau et al., 2002). Users could then literally draw the query subgraphs using a pointing device and a keyboard.

7. References

- Bang-Jensen, J., Gutin, G. Z. (2009). *Digraphs: theory, algorithms and applications*. London: Springer.
- Blau, H., Immerman, N. & Jensen, D. (2002). A Visual Language for Querying and Updating Graphs. Technical Report 2002-037, University of Massachusetts, Amherst.
- Burch, T., Rapp, A. (2007). Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven. In D. Burckhardt, R. Hohls & C. Prinz (eds.) *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006* (= Historisches Forum 10/I). Berlin, pp. 607-627. Accessed at: http://edoc.hu-berlin.de/histfor/10_I/PHP/Woerterbuecher_2007-10-I.php#007001
- de Vincenz, A., Hentschel, G. (2010). *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts.* (= Studia slavica Oldenburgensia, vol. 20). Oldenburg: BIS-Verlag. Online version: <http://www.bis.uni-oldenburg.de/bis-verlag/wdlp>.
- Engelberg, S. (2010). An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6-10 July 2010)*. Ljouwert (Leeuwarden): Fryske Akademy, pp. 639-647.
- Karaulov, J. N. (1979). Obratnyj slovar' zaimstvovanij kak sposob isučenija lingvoëkologii. *Izvestija Akademii Nauk SSSR. Serija Literatury i Jazyka*, 38/6, pp. 552-562.
- Kasneci, G., Suchanek, F.M., Ifrim, G., Ramanath, M. & Weikum, G. (2008). NAGA: Searching and Ranking Knowledge. In *IEEE 24th International Conference on Data Engineering, 2008*. ICDE 2008, pp. 953–962.
- Menzel, Th., Hentschel, G. (2005). *Wörterbuch der deutschen Lehnwörter im Teschener Dialekt des Polnischen. 2nd, enlarged and revised ed. online:* Accessed at <http://www.bkge.de/14451.html>.
- Meyer, P., Engelberg, S. (2011). Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In H. Hedeland, Th. Schmidt & K. Wörner (eds.) *Multilingual Resources and Multilingual Applications*. Hamburg: Universität Hamburg, Sonderforschungsbereich 538 Mehrsprachigkeit, pp. 169-174.
- Meyer, P. (2013). Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation. In S. Kempgen, N. Franz, M. Jakiša & M. Wingender (eds.): *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013*. München: Otto Sagner, pp. 233-242.

- Robinson, I., Webber, J. & Eifrem, E. (in press). *Graph Databases*. Beijing etc.: O'Reilly Media.
- Sanchez, E. & Yamanoi, T. (2006). Fuzzy ontologies for the semantic web. In H.L. Larsen, G. Pasi, D.O. Arroyo, T. Andreasen & H. Christiansen (eds.) *Proceedings of the 7th International Conference on Flexible Query Answering Systems, Milan, Italy*. London etc.: Springer, pp. 691-699.
- Striedter-Temps, H. (1963). *Deutsche Lehnwörter im Slovenischen*. Wiesbaden: Harrassowitz.
- van der Sijs, N. (2010). *Nederlandse woorden wereldwijd*. Den Haag: SDU Uitgever.
- Wood, P. T. (2012). Query Languages for Graph Databases. *SIGMOD RECORD*, 41(1) (March), pp. 50-60.

The lexical editing system of Karp

**Lars Borin, Markus Forsberg, Leif-Jöran Olsson,
Olof Olsson and Jonatan Uppström**

Språkbanken, Department of Swedish, University of Gothenburg, Sweden
{lars.borin, markus.forsberg, leif-joran.olsson, olof.olsson2, jonatan.uppstrom}@gu.se

Abstract

Karp is the open lexical infrastructure of Språkbanken (the Swedish Language Bank). The infrastructure has three main functions: (1) to support the work on creating, curating, and integrating our various lexical resources; (2) to publish the resources, making them searchable and downloadable; and (3) to offer advanced editing functionalities. An important feature of the lexical infrastructure is also that we maintain a strong bidirectional connection to our corpus infrastructure. At the heart of the infrastructure is the SweFN++ project with the goal to create free Swedish lexical resources geared towards language technology applications. The infrastructure currently hosts 23 Swedish lexical resources. The resources are integrated through links to a pivot lexical resource, SALDO, a large morphological and lexical-semantic resource for modern Swedish.

Keywords: lexicon, editing, infrastructure, Swedish language resources, language technology, LMF

1. Introduction

The research and development unit Språkbanken¹ (the Swedish Language Bank) at the University of Gothenburg has since its establishment in 1975 accumulated a large variety of language resources, including corpora of over two billion words of modern and historical Swedish text and a multitude of lexical resources. Some of the lexical resources are digitized dictionaries describing older forms of Swedish, but most of them are contemporary resources intended for NLP use. For most of these, the development of an adequate technical support infrastructure has been hampered by limited research funding, thus leading to the adoption of suboptimal technical solutions such as simple form-based frontends to relational databases or even tab-separated text files, saddling the lexicographers with the responsibility for making sure that any formal requirements are met and for manually weeding out any inconsistencies.

The SweFN++ project (Borin et al., 2010a) had the objective to create, curate, and integrate free Swedish lexical resources with the explicit goal of making them usable for language technology applications. META-NORD² is a broad EC-funded European collaboration with the aim of upgrading and harmonizing language resources and

¹ <<http://spraakbanken.gu.se>>

² <<http://www.meta-nord.eu>>

tools for the Nordic and Baltic languages and making them available across Europe. Thanks to these two, and other externally funded infrastructure initiatives,³ Språkbanken has had the opportunity to focus on safeguarding its existing language technology resources, as well as to develop a generalized lexical infrastructure, referred to as Karp (Borin et al., 2012a). The heart of the lexical infrastructure is a large network of interconnected lexicons (Borin, 2010; Borin et al., 2010a), all encoded in the LMF format (Lexical Markup Framework; see ISO, 2008; Francopoulo, 2013).

Even though our digital lexical resources are primarily intended for use in NLP applications, they are still very much lexicographical entities. Thus, from a linguistic point of view, the work on individual resources as well as on their integration is at heart a genuinely lexicographical activity, to boot one with considerable potential to make significant theoretical contributions to lexicology, lexical semantics and lexical typology because of the large-scale empirical nature of our endeavor and the diversity of the lexical resources involved. In general, working with large amounts of data as we do, requires good tools for interacting with the data, for abstracting, ordering, searching and visualizing the data, for inferring and presenting relations among data items, and for editing the data. The Karp component of our lexical infrastructure has been designed with these aims in mind.

An important feature of the lexical infrastructure is that we maintain a strong bidirectional connection to our corpus infrastructure Korp (Borin et al., 2012b). For example, the corpora are annotated with the lexical information in Karp, and the language examples for the lexical resources in Karp are retrieved from Korp.

A pervasive theme of the infrastructure is openness, which may be seen as a philosophical stance – we believe that research should be carried out in the open to enable inspection and increased collaboration. Openness pervades the infrastructure, in the use of open standards and open-content licenses, as well as the daily publication of not only the resources but everything else that is available in-house, such as formal test protocols, change history and the tools themselves. The tools are available through a set of web services, which are open for others to use, and which provide a convenient way of accessing the lexical information programmatically.

One essential part of this infrastructure is a generic search interface, <<http://spraakbanken.gu.se/karp>>, which provides a plug-and-play search tool for

³In this context, an important initiative is CLARIN <<http://www.clarin.eu>>. Although Sweden is not yet a member of CLARIN, Språkbanken is involved in some CLARIN activities and is also the coordinating partner of a recently submitted funding application for Swedish membership. In the development of Karp and other infrastructure components, we pay close attention to the standards and best practices defined by CLARIN, in order to be able to quickly set up a CLARIN service center when Sweden decides to join CLARIN.

resources already in LMF, where the LMF format is employed both internally within the infrastructure and, trivially, as an export format. As a next logical step, we have augmented the search interface of Karp with editing functionalities, where authorized users may edit and create new lexical entries.

2. The lexical resources

The lexical infrastructure has one primary lexical resource, a pivot, to which all other resources are linked. This is SALDO⁴ (Borin and Forsberg, 2009; Borin et al., 2013b), a large (130K entries and 1.9M wordforms), freely available morphological and lexical-semantic lexicon for modern Swedish. It has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other resources are linked.

Some of the 23 resources (including the pivot resource SALDO) have been created from scratch using existing free resources, both external and in-house. For example, Swesaurus, a Swedish wordnet (Borin and Forsberg, 2010; Borin and Forsberg, 2011b), is being built using not only in-house but also external resources, such as Synlex (Kann and Rosell, 2006), the Swedish Wiktionary,⁵ and more indirectly, using semantic relations extracted from Princeton WordNet (Fellbaum, 1998b) through links between SALDO and Core Princeton WordNet (Boyd-Graber et al., 2006).

Other resources are the result of digitization and (manual and automatic) post-processing of existing paper dictionaries. This holds generally for the historical lexicons and their associated (partial) morphologies (Borin and Forsberg, 2008; Borin and Forsberg, 2011a; Borin et al., 2010b).


As an illustration of the diversity of the resources, here follows a selection of the results of the word form query *springa* in Karp. The selection consists of 13 out of 62 results in seven out of 23 resources.

springa in SALDO

The word sense *springa* ‘run’ in SALDO is described with two semantic relations, the primary relation *röra sig* ‘move’ and the secondary relation *fort* ‘fast’. Furthermore, we have relations where *springa* acts as the primary or secondary, i.e. the reverse relations collectively referred to as children. SALDO has two more word senses of *springa*, one noun and one verb, not shown here, and *springa* is also a component of 11 particle verbs, e.g. *springa bort* ‘run away’.

⁴ <<http://spraakbanken.gu.se/saldo>>

⁵ <<http://sv.wiktionary.org>>


Sense	POS	Primary	Secondary	Children (primary)	Children (secondary)
<i>springa</i> 	verb	<i>röra sig</i>	<i>fort</i>	<ul style="list-style-type: none"> ○ <i>lubba –</i> ○ <i>kesa kalv</i> ○ <i>ifråsprungen ifrån</i> <p>Show all (30)</p>	<i>springa ihop² kollidera</i>

springa in SALDO morphology

The morphological description of SALDO is a separate resource that lists lemgams associated with word senses, where lemgams are form units denoting inflection tables.⁶

The lemgam *springa* (noun), which denotes the inflection table of the two verb senses of *springa*, illustrates the connection to the corpora infrastructure: next to the small raven we see the number of hits in Korp’s corpora collection (307,539 hits), and the table shows which of the word forms are attested: only *sprunges* (passive past subjunctive) is unattested.

springa (verb)

 (307,539)

springa ...

pres ind aktiv	<i>springer</i>
pres ind s-form	<i>springes</i>
pres ind s-form	<i>springis</i>
pres konj aktiv	<i>springe</i>
pres konj s-form	<i>springes</i>
pret ind aktiv	<i>sprang</i>
pret ind s-form	<i>sprangs</i>
pret konj aktiv	<i>sprunge</i>
pret konj s-form	<i>sprunges</i>
imper	<i>spring</i>
inf aktiv	<i>springa</i>
inf s-form	<i>springas</i>

springa in Swedish FrameNet

The word sense *springa* ‘run’ is a lexical unit in the frame Self_motion in the Swedish FrameNet. A click on the frame name takes us to the full description of the frame.

Sense	Frame
<i>springa</i>	<i>Self_motion</i>

A frame is a large information unit: only part of the entry is shown here. Self_motion is a frame from the Berkeley FrameNet (Baker et al., 1998; Ruppenhofer et al., 2005)

⁶ More specifically, a lemgam is a unique combination of a citation form and certain other formal characteristics, in SALDO pronunciation, part of speech, inflectional paradigm and compounding behavior. This corresponds to one usage of the term lemma, but unfortunately this term also is used in other meanings, e.g. ‘citation form’, which is why we have opted for coining a new, unambiguous term.

that the Swedish FrameNet is built upon, and the core and non-core elements have been directly imported from that resource. The word sense *springa* ‘run’ occurs in past tense, *sprang*, in the first annotated example: *Två hästar på rymmen sprang . . .* ‘Two horses on the loose ran (on the roadway in southern Södertälje in the afternoon.)’, but is also listed among the lexical units (not shown).

Self_motion

Domain: **Gen**

Semantic Type: **Move**

Core Elements: **Area Direction Goal Path Self_mover Source**

Peripheral Elements: **Concessive Coordinated_event Cotheme Depictive Distance Duration External_cause Internal_cause Manner Means Path_shape Place Purpose Reason Result Speed Time**

Examples:

Två hästar på rymmen *sprang ute på vägbanan i södra Södertälje på eftermiddagen*.

springa in Swesaurus

The word sense *springa* ‘run’ has seven graded synonymy relations in Swesaurus, all extracted from Synlex (Kann and Rosell, 2006). They are all manners of running, such as *rusa* ‘rush’, roughly corresponding to the troponyms of Princeton WordNet (Fellbaum, 1998a).

Sense	Degree	Sense	Type	Source
<i>springa</i>	↖80	↗ <i>kila</i> ²	syn	fsl
<i>springa</i>	↖92	↗ <i>kuta</i>	syn	fsl
<i>springa</i>	↖94	↗ <i>lubba</i>	syn	fsl
<i>springa</i>	↖88	↗ <i>löpa</i>	syn	fsl
<i>springa</i>	↖80	↗ <i>rusa</i>	syn	fsl
<i>springa</i>	↖90	↗ <i>ränna</i> ²	syn	fsl
<i>springa</i>	↖62	↗ <i>skubba</i>	syn	fsl

springa in IDS/LWT

The IDS/LWT list is a massively multilingual vocabulary of 1,460 word senses used for typological studies. The basic list, 1,310 entries, was first compiled in the Intercontinental Dictionary Series project (Borin et al., 2013a), and new languages are continually being added to the IDS archive at the Max Planck Institute for Evolutionary Anthropology in Leipzig.⁷ Another 150 entries were added when the IDS


⁷ <<http://lingweb.eva.mpg.de/ids/>>

list was used in the recent Loanword Typology project (Haspelmath and Tadmor, 2009).⁸ The new information provided in our version of the Swedish IDS/LWT list is the link between *springa* ‘run’ and the IDS/LWT id S10.460, thereby providing a link from our Swedish lexical resources to a basic vocabulary in over 200 languages.

Sense	FormRepresentation			
<i>springa</i>	LWT ID	English	Definition	Example
	S10.460	to run	(intransitive)	"They ran all the way to school."

springa in Schlyter

Schlyter (1887) is an Old Swedish dictionary describing the vocabulary of Old Swedish law texts, which becomes clear in the definition text: it describes the expression *springa af kaghen* ‘run of the scaffold’, which is a punishment involving a pillory on an elevated platform for public shame or whipping.

springa (verb)
 (40)

springa v. n.

1. *springa*. Chr. * s. af kaghen, o: stå vid skampålen på den upphöjda schavotten för att skämmas l. hudstrykas, och sedan *springa* ned därifrån (jfr. Kgh, och VSt. o. kac), Sk.*; detta straff beskrives tydligare i Dr. Margaretas D. stadsrätt c. 55, där det säges att brottslingen schall settis paa kaget (läs kagen) och self needsprunge (KR. GDL . V. 511; jfr. s. 596).

springa in Diapivot

The Diapivot resource (Borin and Forsberg, 2011a; Andersson and Ahlberg, 2013) provides diachronic links between the lemgrams of the four morphological resources: SALDO, the SALDO morphology (the pivot); Dalin, a 19th century morphology; Swedberg, a 17th century morphology; and finally, an Old Swedish morphology. The linking is done using SKOS (Simple Knowledge Organization System; see Miles and Pérez-Agüera, 2007) where the linking relations are either equivalence or a broader-narrower relationship.

In the lexical entry below we have three lemgrams *springa* (verb), which are not formally the same: they all live in different namespaces. The first one is from the pivot SALDO morphology, followed by a lemgram in Dalin, and three lemgrams in the Old Swedish morphology. All are linked with the equivalence relation.

<i>springa</i> (verb)  (307,539)	<i>springa</i> (verb)  (1) Dalin	<i>löpa</i> (verb)  (158) Old Swedish	<i>skumpa</i> (verb)  (0) Old Swedish	<i>springa</i> (verb)  (40) Old Swedish
--	---	--	--	--

⁸ <<http://wold.livingsources.org/>>

The links between the lemgrams are clearly providing under-specified information. The links are on the sense level where it is proper to talk about equivalence, not on the form level. However, since most words in any of the lexical resources are monosemous (Borin, 2010), most lemgram links are in fact also sense links. Because of this, we have at the moment settled for accepting some degree of under-specification in the Diapivot to allow the resource to grow quickly. Establishing proper word sense links is, of course, part of the ever-growing future work.

3. Search in Karp

Arguably the biggest motivation for building the editing system on top of the existing Karp database is to make use of the extensive and already existing search functionalities. There are four ways to search the Karp lexicons, as described in the following sections. The different ways of searching are available in Karp's search interface and through its web services.

3.1 Basic search

The basic search accepts a wordform, a sense identifier, or a lemgram. The lexical entries containing the requested information are returned.

In addition, the basic search supports full text search in the textual parts of the lexical resources, such as examples and definitions. The full text search, beyond extending the search capabilities, also makes the lexical information lacking wordforms, senses, and lemgrams discoverable.

3.2 Pivot search

The pivot search accepts a wordform that is looked up in all selected morphologies. If one or more lemgrams are found, the lexical entries containing the lemgrams or any of their associated senses are returned.

For example, a search for *katter* 'cats'

- ⇒ finds the lemgram *katt..nn.1* in the SALDO morphology and *katter..nn.1* in the Old Swedish morphology.
- ⇒ finds all senses of *katt..nn.1* and *katter..nn.1*
 - ⇒ searches for *katt..nn.1* and *katter..nn.1* and their senses in the current lexicon selection.

3.3 Diapivot search

As previously mentioned, there is a diachronic pivot resource that links the lemgram units of different morphologies – typically reflecting different historical stages of Swedish, hence the name 'Diachronic pivot' or 'Diapivot' – and thus acts as a middle-layer allowing the location of diachronic lexical information related to the

current search, e.g. a spelling variation or, moving backwards in time, a completely different form unit related to the current search.

Consider *springa* (verb) in the Diapivot resource that was exemplified in section 2. A diapivot search for *springa* (verb) would trigger a search for all lemgrams to which *springa* (verb) is linked in the Diapivot resource, i.e. Dalin *springa* (verb), Old Swedish *löpa* (verb), etc.

The diapivot search has been incorporated into the corpus search interface Korp (Borin et al., 2012b), so that, e.g. a search for *räv* (noun) ‘fox’ also finds words like *räf* (noun) ‘fox’ (a 19th century spelling variant).

3.4 Extended search

The extended search enables search in any of the data fields occurring in the resources of Karp. It uses CQL (Contextual Query Language)⁹ as the query language, which supports complex queries using logical operators, regular expressions, sorting, and more.

The extended search is represented graphically in the search interface. When a query is submitted the interface maps the graphical representation of the query onto a CQL expression that is sent to the Karp web service. For example, in figure 1, we search for the word forms *torsk* ‘cod’ or *lång* ‘ling’ with an exclusion of adjectives (in order to avoid the adjective form *lång* ‘long DEF/PL’).

Search for entries that match

The screenshot shows a search interface with the following elements:

- A title: "Search for entries that match"
- A search box containing: "wordform" (dropdown), "equals" (dropdown), and "torsk" (text input).
- The word "or" below the first search box.
- A second search box containing: "wordform" (dropdown), "equals" (dropdown), and "lång" (text input).
- A plus sign (+) button below the second search box.
- An "except" dropdown menu.
- A third search box containing: "part of speech" (dropdown), "equals" (dropdown), and "adjective" (dropdown).
- A plus sign (+) button below the third search box.
- A plus sign (+) button below the "except" menu.
- A "Sort by" section with "wordform" (dropdown) and "ascending" (dropdown).
- A "Search" button at the bottom.

Figure 1: Extended search

⁹ <<http://www.loc.gov/standards/sru/specs/cql.html>>

4. Towards a generic lexicon editor

Even though all 23 lexicons currently in the Karp system are in LMF format, they contain, as shown in section 2, varying kinds of linguistic information. For example, some contain only morphological descriptions while others contain syntactic and semantic information of different kinds. In order to be useful, a general editing system has to provide synergies but still handle the particularities of each resource and not limit their expressiveness. The editing system should provide methodological support such as additional suggestions and consequence analysis, i.e. the effects one lexical judgement may have on related lexical information. For example, a new synonymy relation may trigger a suggestion in another resource or flag something as being in conflict with the new relation.

To elaborate further, access to the lexical information in all other resources while editing one resource provides rich background information for the lexicographer who is about to make a lexical decision. Moreover, formal inter/intra-resource dependencies can be verified on the fly, and new entries may be derived (semi-)automatically from other resources.

Statements of inter-resource dependencies also function as hypothesis testing: what intuitively seems true, and hence stated as a formal requirement, may instead illustrate important yet subtle differences in the resources. For example, it may seem intuitively evident that the frame hierarchy in the Swedish FrameNet should respect the hyponym relations in Swesaurus, such that if w_1 is a hyponym of w_2 , then w_1 should never occur higher in the frame hierarchy than w_2 . However, if this is a reasonable assumption or not is an empirical question.

Another important challenge is to allow lexical editing systems to take advantage of the massive amounts of linguistically-annotated text which are available in the corpora infrastructure Korp (Borin et al., 2012b) at Språkbanken, for example, when annotating examples or writing sense definitions. The information is, in principle, already available since the corpora have been annotated with lemmata and sense identifiers occurring in the resources of Karp, but it is still an open question how this information is best utilized and presented in the editor.

4.1 Editing the Swedish Constructicon

As an example of the current state of affairs of the lexicon editor, let us consider how editing of the Swedish Constructicon (Lyngfelt et al., 2012) is performed in Karp.

To start editing an existing lexicon entry in the Swedish Constructicon, the user has to log in and look up the particular entry using any of the provided search tools (see section 3). In the presentation view, the user clicks a button to open up a new editor tab with a slightly different presentation more suited for editing. Having different modes for normal presentation and for editing has the advantage that the editing mode can be generalized for all lexicons, while the presentation view may be tailored

for a specific audience or presentation style.

In the editor mode, the different fields of the lexicon are presented as a list. The exact presentation can be specified in a configuration file for each lexicon since the kinds of data can differ considerably. In figure 2, an entry of the Swedish Constructicon is being edited: the REFLEXIVE RESULTATIVE construction. It is a partially schematic construction expressed formally as VB REFL AP and semantically as ACTION ACTOR RESULT, with constructs such as *äta sig mätt* 'eat oneself full' and *skrika sig hes* 'shout oneself hoarse'. Moreover, the example section contains annotated authentic example sentences that illustrate the construction, e.g. *Drick dig smal i vår* 'Drink yourself thin this spring'. For more detailed information about this entry, please consult Lyngfelt et al. (2012).

All entries in this resource are quite similar in structure and have a small number of fields, making it sensible to also show the unfilled ones (toned down in gray). For other resources, however, the set of possible fields is much larger and there is more hierarchy in the entries. For such resources, fields can instead be added to the view upon request.

4.2 Technical details

The editing functionality of Karp is divided into three technical components: a backend with a REST-based web service API, a user authentication service, and a graphical frontend.

The backend recognizes a set of generic commands for adding, removing, updating lexical entries, and for manipulating the edit queue. A special **updates** command enables multiple actions at once, making it easier to log and backtrack changes by session.

Since the operations are generic, it is possible to use the Karp API for other applications as well. One such example is the language exercise platform Lärka 'lark' (Volodina et al., 2012) which uses the Karp editing API for logging user input.

The backend hosts two different database layers. The first one gets updated directly when a user edits a lexicon. If the user is fully authorized, the modified entries will eventually be copied to the main database layer, but only after they have been batch processed on a regular interval to ensure global consistency. If the user is not fully authorized, however, the changes will be put into the edit queue, waiting for a fully authorized user to accept or reject the changes.

Having multiple users working with the same lexicon may lead to the same problems as for any multi-user project. Changes may need to be undone while not altering other changes. Instead of reinventing the wheel, Karp makes use of an off-the-shelf version control system (VCS) inside the database. With each update the particular

lexicon change is checked into the VCS repository.

Although no two lexicons have exactly the same structure, they typically have certain traits in common that manifest themselves as similar frontend requirements, such as sharing the same settings or editing logic. For that purpose, Karp uses a class hierarchy for handling data structures in the frontend. The most basic class type is a string which is represented to the user as a simple text field. This can be extended to handle more complicated data structures, and modify the graphical user interface for editing the data. For example, the basic text widget can be subclassed to allow the user to select from a drop-down value list, that can be further subclassed to add consistency checks and other functionalities.

5. Conclusions and future work

We have briefly presented the ongoing work on adding editing functionality to the open lexical infrastructure at Språkbanken. It is still under active development, but is already a versatile tool for our work on the lexical resources.

The technical foundation is now in place, so our next step is to make all lexical resources of Karp editable. We will also explore the methodological details to ensure that the lexicographic work becomes as efficient as possible, and to secure the consistency and completeness of each resource by employing both internal and external lexical information.

6. Acknowledgements


The research presented here was supported by the Swedish Research Council (the projects Safeguarding the future of Språkbanken, VR dnr 2007-7430 and Swedish FrameNet++, VR dnr 2010-6013), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken (the Swedish Language Bank), and by the European Commission through its support of the META-NORD project under the ICT PSP Programme, grant agreement no 270899.


7. References

- Peter Andersson and Malin Ahlberg. 2013. Towards automatic tracking of lexical change: Linking historical lexical resources. NEALT Proceedings Series, 18.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In Proceedings of the 17th international conference on Computational linguistics, pages 86–90, Morristown, NJ, USA.
- Lars Borin and Markus Forsberg. 2008. Something old, something new: A computational morphological description of Old Swedish. In LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), pages 9–16.
- Lars Borin and Markus Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In Proceedings of the Nodalida 2009 Workshop on WordNets and

- other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies, Odense.
- Lars Borin and Markus Forsberg. 2010. From the People’s Synonym Dictionary to fuzzy synsets – first steps. In Proceedings of the LREC 2010 workshop Semantic relations. Theory and Applications, pages 18–25, Valletta. ELRA.
- Lars Borin and Markus Forsberg. 2011a. A diachronic computational lexical resource for 800 years of Swedish. In Language technology for cultural heritage, pages 41–61. Springer, Berlin.
- Lars Borin and Markus Forsberg. 2011b. Swesaurus – ett svenskt ordnät med fria tyglar. *LexicoNordica*, 18:17–39.
- Lars Borin, Dana Danélls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010a. The past meets the present in Swedish FrameNet++. In 14th EURALEX International Congress, pages 269–281, Leeuwarden. EURALEX.
- Lars Borin, Markus Forsberg, and Dimitrios Kokkinakis. 2010b. Diabase: Towards a diachronic BLARK in support of historical studies. In Proceedings of LREC 2010, pages 35–42, Valletta. ELRA.
- Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012a. The open lexical infrastructure of Språkbanken. In Proceedings of LREC 2012, pages 3598–3602, Istanbul. ELRA.
- Lars Borin, Markus Forsberg, and Johan Roxendal. 2012b. Korp – the corpus infrastructure of Språkbanken. In Proceedings of LREC 2012, Istanbul. ELRA.
- Lars Borin, Bernard Comrie, and Anju Saxena. 2013a. The Intercontinental Dictionary Series – a rich and principled database for language comparison. In Lars Borin and Anju Saxena (eds.), *Approaches to measuring linguistic differences*, pages 285–302. Mouton de Gruyter, Berlin.
- Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013b. SALDO – a touch of yin to WordNet’s yang. *Language Resources and Evaluation*.
- Lars Borin. 2010. Med Zipf mot framtiden – en integrerad lexikonresurs för svensk språkteknologi. *LexicoNordica*, 17:35–54.
- Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to WordNet. In Proceedings of the Third International WordNet Conference.
- Christiane Fellbaum. 1998a. A semantic network of English verbs. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 69–104. MIT Press, Cambridge, Mass.
- Christiane Fellbaum, editor. 1998b. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.
- Gil Francopoulo, editor. 2013. *LMF: Lexical Markup Framework*. ISTE/Wiley, London/Hoboken, NJ.
- Martin Haspelmath and Uri Tadmor, editors. 2009. *Loanwords in the World’s Languages: A Comparative Handbook*. Mouton de Gruyter, Berlin.
- ISO. 2008. *Language resource management – Lexical Markup Framework (LMF)*. International Standard ISO 24613:2008.

- Viggo Kann and Magnus Rosell. 2006. Free construction of a free Swedish dictionary of synonyms. In Proceedings of the 15th NODALIDA conference, Joensuu 2005, pages 105–110. Department of Linguistics, University of Joensuu.
- Benjamin Lyngfelt, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldbberg, and Sofia Tingsell. 2012. Adding a constructicon to the Swedish resource network of Språkbanken. In Proceedings of KONVENS 2012 (LexSem 2012 workshop), pages 452–461, Vienna.
- Alistair Miles and José R. Pérez-Agüera. 2007. SKOS: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3):69–83.
- Josef Ruppenhofer, Michael Ellsworth, R. L. Miriam Petruck, R. Christopher Johnson, and Jan Scheffczyk. 2005. *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.
- C.J. Schlyter. 1887. *Ordbok till Samlingen af Sweriges Gamla Lagar*. (Saml. af Sweriges Gamla Lagar 13). Lund, Sweden.
- Elena Volodina, Lars Borin, Hrafn Loftsson, Birna Arnbjörnsdóttir, and Guðmundur Örn Leifsson. 2012. Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use. In Proceedings of the SLTC 2012 Workshop on NLP for CALL, Lund, 25th October, 2012, number 80 in Linköping Electronic Conference Proceedings, pages 47–58.

✖  **ID** reflexiv_resultativ

Type ▾   cx, refl



✖  **Cat** VP

✖  **Evokes** Causation_scenario

✖  **Definition** [Någon]_{Actor} eller [något]_{Theme} utför eller undergår [en aktion]_{Activity} som leder (eller antas leda) till att [aktören]_{Actor} / [temat]_{Theme}, uttryckt med reflexiv, uppnår [ett tillstånd]_{Result}.

✖  **Structure** [V refl AP]

 Inheritance

Cee ▾   refl



Coll ▾   {äta¹ : mätt¹}

  {supa¹ : full²}

  {skrika¹ : hes¹}

  springa¹



Internal construction elements ▾   name=Activity, cat=vb

  cx=refl, name=Actor

  cx=refl, name=Theme

  name=Result, cat=AP



External construction elements ▾   name=Actor, cat=NP

  name=Theme, cat=NP




Example ▾   [Vi åskådare]_{Actor} [[springer]_{Activity} [oss]_{Actor} inte [varma]_{Result}]resultativ_reflexiv direkt.

  [Kornet och havren]_{Theme} får [[frysa]_{Activity} [sig]_{Theme} [mogen]_{Result}]resultativ_reflexiv ·

  [[Drick]_{Activity} [dig]_{Actor} [smal]_{Result}]resultativ_reflexiv i vår.



✖  **Comment** Det finns också en PP-variant med resultativ betydelse, t.ex. "träna sig i form", som ev. bör inkorporeras här - alt. betraktas som en metaforisk utvidgning av någon rörelse-cx.

 Internal comment (hidden)

✖  **Reference** Jansson, Håkan (2006): Har du ölat dig odödlig? En undersökning av resultativkonstruktioner i svenskan. (D-uppsats, även publicerad som MISS 57) <http://hdl.handle.net/2077/19000> Lyngfelt, Benjamin (2007): Mellan polerna. Reflexiv- och deponenskonstruktioner i svenskan. Språk och stil NF 17: 86–134. <http://hdl.handle.net/2077/21731>

 Status

Figure 2: Editing the SweCxn entry reflexiv_resultativ in Karp