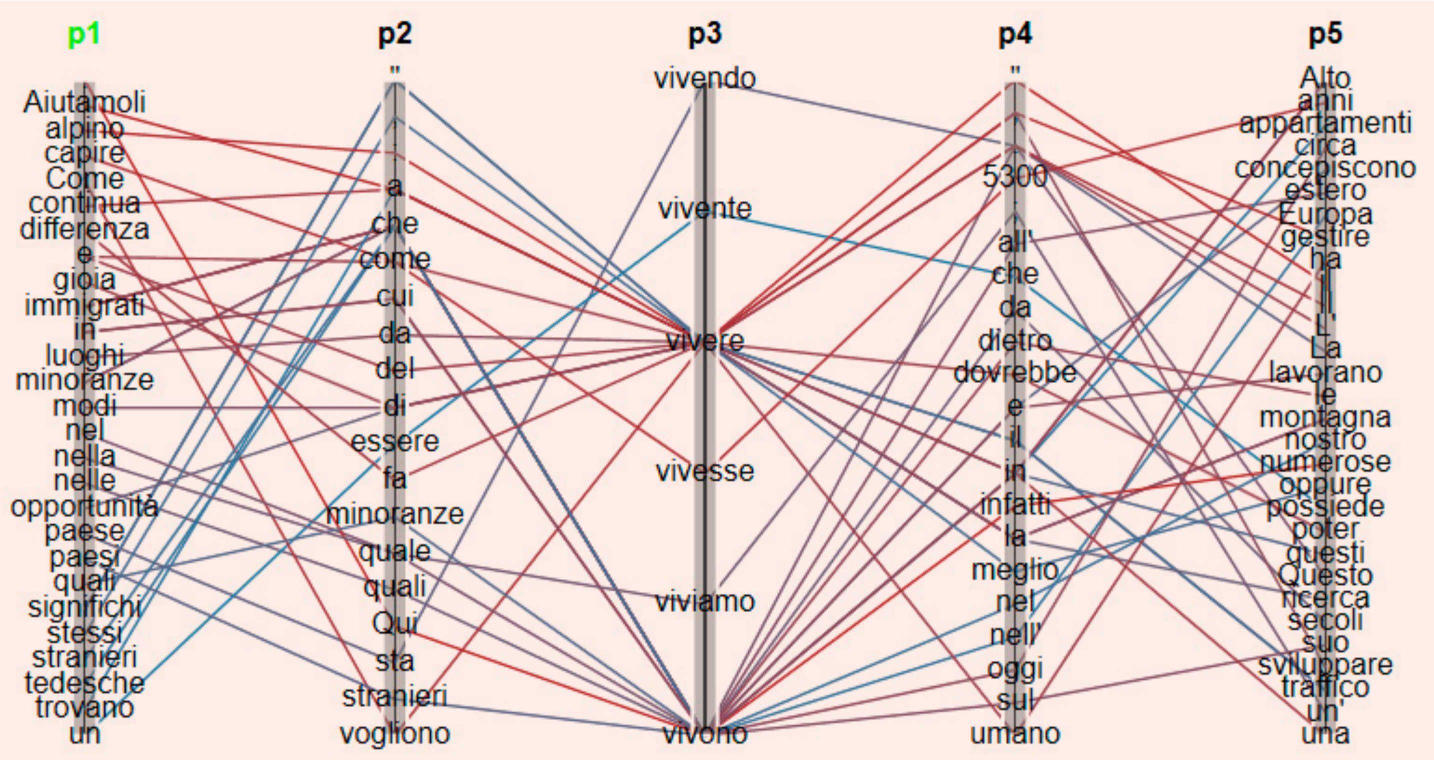


Proceedings of the XVI EURALEX International Congress:

The User in Focus

15-19 July 2014, Bolzano/Bozen

Edited by Andrea Abel, Chiara Vettori, Natascia Ralli

Proceedings of the XVI EURALEX International Congress: The User in Focus

Acknowledgements

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

A.S. Hornby Educational Trust
Accademia della Crusca (scientific patronage)
Autonomous Province of Bolzano/Bozen (patronage)
Consortium Südtiroler Speck Alto Adige
Dr. Schär AG/SPA
Istitut Ladin Micurà de Rü
k Dictionaries
Municipality of Bolzano/Bozen (patronage)
Oxford University Press
SDL
Sketch Engine
Sprachstelle im Südtiroler Kulturinstitut
Südtirol Marketing (SMG)
Verband der Südtiroler Obstgenossenschaften (VOG)

Programme Committee

Andrea Abel (EURAC)
Janet DeCesaris (Universitat Pompeu Fabra)
Tinatin Margalitadze (Ivane Javakhishvili Tbilisi State University - Next EURALEX Congress Organizer)
Natascia Ralli (EURAC)
Marcello Soffritti (Università di Bologna/EURAC)
Pius ten Hacken (Universität Innsbruck)
Ruth Vatvedt Fjeld (Universitetet i Oslo)
Chiara Vettori (EURAC)

Reviewers (with number of reviewed papers)

Andrea Abel (EURAC, 18), Hauke Bartels (Serbski Institut, 5), Maria Paz Battaner Arias (Catedrática jubilada - Universitat Pompeu Fabra, 10), Patrizia CORDIN (Università degli Studi di Trento, 8), Janet DeCesaris (Universitat Pompeu Fabra, 19), Anne Dykstra (Fryske Akademy, 15), Ruth Vatvedt Fjeld (Universitetet i Oslo, 13), Thierry Fontenelle (Translation Centre for the Bodies of the European Union, 5), Alexander Geyken (Berlin-Brandenburgische Akademie der Wissenschaften, 7), Rufus Gouws (Universiteit Stellenbosch University, 7), Sylviane Granger (Université catholique de Louvain, 3), Reinhard Hartmann (University of Exeter, emeritus, 5), Ulrich Heid (Universität Hildesheim, 6), Adam Kilgarriff (Lexical Computing Ltd, 7), Annette Klosa (Institut für Deutsche Sprache, Mannheim, 8), Simon Krek (Jozef Stefan Institute & Amebis, 7), Iztok Kosem (Trojina, Institute for Applied Slovene Studies, 14), Lothar Lemnitzer (Berlin-Brandenburgische Akademie der Wissenschaften, 8), Robert Lew (Uniwersytet im. Adama Mickiewicza w Poznaniu, 7), Carla Marello (Università di Torino, 8), Tinatin Margalitadze (Ivane Javakhishvili Tbilisi State University, 11), Carolin Müller-Spitzer (Institut für Deutsche Sprache, Mannheim, 8), Martina Nied Curcio (Università degli Studi Roma Tre, 7), Magali Paquot (Université catholique de Louvain, 8), Natascia Ralli (EURAC, 21), Michael Rundell (Lexicography MasterClass, 9), Felix San Vicente (Università di Bologna, 9), Marcello Soffritti (Università di Bologna/EURAC, 14), Joan Soler Bou (Institut d'Estudis Catalans, 9), Pius ten Hacken (Universität Innsbruck, 19), Carole Tiberius (Instituut voor Nederlandse Lexicologie, 7), Lars Trap Jensen (Det Danske Sprog- og Litteraturselskab, 16), Serge Verlinde (KU Leuven, 7), Chiara Vettori (EURAC, 22), Geoffrey Williams (Université de Bretagne Sud, 16).

Proceedings of the XVI EURALEX International Congress: The User in Focus

Proceedings of the XVI EURALEX International Congress:

The User in Focus

15-19 July 2014, Bolzano/Bozen

Edited by Andrea Abel, Chiara Vettori, Natascia Ralli

Hornby
A. S. Hornby Educational Trust

Accademia della Crusca

AUTONOME PROVINZ BOZEN SÜDTIROL
PROVINCIA AUTONOMA DI BOLZANO ALTO ADIGE
PROVINZIA AUTONOMA DE BULSAN SÜDTIROL

Speck Alto Adige I.G.P. Südtiroler Speck g.g.A.
Südtirol
Indicazione Geografica Protetta Geschützte Geographische Angabe

ISTITUT LADIN MICURÀ DE RÜ

KDICTIONARIES

Città di Bolzano
Stadt Bozen

OXFORD
UNIVERSITY PRESS

SDL

Sketch Engine

Sprachstelle
im Südtiroler KULTURinstitut

Südtirol

Marlene®
Südtirol

Institute for Specialised Communication and Multilingualism
Viale Druso/Drususallee 1, 39100 Bolzano/Bozen
Tel. +39 0471 055100
Fax +39 0471 055199
communication.multilingualism@eurac.edu
www.eurac.edu

The digital copy of the Proceedings of the XVI EURALEX International Congress can be downloaded at the following URL
http://www.eurac.edu/en/research/institutes/multilingualism/Publications

Director: Stephan Ortner
Prepress: Pluristamp
Press: EURAC research

ISBN: 978-88-88906-97-3

Foreword

The XVI edition of the EURALEX International Congress took place in Italy at the European Academy of Bolzano/Bozen (EURAC) from 15 to 19 July 2014. On behalf of the organising committee, I am extremely pleased to present this volume of proceedings. A novelty of this latest EURALEX edition is our tribute to the digital age: the proceedings collect selected contributions presented at the congress and are published exclusively in digital format on the Internet.

EURALEX congresses take place every second year and usually attract a large international audience. This was true also for the XVI edition. You will find that the authors contained in this volume come from all over the world and have drafted their papers in several languages. Even though English remains the most popular, we have received contributions written in other major European languages, such as French, German, Italian and Spanish. The number of languages that the authors treat within their papers is even higher and goes well beyond the frontiers of the European continent, covering a wide range of languages.

Being a European organisation devoted to lexicography and related fields, EURALEX is very sensitive to issues of language diversity. Regional and minority languages on the one hand and large international languages on the other hand therefore receive equal consideration. In this respect, South Tyrol (called *Alto Adige* in Italian and *Südtirol* in German) represents the ideal location for an edition of the EURALEX congress. The region being officially trilingual, South Tyrol is a perfect example of language diversity. Italian and German are both official languages of the province of Bolzano/Bozen, but also the local minority language Ladin has gained official recognition. Moreover, South Tyrol is becoming growingly multilingual and multicultural, thus adding new languages from all over the world to an already quite colourful linguistic landscape.

EURAC's Institute for Specialised Communication and Multilingualism (ISCM) was particularly honoured to host the XVI EURALEX International Congress. The main aim of our institute's research is to provide scientific answers to current issues of language and education policy as well as to economic and social questions. Our activities at local and international level comprise applied research (research projects and networking), training and consulting (consulting services, monitoring, seminars) as well as dissemination (scientific publications, dictionaries, databases, corpora) in three main research areas: bilingualism and multilingualism, specialised communication and language technologies. Our interests range from general to special languages, from old to new minorities and from language policy to language planning. Our research activities also centre on the observation of language usage, language documentation, language consulting, multilingual knowledge management and the management of linguistic and cultural diversity. To support our research initiatives, we also create and develop dedicated research infrastructures.

Against this background, at the ISCM we have developed long-term experience and in-depth expertise into different types of lexicographic activities concerning both general language and special language. We particularly focus on terminology and electronic learner lexicography, but have also produced reference works for the Italian Sign Language and Ladin terminology. A pioneering feature of our work consists in applying new technologies to lexicography, as in the Information system for legal terminology bistro, the electronic learner dictionary for German and Italian eldit and the first bilingual electronic dictionary Italian Sign Language - Italian e-LIS. To this end, we have created and developed resources to support lexicography and terminography, such as online corpora (e.g. Korpus Südtirol and PAISÀ) and tools for visualising linguistic data (LinfoVis).

Our researchers at the ISCM are active in large international projects and networks, such as the academic DFG-network "Internet Lexicography" and the COST action "European Network of e-Lexicography". They are members of various international organisations like the Council for German language terminology RaDT and the International Information Centre for Terminology Infoterm. Given such fitting background, the EURAC Institute for Specialised Communication and Multilingualism seemed particularly appropriate as a host and organiser of a EURALEX International Congress.

The 2014 edition was driven by the motto "the user in focus", since one of the challenges of lexicographic products is to respond to specific and diverse user needs. Therefore, we particularly welcomed submissions focusing on the user perspective at all levels of dictionary production and consultation (e.g. experts vs. laymen as users and/or as producers, professional lexicographers as users of dedicated lexicographic tools, user needs before and during dictionary consultation, considering user needs during the whole lexicographic process, groups with specific needs). Nevertheless, we did not wish to exclude any areas of potential interest for lexicography, intending to cover – as for all preceding editions of the EURALEX congress – a large and varied array of relevant and current themes, which you consequently find in the present proceedings.

This volume contains the four plenary lectures given at the congress. The plenaries represent significant contributions by well-known experts in their fields. Thierry Fontenelle of the Translation Centre for the Bodies of the European Union in Luxembourg argued for considering the space between lexicography and terminology as a continuum rather than a hard-and-fast dichotomy. Ulrich Heid of the Universität Hildesheim analysed natural language processing techniques with respect to their impact on the user friendliness of electronic dictionaries. Carla Marello of the Università di Torino discussed whether the reference skills of native digital EFL students are developed well enough for them to take advantage of large bilingual dictionaries on their smartphones. Rosamund Moon of the University of Birmingham explored ideological meanings in learner dictionaries with a particular focus on age and ageism. As in past EURALEX editions, the Hornby Trust generously sponsored one of the plenary lectures (R. Moon) in honour of A.S. Hornby, a pioneering figure in learners' dictionaries for non-native speakers. The organising committee would like to extend its sincere thanks to all plenary speakers for setting the tone of the congress and of this volume, which we believe represents a significant contribution to the literature of dictionary making theory and practice. Also, we were particularly glad that they all most excellently considered and discussed the motto of the XVI edition.

Submissions to EURALEX 2014 were filed under several categories: full papers, short papers, posters and software demonstrations. The contributions in these proceedings are grouped according to the following sections, which set the frame for the call-for-papers:

- The Dictionary-Making Process
- Research on Dictionary Use
- Lexicography and Language Technologies
- Lexicography and Corpus Linguistics
- Bi- and Multilingual Lexicography
- Lexicography for Specialised Languages, Terminology and Terminography
- Lexicography of Lesser Used languages
- Phraseology and Collocation
- Historical Lexicography and Etymology
- Lexicological Issues of Lexicographical Relevance
- Reports on Lexicographical and Lexicological Projects
- Others

Within each section, the papers are organised alphabetically by the surname of first author.

EURALEX Congress proceedings have become an important reference for dictionary research, based on the high quality of the papers selected. During the review process every submitted abstract was evaluated by two independent blind referees. In case of doubt, a third independent opinion was sought. This process lead to rejecting about one third (35%) of the papers originally submitted.

On behalf of everyone associated with the organisation of EURALEX 2014 at EURAC, I would like to express our gratitude to all the contributors for submitting relevant and interesting work as well as for meeting the tight production schedule of this online publication. I personally would like to thank also all the colleagues who participated in the review process and those who joined me on the EURALEX 2014 programme committee to jointly prepare the final congress programme. The generous patrons and sponsors who supported us for this edition are all listed on a dedicated page within these proceedings. Last but not least, I am particularly indebted to the three members of the organising committee who joined efforts with me to make EURALEX 2014 a successful event: Chiara Vettori, Natascia Ralli and Daniela Gasser from EURAC. Their dedication and constant commitment deserve a special mention. Chiara Vettori is the main responsible for the timely production of these online proceedings. It is a pleasure to acknowledge their precious work here, but also that of all staff who contributed to a successful congress week and who cannot be named here.

Andrea Abel

Chair, XVI EURALEX International Congress

May, 2014

Index

Lexicography and Corpus Linguistics 415

Bi-and Multilingual Lexicography 493

Lexicography for Specialised Languages, Technology and Terminography 599

Lexicography of Lesser Used Languages 737

Phraseology and Collocation **837**

Historical Lexicography and Etymology **939**

Plenary Lectures

From Lexicography to Terminology: a Cline, not a Dichotomy

Thierry Fontenelle
Translation Centre for the Bodies of the European Union, Luxembourg
thierry.fontenelle@cdt.europa.eu

Abstract

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the OED's problematic claim to be the "definitive record of the English language". He argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

In this paper, I would like to extend the parallel made by ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation. In the course of their careers, translators will indeed need to make use of a whole range of dictionaries, starting from learners' dictionaries when they learn foreign languages, to monolingual dictionaries and bilingual dictionaries to learn translation techniques, to term bases as soon as they start translating highly specialized and technical texts. We will focus on terminology databases such as IATE, the European Union's interinstitutional term base, which is the natural tool to which they turn to obtain information about technical terms in the medical, legal, environmental, chemical fields, to cite only a few domains covered by this resource. With 8.7 million terms covering the 24 official languages of the European Union, including 1.4 million English terms and half a million abbreviations, this database is a highly popular tool in the translation world (44 million queries in 2013).

In addition to general term bases such as IATE, we will also discuss other specialized EU terminological databases such as ECHA-Term, a term base compiled for the European Chemicals Agency (ECHA) to help industry comply with the legal requirements of the REACH Directive and of the regulation on the classification, labeling and packaging of chemicals. We will show how user requirements have been taken into account to meet the needs of the users of the database, who resort to ECHA-Term to obtain reliable, coherent and up-to-date multilingual terminology in the chemicals field, a sine qua non for clear specialized communication. The description of these databases will make it clear that the distinction between 'traditional' lexicography and terminology is more a cline than a dichotomy, insofar as the types of linguistic information included in the respective products created by both disciplines all correspond to the specific needs of their users.

Keywords: term banks; terminological database; translation; European Union; LSP dictionaries; IATE; ECHA-term

1 Introduction

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the Oxford English Dictionary's problematic claim to be the "definitive record of the English language". He points out that the OED is often regarded as authoritative and that one of the aspects of authority is the comprehensive lexical coverage of the dictionary. Yet, he argues, even if lexicographers such as Simpson (2000:1) call the OED "the principal dictionary of record for the English language" , there is in fact no empirical entity corresponding to "the English language" for which the OED could be taken as a description (ten Hacken 2012: 838). Simpson himself is aware of the impossibility of providing a comprehensive coverage in a dictionary. It is therefore a myth to assume that a dictionary should contain every word. A dictionary can therefore only be a partial record and it is unrealistic to assume that a dictionary can provide a full record of a language.

Ten Hacken argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

It is interesting to note that the controversy around the comprehensive nature of a dictionary such as the OED usually involves a comparison with other general-purpose monolingual dictionaries, including learners' dictionaries. The inclusion of usage notes is also used as a criterion to distinguish descriptive and prescriptive dictionaries. Surprisingly, no mention is usually ever made of a different type of dictionary, namely terminological databases, which should also be seen as specific types of dictionaries designed to solve specific types of linguistic problems. Why is that nobody ever questions the comprehensiveness of a terminological database? Why would anybody expect the OED to include "all possible" words of the English language, while recognizing that the list of acronyms and abbreviations in a language is potentially infinite and that even a huge term base such as IATE, the interactive terminology database of the European Union described below, can only provide a partial record of specialized terminology, even with 1.4 million English terms and half a million abbreviations and acronyms?

2 From Learner's Dictionaries to Mono- and Bilingual Dictionaries to Terminological Bases

I would like to extend the parallel made by ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation.

If, as proposed by Ten Hacken (2012, 2009), dictionaries are tools with which users solve problems of a particular type, the various types of dictionaries available on the market actually correspond to the range of problems users are faced with at different moments of their career. A teenager or a university student who learns a foreign language will most probably require a small bilingual dictionary at the beginning of the learning process because a learner's dictionary can only be used by someone who is not a total beginner in this foreign language. Once knowledge of the foreign language reaches a certain level, the student will be encouraged to make use of a learners' dictionary, which will provide useful information in a decoding and an encoding perspective, thanks to its simplified definitions, its system of grammar codes, its illustrative examples, etc. General-purpose monolingual dictionaries target a different kind of public, made up of advanced native speakers or of non-native speakers who have an in-depth knowledge of the language of the dictionary. Historical dictionaries such as the OED, with their focus on etymology and the evolution of words, are yet for other users who expect the dictionary to provide descriptive records of the development and use of words over time.

A parallel may be drawn with the tools used by translators. At the beginning of their career, students in translation will primarily make use of general-coverage bilingual dictionaries which will provide them with information about collocations, idioms, sense distinctions, etc. The role of translations in bilingual dictionaries is to provide target-language equivalents of the source-language headword (see also Fontenelle forthcoming). At a later stage, however, seasoned translators will tend to consult their bilingual dictionaries less and less, and will turn more frequently to terminological databases (a.k.a. term banks), which will enable them to translate highly-specialized texts and to make communication possible between specialists, or between specialists and the general public. Such tools have become a sine qua non in our multilingual world where access to technical information across multiple domains is a must.

3 Terminology and Term Banks

Understanding terminology, i.e. the specialized vocabulary which is used in a specific domain, is a key element in communication. This poses a number of challenges in the case of translation in a multilingual context, since knowing the exact meaning of a technical term is necessary to understand a text, but also to reproduce the text as faithfully as possible in another language. It therefore no surprise that the various translation services of the European Union have traditionally dedicated significant resources to the compilation of terminological information in the official languages of the EU. In order to describe the vocabulary of special subject fields, terminologists create term banks, which are compilations of the collections of words associated with a given domain. A terminology database will then be seen as a repository of descriptions of concepts, which are seen as mental constructs which are distinct from the terms they correspond to in a given language (see also Fontenelle and Rummel, forthcoming). The traditional approach assumes that terms can be organised into networks of concepts to structure a given domain. This is indeed well-suited for normalization, but, as is pointed out by Jacquemin and Bourrigault (2003), there seems to be a flaw in this reasoning, because this approach is not really suitable for computational term analysis. A terminologist indeed bases his or her work upon the analysis of textual data (corpora) and the term base is actually the result of this analysis, and not the result of some introspection whereby abstract conceptual maps would be derived.

Even if terminological databases are traditionally seen as distinct from dictionaries, it cannot be denied that terminological entries have a lot in common with dictionary entries. Of course, at the macrostructural level, it is clear that some entries found in a traditional dictionary will not be found in a term base: some parts of speech will be absent from term banks. Prepositions or adverbs for instance will most probably not be found in term banks. Even verbs are traditionally underrepresented in such resources because the vast majority of terms are noun phrases. This is not enough to consider that a term base is not a dictionary, however. After all, rare scientific words will also be excluded from learners' dictionaries. At the microstructural level, definitions will be as essential in term bases as in a traditional monolingual dictionary and the NLP community has always been interested in how candidate terms could be extracted from corpora, together with possible definitions (see Person 1998, who proposed an analysis of the defining mechanisms signalling the presence of a term in a corpus, using linguistic patterns such as 'X is known as Y' or "X is called Y" to link a definiens and a definiendum).

4 IATE: The European Union Terminological Database

4.1 An Interinstitutional Database

IATE stands for 'Inter-Active Terminology for Europe' and is the term base of the language services of the European Union. This concept-oriented, large-scale multilingual database covers all fields of activity of the European Union. IATE was initially launched in 1999 by the Translation Centre for the Bodies of the European Union, located in Luxembourg. Today, the Translation Centre manages the technical aspects of the project on behalf of the project partners: the European Commission, the European Parliament, the Council, the Court of Auditors, the Court of Justice of the European Union, the European Investment Bank, the European Central Bank, the Economic and Social Committee, the Committee of the Regions and the Translation Centre. Before the launch of the project and its opening to the public in 2007, nearly every institution had its own term base (Fontenelle and Mergen 1998; Reichling 1998), while, today, IATE can be seen as the shared terminology database of all existing terminology collections of the translation services of all EU institutions and bodies. It can be consulted free of charge at http://iate.europa.eu.

IATE can be searched for specific terms or abbreviations in a given source language and for its equivalent(s) in any of the 23 other languages (IATE contains mainly terminology in the 24 official languages of the EU, as well as some content in non-official languages). After the accession of Croatia on 1 July 2013, the 24 official languages are: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish.

The database is constantly updated by the terminologists and translators of the various participating institutions. In 2013, around 97 000 new terms were added and over 150 000 existing terms were modified. The contents of the database as of 1 January 2014 can be broken down as in Table 1.

Language	Number of terms
en – English	1402006
fr – French	1339496
de – German	1034088
it – Italian	701735
nl – Dutch	691801
es – Spanish	617124
da – Danish	603481
pt – Portuguese	533623
el – Greek	523086
fi – Finnish	329491
sv – Swedish	314220
la – Latin	64159
pl – Polish	59576
ga – Irish	57879
lt – Lithuanian	53802
hu – Hungarian	49858
et – Estonian	41489
sl – Slovenian	41337
cs – Czech	38043
sk – Slovak	37219
mt – Maltese	35732
ro – Romanian	35451
bg – Bulgarian	34420
lv – Latvian	28617
hr – Croatian	8863

Table 1: IATE: Number of terms (1/1/2014).

The table reflects the history of the European Union and its successive enlargements, of course. Altogether, the IATE database contained 8,705,334 terms on 1/1/2014. The languages with the most terms were the European Union's most often used working languages, viz. English, with more than 1.4 million terms in first place, followed by French (1.3 million terms) and then German (1 million terms). It is interesting to note that Latin is also fairly well-represented, with 64159 terms: taxonomies of animal and vegetal species are crucial for translators who deal with the translation of texts related to the Common Fisheries Policy or the Common Agricultural Policy, as well as texts written by the European Environment Agency, the Community Plant Variety Office or the European Food Safety Agency.

The public version of IATE received 44 million queries in 2013 and the number of queries in the internal version of IATE (accessible only to EU staff) amounted to 14,206,137 (vs. 11,178,323 queries in 2012). Queries in the public version of IATE came from all over the world, from France to Somalia, Argentina to South Korea. The country where the most queries originated from in 2013 was Italy, followed by France, Spain, Germany, Belgium, Greece, Portugal, the United Kingdom, the Netherlands and Switzerland.

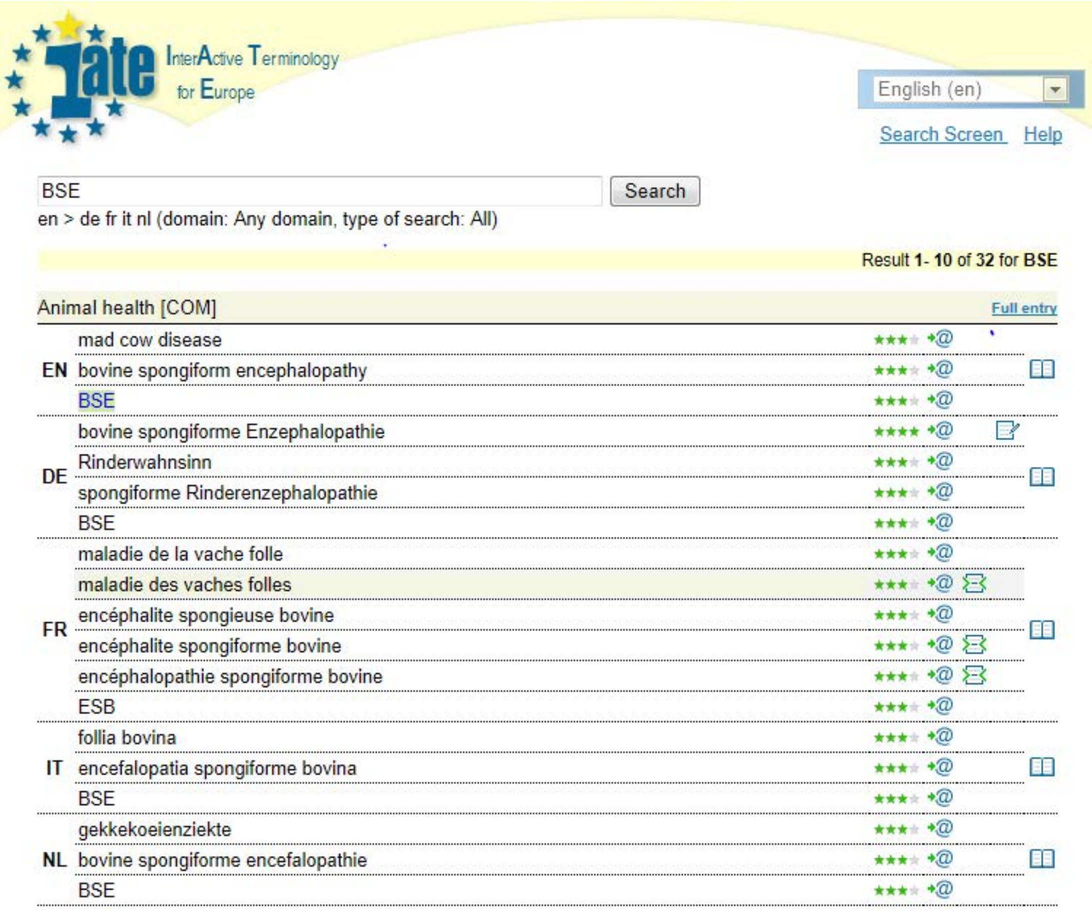
Terms in IATE have a fairly standard data structure. One of the challenges which the creators of the database faced was the mapping rules between the data structures of the existing databases and the new format of this interinstitutional database. A concept-oriented approach was adopted to express the various aspects of concepts via a series of three interrelated levels:

(1) a language-independent top level, containing information pertaining to the whole concept. Information about domains is a case in point (i.e. the field of knowledge in which the concept is used). Other types of information can also be stored at that level, including pictures or images;

(2) An intermediate 'language' level for definitions, explanations and comments, which can be stored for each of the languages of the terminological record;

(3) The 'term' sub-level, at which several terms can be distinguished to store synonyms of a given concept or abbreviations. Reliability codes, term references and the date when a record was last edited will typically appear at that level.

For instance, *BSE, bovine spongiform encephalopathy* and *mad cow disease* are three distinct terms that are synonyms and refer to the same concept. The definitions would be coded at level 2 (the three terms will have the same definition in a given language), while the domain (Animal health) is at level 1, the conceptual level, as is illustrated in Figure 1 and in Figure 2 below.



BSE

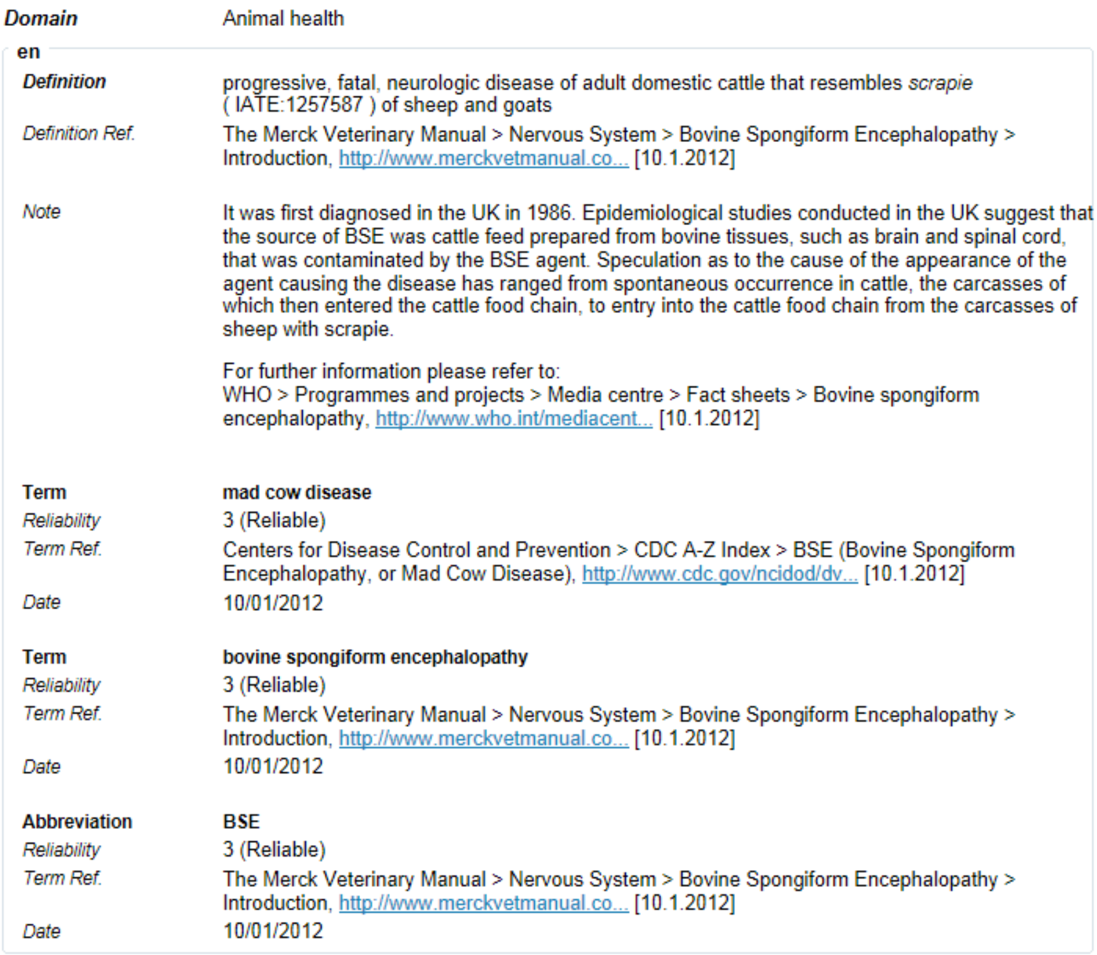
en > de fr it nl (domain: Any domain, type of search: All)

Result 1- 10 of 32 for BSE

Animal health [COM]

EN	mad cow disease	★★★
	bovine spongiform encephalopathy	★★★
	BSE	★★★
DE	bovine spongiforme Enzephalopathie	★★★★
	Rinderwahnsinn	★★★
	spongiforme Rinderenzephalopathie	★★★
	BSE	★★★
FR	maladie de la vache folle	★★★
	maladie des vaches folles	★★★
	encéphalite spongieuse bovine	★★★
	encéphalite spongiforme bovine	★★★
	encéphalopathie spongiforme bovine	★★★
	ESB	★★★
IT	follia bovina	★★★
	encefalopatia spongiforme bovina	★★★
	BSE	★★★
NL	gekkekoeienziekte	★★★
	bovine spongiforme encefalopathie	★★★
	BSE	★★★

Figure 1: Query on BSE in IATE.



Domain	Animal health
en	
Definition	progressive, fatal, neurologic disease of adult domestic cattle that resembles *scrapie* (IATE:1257587) of sheep and goats
Definition Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Note	It was first diagnosed in the UK in 1986. Epidemiological studies conducted in the UK suggest that the source of BSE was cattle feed prepared from bovine tissues, such as brain and spinal cord, that was contaminated by the BSE agent. Speculation as to the cause of the appearance of the agent causing the disease has ranged from spontaneous occurrence in cattle, the carcasses of which then entered the cattle food chain, to entry into the cattle food chain from the carcasses of sheep with scrapie. For further information please refer to: WHO > Programmes and projects > Media centre > Fact sheets > Bovine spongiform encephalopathy, http://www.who.int/mediacent... [10.1.2012]
Term	**mad cow disease**
Reliability	3 (Reliable)
Term Ref.	Centers for Disease Control and Prevention > CDC A-Z Index > BSE (Bovine Spongiform Encephalopathy, or Mad Cow Disease), http://www.cdc.gov/ncidod/dv... [10.1.2012]
Date	10/01/2012
Term	**bovine spongiform encephalopathy**
Reliability	3 (Reliable)
Term Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Date	10/01/2012
Abbreviation	**BSE**
Reliability	3 (Reliable)
Term Ref.	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, http://www.merckvetmanual.co... [10.1.2012]
Date	10/01/2012

Figure 2: set of synonyms for bovine spongiform encephalopathy.

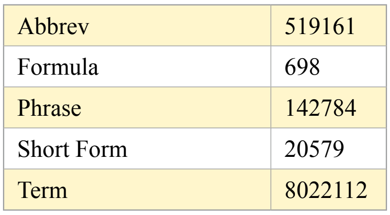
The system offers today the following features (see also Fontenelle and Rummel in press):

- One common database for all institutions and agencies containing all legacy data;
- Basic and advanced search features (including stemming and base character conversion);
- On-line access in read and write mode, i.e. the possibility for users to carry out modifications, to add entries directly to the central database and hence to allow their colleagues to benefit from this work in real time;
- A validation workflow that ensures that all newly added or modified terminology is reviewed;
- Role-based user management;
- Auditing features that keep track of all changes made to the terminology in the database;
- Features for the export and import of data;
- Statistics on the content of the database and user activity;
- A basic messaging system as communication mechanism between the actors in the terminology workflow.

In 2014, a new functionality will also be offered to allow users to download or copy the contents of the IATE database which is not protected by third-party copyrights, for research or for commercial purposes.

4.2 Language and Terminology: A Dynamic Organism

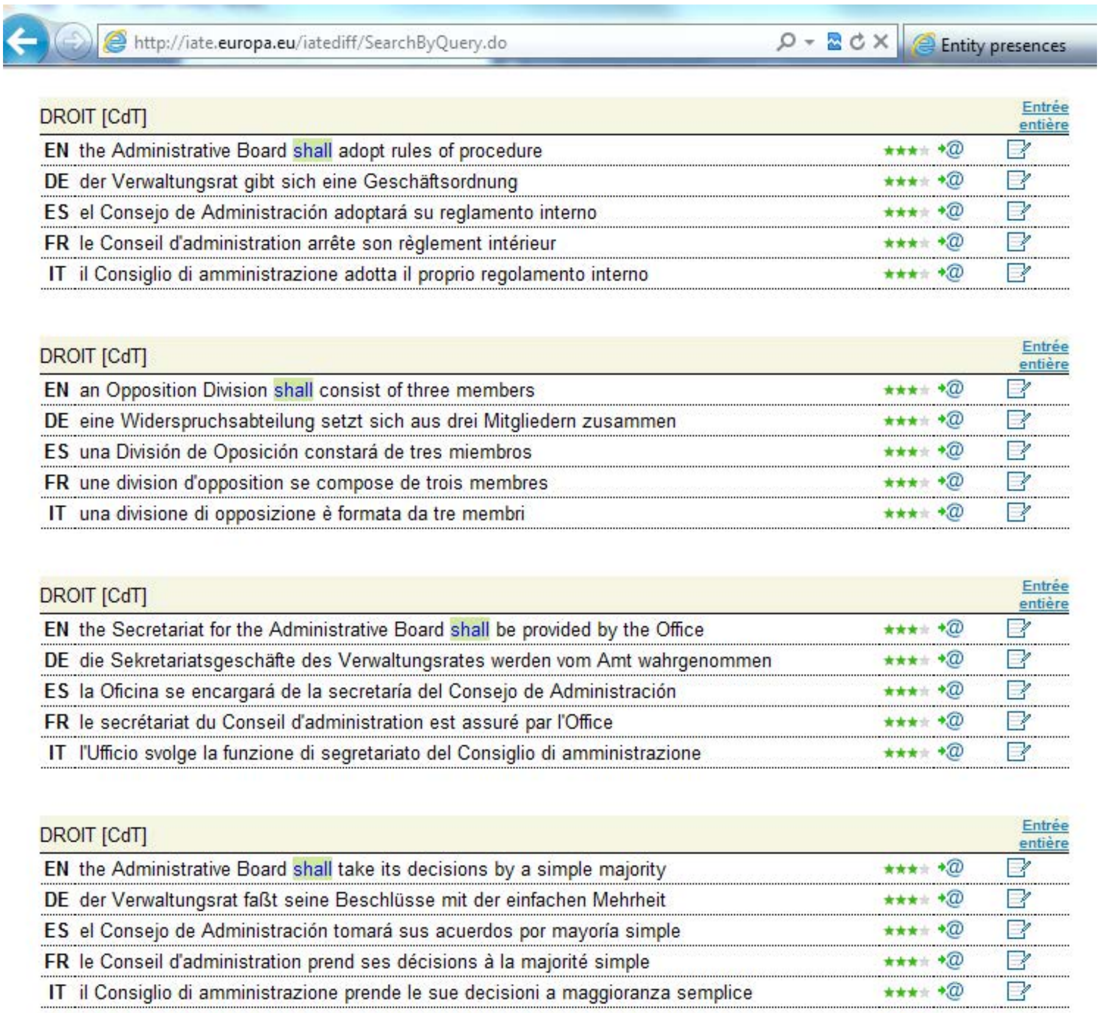
Languages are dynamic organisms. New terms are created every day (nobody was talking about 3D printers five years ago and the translation of selfie is a hot topic in many linguistic communities in 2014). It is therefore not surprising to see that IATE is not just concerned with terms proper, but also with abbreviations and acronyms, which are condensed versions of often long and complex terms. Table 2 shows the contents of the IATE database per term type.



Abbrev	519161
Formula	698
Phrase	142784
Short Form	20579
Term	8022112

Table 2: Number of terms per type in IATE.

With more than half a million abbreviations, one can immediately see that the question of the 'completeness' of a dictionary discussed in the context of the OED above is a myth, indeed, and a quick glance at the list of abbreviations included in such a database will convince anybody that there cannot be such a thing as a "definitive record" of any language.



DROIT [CdT]			Entrée entière
EN	the Administrative Board shall adopt rules of procedure	★★★☆ +@	✎
DE	der Verwaltungsrat gibt sich eine Geschäftsordnung	★★★☆ +@	✎
ES	el Consejo de Administración adoptará su reglamento interno	★★★☆ +@	✎
FR	le Conseil d'administration arrête son règlement intérieur	★★★☆ +@	✎
IT	il Consiglio di amministrazione adotta il proprio regolamento interno	★★★☆ +@	✎

DROIT [CdT]			Entrée entière
EN	an Opposition Division shall consist of three members	★★★☆ +@	✎
DE	eine Widerspruchsabteilung setzt sich aus drei Mitgliedern zusammen	★★★☆ +@	✎
ES	una División de Oposición constará de tres miembros	★★★☆ +@	✎
FR	une division d'opposition se compose de trois membres	★★★☆ +@	✎
IT	una divisione di opposizione è formata da tre membri	★★★☆ +@	✎

DROIT [CdT]			Entrée entière
EN	the Secretariat for the Administrative Board shall be provided by the Office	★★★☆ +@	✎
DE	die Sekretariatsgeschäfte des Verwaltungsrates werden vom Amt wahrgenommen	★★★☆ +@	✎
ES	la Oficina se encargará de la secretaría del Consejo de Administración	★★★☆ +@	✎
FR	le secrétariat du Conseil d'administration est assuré par l'Office	★★★☆ +@	✎
IT	l'Ufficio svolge la funzione di segretariato del Consiglio di amministrazione	★★★☆ +@	✎

DROIT [CdT]			Entrée entière
EN	the Administrative Board shall take its decisions by a simple majority	★★★☆ +@	✎
DE	der Verwaltungsrat faßt seine Beschlüsse mit der einfachen Mehrheit	★★★☆ +@	✎
ES	el Consejo de Administración tomará sus acuerdos por mayoría simple	★★★☆ +@	✎
FR	le Conseil d'administration prend ses décisions à la majorité simple	★★★☆ +@	✎
IT	il Consiglio di amministrazione prende le sue decisioni a maggioranza semplice	★★★☆ +@	✎

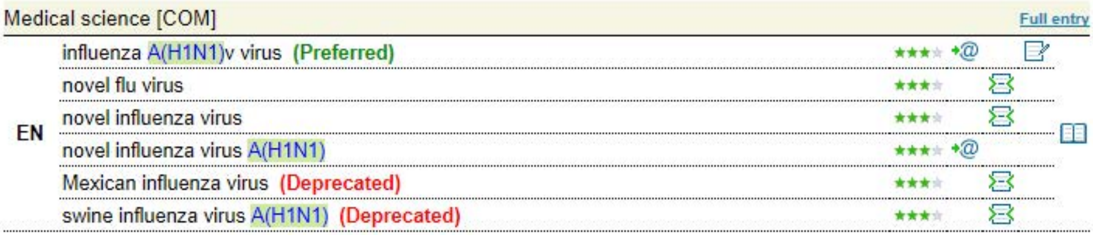
Figure 3: semi-fixed phrases including shall in IATE.

The inclusion of "formulae" and "phrases" in the term base is also a sign that terminologists are more and more concerned with phraseological units and various types of formulaic expressions. The need to standardize language (especially in legal texts, but also in technical fields such as aviation and aeronautics) encourages people to make use of ready-made phraseological patterns which sometimes go way beyond the traditional notion of terms viewed as a noun phrase. In this context, the new European financial supervision authorities which organize stress tests at the European level in coordination with the European Central Bank regularly issue Guidelines which must be translated in all the official EU languages. The translators working in the financial field to provide the various linguistic versions of these Guidelines have received strict recommendations concerning the translation of modal auxiliaries like "shall" and "should" (e.g. should needs to be translated by devoir in French in the specific context of Guidelines for financial supervision). These questions are debated by the terminologists of the various language teams of the European Union and IATE now includes a fair amount of "fixed phrases" which show how these modal auxiliaries should be translated (e.g. "shall" in English most frequently translates as a present tense in French legal texts, when in normal, non-legal language, the default is usually a future tense in French). Figure 3 above illustrates some of these "semi-fixed phrases" included in IATE (see "shall be provided by the Office" à "est assuré par l'Office"). It is clear that there has been an evolution over the last few years with respect to the traditional distinction between lexical items included in dictionaries and terms included in terminology databases and what used to be a clear-cut distinction now increasing appears as a cline between lexicography and terminology.

4.3 Metadata for User Preferences

Very much like a traditional dictionary which makes use of usage notes and a variety of labels aimed at capturing levels of formality (formal, informal, slang, taboo...), a terminological database such as IATE makes extensive use of metalinguistic labels which guide translators in their daily work. Official terminology in any field may indeed change rapidly and terms which are commonly used today may become deprecated tomorrow. It is therefore important to capture the preferences expressed by the main "consumers" of a translation (official organizations and administrations, public bodies and authorities, scientific communities, etc). Such preferences may result from simple stylistic or even sometimes arbitrary preferences and conventions. They may also reflect historical evolutions and a need to avoid geo-political problems. IATE will therefore make use of metalinguistic labels such as "preferred", "obsolete" or "deprecated" to provide information about changes in the pragmatic use that is made of these terms.

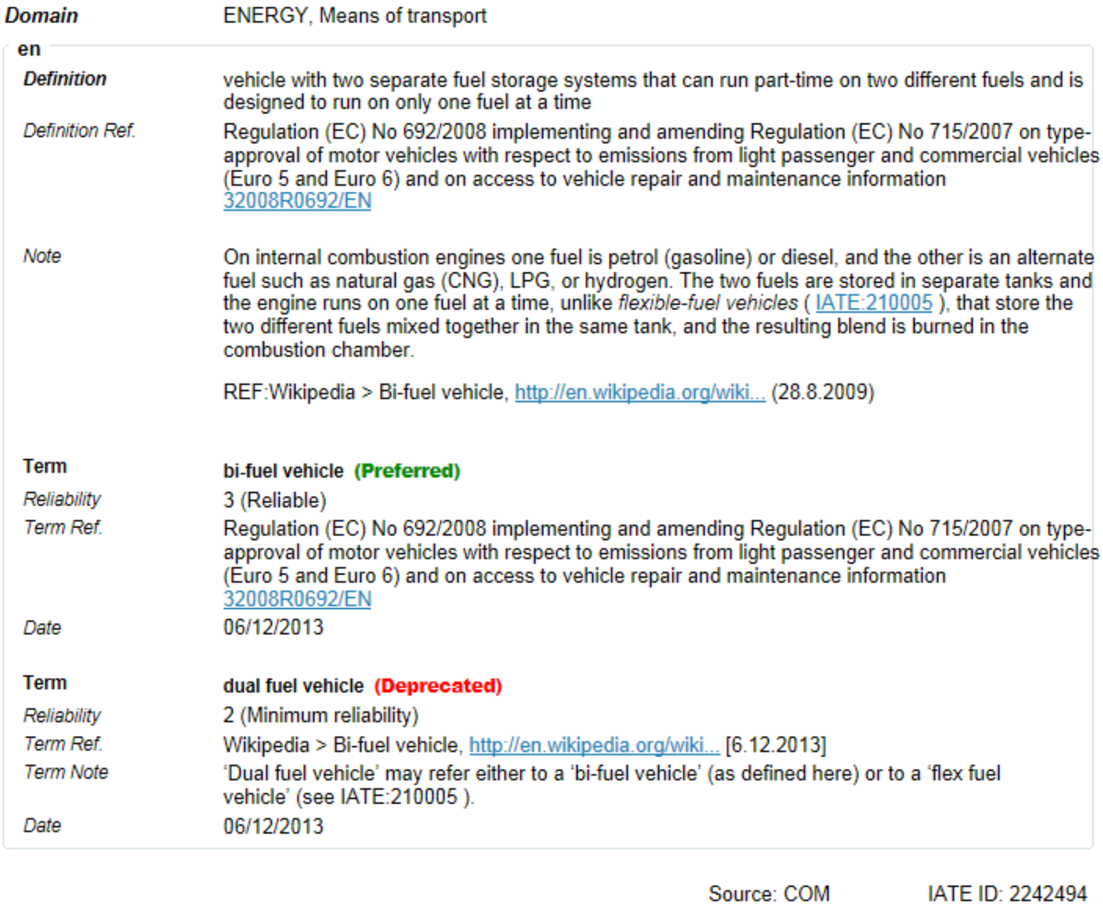
The concept for the disease known as A(H1N1) is a case in point. IATE indicates that several "synonymous" terms can be used to refer to that disease, which was first observed in Mexico in 2009, including the term Mexican flu as well as swine flu. The term Mexican flu was used at the very beginning of what later became a worldwide epidemic, but international organisations such as the European Centre for Disease Prevention and Control (ECDC), one of the Translation Centre's clients, expressed a strong preference in favour of the term A(H1N1)v, rather than Mexican flu or swine flu. Such preferences are captured through the use of labels like Preferred (appearing in green on the IATE web site) or Deprecated (in red on the IATE web site), as illustrated in Figure 4.



Medical science [COM]			Full entry
EN	influenza A(H1N1)v virus (Preferred)	★★★☆	+@
	novel flu virus	★★★☆	
	novel influenza virus	★★★☆	
	novel influenza virus A(H1N1)	★★★☆	+@
	Mexican influenza virus (Deprecated)	★★★☆	
	swine influenza virus A(H1N1) (Deprecated)	★★★☆	

Figure 4: Preferred terms vs Deprecated terms.

The usage note included at the conceptual level of the terminological record reads as follows:
ECDC prefers to use the term influenza A(H1N1)v (where v indicates variant), which has been chosen by WHO's Global Influenza Surveillance Network and helps distinguish the virus from seasonal influenza A(H1N1) viruses and A(H1N1) swine influenza viruses. A name for the disease caused by the virus has yet to be determined by WHO but the term 'swine flu' is inaccurate for what is now a human influenza. REF: European Centre for Infectious Disease Control ECDC Interim Risk Assessment > Health Topics > Documents > Human cases of influenza A(H1N1)v, http://www.ecdc.europa.eu/en... (27.8.2009)
Metalinguistic labels such as Deprecated/Obsolete or Preferred are also complemented by the systematic use of reliability codes, which are captured in a different field in the database. Less reliable terms may indeed be included in the database in order to offer as much information as possible to the translators, but a low reliability code provides an indication that the term is based on information coming from less trusted sources, as is the case for the distinction between the deprecated term *dual fuel vehicle* (reliability =2) and the preferred term *bi-fuel vehicle* (reliability =3), as in Figure 5 below.



Domain	ENERGY, Means of transport
en	
Definition	vehicle with two separate fuel storage systems that can run part-time on two different fuels and is designed to run on only one fuel at a time
Definition Ref.	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information 32008R0692/EN
Note	On internal combustion engines one fuel is petrol (gasoline) or diesel, and the other is an alternate fuel such as natural gas (CNG), LPG, or hydrogen. The two fuels are stored in separate tanks and the engine runs on one fuel at a time, unlike *flexible-fuel vehicles* (IATE:210005), that store the two different fuels mixed together in the same tank, and the resulting blend is burned in the combustion chamber. REF:Wikipedia > Bi-fuel vehicle, http://en.wikipedia.org/wiki... (28.8.2009)
Term	**bi-fuel vehicle (Preferred)**
Reliability	3 (Reliable)
Term Ref.	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information 32008R0692/EN
Date	06/12/2013
Term	**dual fuel vehicle (Deprecated)**
Reliability	2 (Minimum reliability)
Term Ref.	Wikipedia > Bi-fuel vehicle, http://en.wikipedia.org/wiki... [6.12.2013]
Term Note	'Dual fuel vehicle' may refer either to a 'bi-fuel vehicle' (as defined here) or to a 'flex fuel vehicle' (see IATE:210005).
Date	06/12/2013

Source: COM IATE ID: 2242494

Figure 5: Preference labels and reliability codes.

A label such as Obsolete may be used to mark terms which are no longer in official use. Official denominations for countries or cities may indeed change, as was the case for Bombay, which was changed to Mumbai. Depending upon the context, translators may need to preserve the former name (in historical documents, for instance), which is why the two terms need to coexist in the database, with labels distinguishing them.

5 ECHA-term

IATE is a "generalist" database covering many subject fields. It is one of the key resources used by translators, who contribute to its enrichment alongside the terminologists who manage the database. The Translation Centre for the Bodies of the European Union also created much more specialized databases, such as the ECHA-term database, which was developed for the European Chemicals Agency (ECHA), located in Helsinki, Finland. ECHA was created in 2007 to implement the European Union's chemical legislation, and more particularly the REACH Directive which was adopted to improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemicals industry. REACH also promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals. With the REACH regulation, companies are responsible for providing information on the hazards, the risks and the safe use of the chemical substances they manufacture, import and transport throughout the European Union. The Classification, Labelling and Packaging (CLP) Regulation introduced a globally harmonised system for classifying and labelling chemicals in the EU, thereby ensuring that the hazards presented by these chemicals are clearly communicated to workers and consumers.

The European Chemicals Agency therefore has very important multilingual communication tasks (the Translation Centre produced over 25,000 pages of translations for ECHA in 2013, mainly leaflets, technical guidance, web content, IT manuals, administrative documents and news items translated into all the official languages of the European Union). In 2009, the ECHA-term project was launched with the objective to provide ECHA and its stakeholders with a reliable, coherent, and up-to-date source of terminology to harmonise the use of terminology in the REACH and CLP context, to enhance clear communication and ultimately to reduce costs for the stakeholders. The more general aim was to help industry comply with the legal requirements, to support the national authorities in their work and to improve the quality of the translated material. The Translation Centre collaborated closely with ECHA to create a terminological database, which included the compilation of a multilingual database of over 1200 terms related to REACH, CLP and the biocides regulation, as well as “substances of very high concern” (over 50 terms) in 23 languages. The project also included the development of a platform for the dissemination of the contents.

5.1 Compilation of Contents for ECHA-term

The general process for the creation of the terminological contents of the database can be described as follows:

- Definition of a relevant corpus in the source language (usually English);
- Semi-automatic extraction of concepts and completion with definition, reference, context, note etc. by terminologists;
- Validation of the monolingual glossary by 2 or 3 translators to ensure that the data is relevant (are any key concepts missing)
- Formal revision by English terminologist;
- Validation of the monolingual glossary by ECHA’s experts;
- Multilingual phase: target equivalents and relevant information are completed by the Translation Centre’s terminologists;
- Ideally target language equivalents are validated by experts (ECHA);

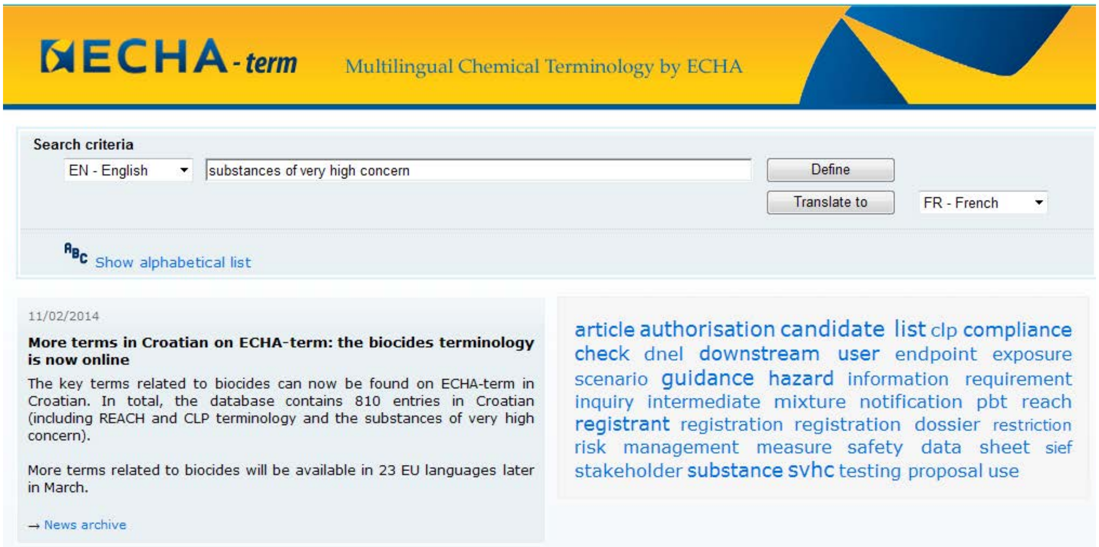
- Import of data in ECHA-term;
- Maintenance of data following user feedback

5.3 Main Features

The database, which can be consulted free of charge at http://echa.cdt.europa.eu, offers a range of search options. Users can search by:

- Terms
- EC numbers
- CAS numbers (unique numeric identifier designating one substance in the Chemical Abstracts Service)
- GHS codes (Globally Harmonised System)
- Hazard
- Precautionary statements

In addition to the search criteria above, an alphabetical list of terms can also be browsed. A word cloud also appears to the left of news items on the home page, displaying the most-frequently searched terms, as illustrated in Figure 6. A lemmatization tool is also included to provide a match with respect to the contents of the database even if the query is inflected (as in a query on the plural form in substances of very high concern).


ECHA-term
Multilingual Chemical Terminology by ECHA
Define
Translate to
Show alphabetical list
article authorisation candidate list clp compliance check dnel downstream user endpoint exposure scenario guidance hazard information requirement inquiry intermediate mixture notification pbt reach registrant registration registration dossier restriction risk management measure safety data sheet sief stakeholder substance svhc testing proposal use
→ News archive

Search criteria
EN - English | substances of very high concern
FR - French

11/02/2014

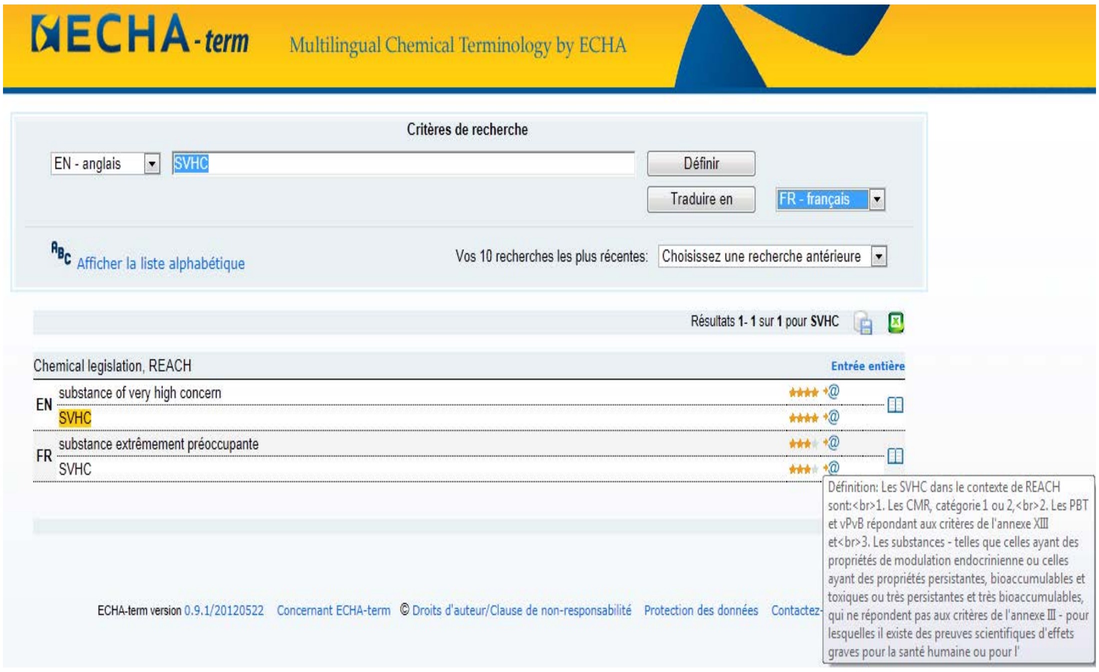
More terms in Croatian on ECHA-term: the biocides terminology is now online

The key terms related to biocides can now be found on ECHA-term in Croatian. In total, the database contains 810 entries in Croatian (including REACH and CLP terminology and the substances of very high concern).

More terms related to biocides will be available in 23 EU languages later in March.

Figure 6: ECHA-term home page.

The user can choose to either define the monolingual term or to translate it into one of the 23 supported languages. The presentation of term entries is very similar to the layout offered by IATE, as one can see in Figure 7 below.


ECHA-term
Multilingual Chemical Terminology by ECHA
Critères de recherche
Définir
Traduire en
Afficher la liste alphabétique
Entrée entière
Concernant ECHA-term
© Droits d'auteur/Clause de non-responsabilité
Protection des données
Contactez-

EN - anglais: SVHC

Traduire en: FR - français

Vos 10 recherches les plus récentes: Choisissez une recherche antérieure

Résultats **1**- **1** sur **1** pour **SVHC**

Chemical legislation, REACH

EN	substance of very high concern	★★★★
	SVHC	★★★★
FR	substance extrêmement préoccupante	★★★☆
	SVHC	★★★☆

Définition: Les SVHC dans le contexte de REACH sont:
1. Les CMR, catégorie 1 ou 2,
2. Les PBT et vPvB répondant aux critères de l'annexe XIII et
3. Les substances - telles que celles ayant des propriétés de modulation endocrinienne ou celles ayant des propriétés persistantes, bioaccumulables et toxiques ou très persistantes et très bioaccumulables, qui ne répondent pas aux critères de l'annexe III - pour lesquelles il existe des preuves scientifiques d'effets graves pour la santé humaine ou pour l'

ECHA-term version 0.9.1/20120522

Figure 7: EN->FR translation of SVHC ("substance of very high concern" - "substance extrêmement préoccupante" in French).

Figure 8 below illustrates the inclusion in ECHA-term of precautionary statements. Such statements are phrases that describe the recommended measures to minimise or prevent adverse effects resulting from exposure to a hazardous substance or mixture due to its use or disposal. It is easy to understand why standardisation is crucial in this context, since the manufacturers and the chemical industry need to make use of "pre-fabricated language" which is in a way similar to the use of controlled language and vocabulary in the aviation and aeronautic domain. The labelling and packaging of dangerous goods need to be clear and unambiguous, which leaves little room for creativity and stylistic variations. It is therefore not surprising to find fairly lengthy phrases and linguistic material in this terminology database which goes beyond the traditional notion of a term typically viewed as a noun phrase. Consider the following examples of precautionary statements whose translations are provided into all the EU official languages:

- Do not spray on an open flame or other ignition source. [see Figure 8]
- Avoid contact during pregnancy/while nursing.
- Keep cool. Protect from sunlight.
- Rinse cautiously with water for several minutes.

- Do not pierce or burn, even after use.
- Keep away from any possible contact with water, because of violent reaction and possible flash fire.

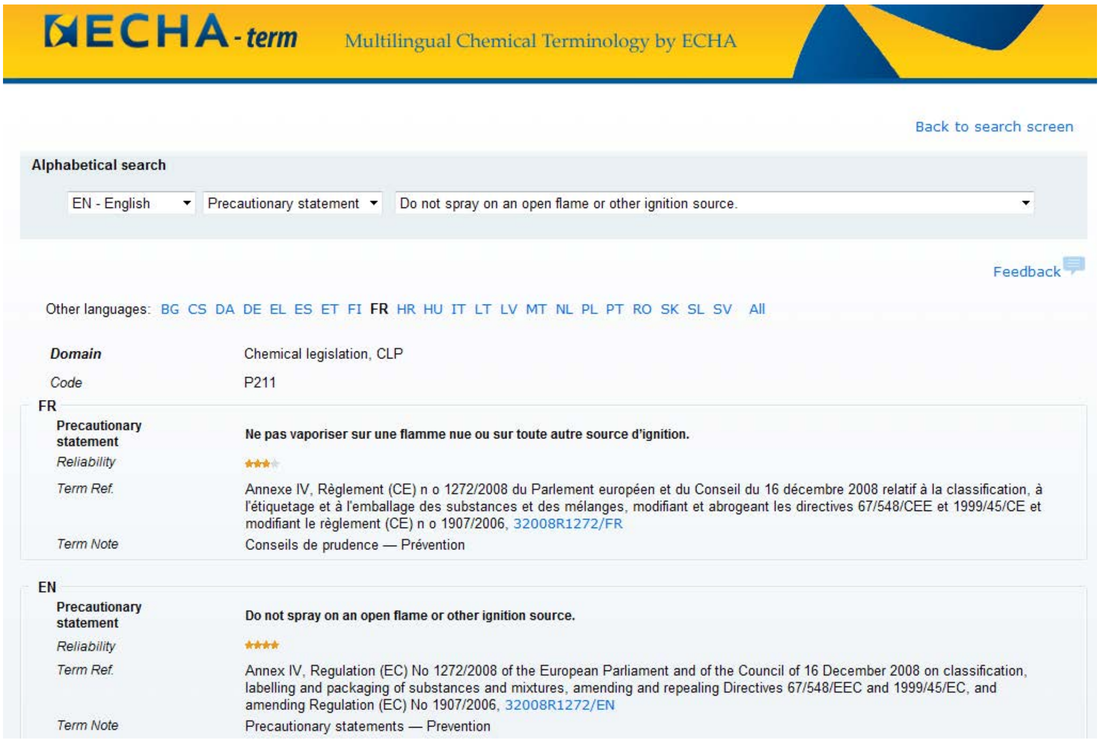
The database also includes hazard statements such as:

- Contains gas under pressure, may explode if heated
- In contact with water releases flammable gases which may ignite spontaneously.
- Harmful to aquatic life with long lasting effects.
- Toxic by eye contact.

The examples above make it abundantly clear that the traditional distinction between lexicography and terminology is more and more blurred. The phraseological patterns displayed in the precautionary statements above sometimes correspond to entire sentences (in the imperative form). In some cases, they correspond to what would be considered a collocation in a traditional dictionary. Keep cool is a case in point: absent any context, the phrase is highly ambiguous ('stay calm and relaxed', 'gardez votre calme' in French, is a possible meaning). In the chemical legislation covered by our specialized database, the advice is not ambiguous and indicates that a product should not be exposed to high temperatures ('tenir au frais' in French). One can imagine the possible disastrous consequences of mistranslations of such labels if the appropriate terminology is not respected.

5.3 Pictograms as Terms

Terms included in terminology databases are most often noun phrases, although, as we have seen above, verbal collocations such as *keep cool,* complex imperative sentences, other types of prefabricated phraseological patterns and even modal auxiliaries may be granted term status in such databases. A more recent trend appears to be the inclusion of items which are traditionally seen as non-translatable material, such as images or diagrams. ECHA-term has innovated in this context, with the inclusion of pictograms used in the chemical CLP legislation. Figure 8 illustrates the information displayed for a pictogram corresponding to corrosion, which refers to a type of physical or health hazard. Other pictograms such as skull and crossbones, exclamation marks, gas cylinders or exploding bombs will allow users to quickly find out the meaning of such graphical representations, a crucial element for the industry, but also for firemen and civilian protection specialists in emergency situations. Such CLP pictograms can be used when creating safety data sheets and training material in national languages in the various Members States of the European Union.



ECHA-term Multilingual Chemical Terminology by ECHA

Back to search screen

Alphabetical search

EN - English ▾ | Precautionary statement ▾ | Do not spray on an open flame or other ignition source. ▾

Feedback

Other languages: BG CS DA DE EL ES ET FI **FR** HR HU IT LT LV MT NL PL PT RO SK SL SV All

Domain	Chemical legislation, CLP
Code	P211

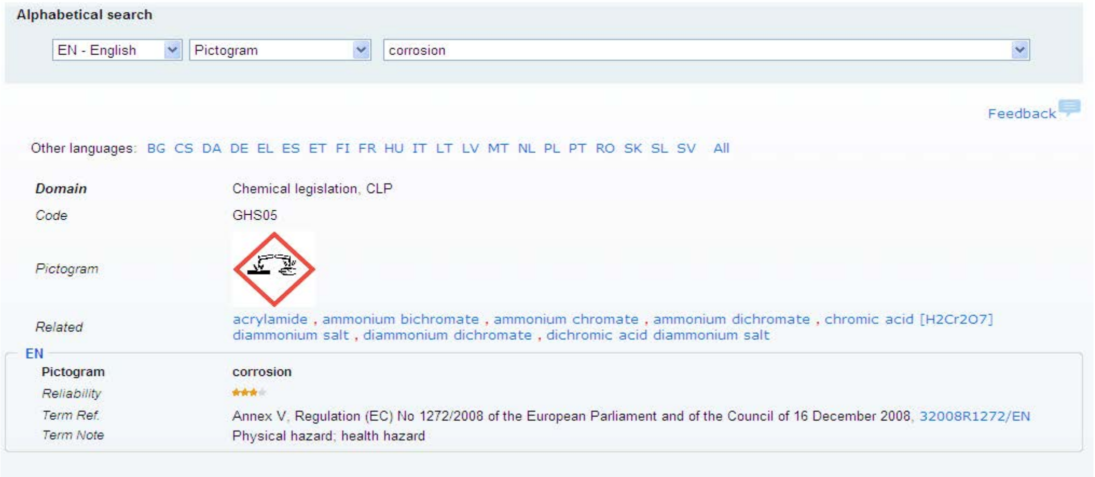
FR

Precautionary statement	**Ne pas vaporiser sur une flamme nue ou sur toute autre source d'ignition.**
Reliability	★★★☆
Term Ref.	Annexe IV, Règlement (CE) n o 1272/2008 du Parlement européen et du Conseil du 16 décembre 2008 relatif à la classification, à l'étiquetage et à l'emballage des substances et des mélanges, modifiant et abrogeant les directives 67/548/CEE et 1999/45/CE et modifiant le règlement (CE) n o 1907/2006, 32008R1272/FR
Term Note	Conseils de prudence — Prévention

EN

Precautionary statement	**Do not spray on an open flame or other ignition source.**
Reliability	★★★★
Term Ref.	Annex IV, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006, 32008R1272/EN
Term Note	Precautionary statements — Prevention

Figure 8: Precautionary statements.



Alphabetical search

EN - English ▾ | Pictogram ▾ | corrosion ▾

Feedback

Other languages: BG CS DA DE EL ES ET FI FR HU IT LT LV MT NL PL PT RO SK SL SV All

Domain	Chemical legislation, CLP
Code	GHS05
Pictogram	
Related	acrylamide , ammonium bichromate , ammonium chromate , ammonium dichromate , chromic acid [H2Cr2O7] diammonium salt , diammonium dichromate , dichromic acid diammonium salt

EN

Pictogram	**corrosion**
Reliability	★★★☆
Term Ref.	Annex V, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008, 32008R1272/EN
Term Note	Physical hazard; health hazard

Figure 9: pictograms in ECHA-term (corrosion).

5.4 A User Perspective

In 2012, the European Chemicals Agency conducted a survey in order to better understand to what extent the ECHA-term users were satisfied with the glossary and who these users were. With around 3200 visitors per month and about 300 search queries per day, ECHA-term is obviously a much more 'confidential' database than IATE. Yet, contrary to original expectations, only 22% of the users who responded to the survey are translators. The majority (56%) of the respondents indicated that they actually work in the chemical industry, the remainder working for EU Member States, for international organisations and for NGOs such as environmental protection agencies. The vast majority of the users indicated that the database helps them understand the REACH and CLP Regulations and stressed that the terms were relevant to them. The multilingual nature of the database is considered a key feature for the users, who mainly visit the ECHA-term web site in order to look up the translation of specialized terminology. A feedback mechanism also allows users to suggest new terms (around 100 terms are added every year) and to provide comments about existing entries. Users can also download the entire database.

Such results explain why this project has been considered a success. They show that a small, but well-maintained glossary can make a difference for the Agency's stakeholders, who use it on a daily basis. It contributes to the use of a unified and consistent terminology in all the translations related to the regulations in the chemical field, a sine qua non in multilingual communication.

6 Conclusion

I started this paper with a reference to ten Hacken's remark that "dictionaries are not descriptions of a language, but tools with which users of the dictionary solve problems of a particular type". The nature of the dictionary therefore determines which types of problems it can solve, he argued in his 2012 Euralex paper. The same is true of terminology databases, as we have seen. The use of such databases in today's globalized economy and multilingual world accounts for the nature of the linguistic information included in these electronic resources. Translations, definitions, acronyms, subject field labels, usage notes and examples are similar to what can be found in monolingual or bilingual dictionaries. Some other types of information are used somewhat differently or to a larger extent in terminology databases, however. References and reliability codes are crucial in term bases, although they are virtually absent in traditional monolingual dictionaries, even if historical dictionaries such as the OED do make use of reference information, an essential element when sketching the historical development of a given lexical item. Information about the author of a term entry is also important in term bases, given that terminology is frequently associated with standardized language used by specific communities of users.

Metalinguistic labels, which are not different from prescriptions, as is pointed out by ten Hacken (2012: 843), are found both in traditional dictionaries and in terminology databases. Labels such as *nonstandard, preferred or obsolete* reflect the lexicographer's or the terminologist's attempt to capture a judgment which will be exploited by the end user of the linguistic resource. This label will help the user decide whether it is pragmatically appropriate to use a given form (e.g. a translation), depending upon the context, the nature of the document produced, the client for whom the translation is made, etc.

The distinction between terminological items and lexical items is also more and more blurred. The nature of the linguistic items discussed by terminologists has undoubtedly evolved over the last 10-15 years. The inclusion in specialized electronic glossaries and term bases of items such as modal auxiliaries, complete sentences, collocational or phraseological patterns, images, diagrams and pictograms is driven by the needs of the target users and the requirements of modern multilingual communication. In this respect, the road from lexicography to terminology is more a continuum, a cline, rather than a hard-and-fast dichotomy.

Another issue which will also need to be addressed in the future is the level at which terminology should be managed (see also Fontenelle and Rummel, in press). Should terminology management be centralized or should it rather be done at the local level, down to the level of individual translators in big translation services? How then should the data be made available to its users? Clearly, web technologies have made it possible to disseminate terminological knowledge to millions of users (the publicly available version of IATE received 44 million queries in 2013). However, one of the major stumbling blocks in the dissemination process is that it is still up to the individual translator or user to 'suspect' that a term base such as IATE or ECHA-term is able to provide interesting and useful information about a given term. What is therefore needed is a mechanism which can alert a translator that a word or a sequence of words appearing in the source text she is dealing with corresponds to a term entry in a specialized database for which an equivalent exists in the target language. Such tools exist at the local level, but will need to be linked to huge databases like IATE, without forcing the translator to host a local copy of a 9-million-term database, which is not recommended for obvious performance reasons. A number of initiatives are currently under way to tackle this crucial issue. Another burning issue is also related to the use of term checkers which ensure that only recommended (read 'validated') terminology is used and that 'dispreferred', obsolete or deprecated terms are not used by the translator. Once again, such obstacles require some level of linguistic processing to match inflected forms in a text and the canonical forms recorded in the quality assurance mechanisms. Organization challenges are also at stake here, since it is crucial to determine who is doing what. Such challenges are different from the question revolving around the distinction between terminological and lexical items, but they are equally important from a user's perspective. Should translators themselves take care of the terminological work? Where should they capture the preferences expressed by the "clients"? Should it be done centrally or locally? How can we make sure these preferences are not one-off information, but can be recycled in future translations to avoid repeating the same mistakes?

These questions do not seem to have clear-cut answers: what is clear, however, is that the solutions can only be effective if they combine technological innovation, using the appropriate amount of linguistic processing, together with organizational changes to make the best use of what modern technology can offer to language workers.

7 References

Fontenelle, Th. and Mergen, C. (1998): « Les interfaces terminologiques au Service de Traduction de la Commission européenne », Terminologie et Traduction, 1.1998, Commission des Communautés Européennes, Luxembourg, 210-221.

Fontenelle, T. (in press) 'Bilingual Dictionaries'. In Durkin, P. (ed.) The Oxford Handbook of Lexicography. Oxford: Oxford University Press.

Fontenelle T., Rummel, D. (in press) 'Term banks'. In Hanks, P., De Schrijver, G.-M. (eds) International Handbook of Modern Lexis and Lexicography. Springer.

Jacquemin, C. and Bourrigault, D. (2003) 'Term extraction and automatic indexing', in Mitkov, R. (ed.) The Oxford Handbook of Computational Linguistics. Oxford: Oxford University Press. 599-615.

Pearson, J. (1998) Terms in Context. Studies in Corpus Linguistics. John Benjamins Publishing Co. Amsterdam.

Reichling, A. (1998) 'Gestion centrale de la terminologie, EURODICAUTOM et ses outils satellites'. Terminologie et Traduction, 1.1998, Commission des Communautés Européennes, Luxembourg, 172-201.

Rey, A. (1992) La terminologie – Noms et notions. 2ème édition. Collection « Que sais-je ? ». Presses Universitaires de France, Paris.

Sager, J. (1990) A Practical Course in Terminology Processing. John Benjamins Publishing Company. Amsterdam.

Simpson, J. (2000) 'Preface to the third edition of the OED'. Accessed at: http://www.oed.com/public/oed3preface/preface-to-the-third-edition-of-the-oed. [05/04/2014]

Ten Hacken, P. (2012) 'In what sense is the OED the definitive record of the English language?', Fjeld, R., Torjusen, J. M. (eds) Proceedings of the 15th EURALEX International Congress, University of Oslo, 834-845.

Ten Hacken, P. (2009) 'What is a Dictionary? A View from Chomskyan Linguistics'. International Journal of Lexicography. 22.4: 399-421.

Wright, S. E. and Budin, G. (2001) Handbook of terminology management (Volume 2): Application-oriented terminology management. John Benjamins Publishing Company.

Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries

Ulrich Heid
Universität Hildesheim, Germany
heid@uni-hildesheim.de

Abstract

We discuss Natural Language Processing (NLP) tools and techniques which may be used to enhance the user friendliness of electronic dictionaries. Intended properties of electronic dictionaries on which we focus are improved guidance for text production, as well as easy and efficient access to lexical data in text reception dictionaries. In this talk, we focus on those NLP techniques which are mostly available for the major European languages: morphological analysis of inflection and word formation, as well as syntactic analysis. We also address the relevance of a detailed classification and representation of lexical data categories within the dictionary: this is a central prerequisite for any integration of dictionaries and NLP tools. Our discussion is embedded in an interpretation of the claims of the lexicographic Function Theory with respect to user orientation in dictionaries.

Keywords: electronic dictionaries; NLP tools and techniques; user orientation

1 Introduction

In this article, we give a short overview of existing and possible applications of Natural Language Processing (NLP) that could be used to enhance the user friendliness and usability of electronic dictionaries. Enhancing user friendliness in this context means providing better guidance in text production dictionaries, as well as improving access to the data provided by reception dictionaries. In the long term, we envisage integrated lexical information systems that combine a dictionary with a number of Natural Language Processing components.

We first situate our discussion in the framework of our view on the lexicographic Function Theory (cf. e.g. Tarp 2008), and we summarize those aspects of the Function Theory that are directly relevant for the integration of language processing and dictionaries (section 2). As the internal representation of lexical and lexicographical data is a key element in the interaction between the lexicographic and the language processing components, we devote a short section to the issue of data categories and markup (section 3). In section 4, we then discuss existing and likely upcoming language processing devices for text production dictionaries (section 4.1) and for text reception dictionaries (section 4.2). Before we conclude, we address a few general design issues and questions related with the presentation of language processing results to the user (section 5).

2 User Orientation in Dictionaries

Requesting user friendliness of printed and in particular of electronic dictionaries has almost become a common place of lexicographic theory, but also of the advertisements for dictionaries. An example from metalexicography is the lexicographic Function Theory (cf. e.g. Tarp 2008) which places the user and his1 needs in the centre of its reasoning about dictionary design.

2.1 Metalexicographic Viewpoint

As stated by Tarp (2008), dictionaries are to be seen as utility tools. The dictionary is a (possibly network-like) set of structured texts from which a user may be able to extract textual data that allow him, by means of an interactive interpretation, to derive information. The user will go through this interpretation process in response to a need that arises from a non-lexicographic situation. Tarp classifies these needs into several types; the needs immediately relevant to the discussion in this paper are either cognitive or communicative in nature. Cognitive needs arise in situations where the user wants to know about or to learn certain facts, be they about things, concepts or words. Communicative needs arise in (the preparation for) communication activities, i.e. text reception (reading or hearing – and understanding) or text production (writing or speaking – and lexical or grammatical choice). Translation, the revision of texts in a foreign language etc. are also communicative situations, and thus translation towards the mother tongue is a receptive activity, and translation to a foreign language is a production-oriented one.

Dictionaries are supposed to provide appropriate data for users to satisfy needs of the above types. An ideal dictionary, according to the lexicographic Function Theory (henceforth: FT), satisfies exactly one type of need. In terms of FT, the "dictionary function" is to satisfy such a need, and the optimal dictionary is monofunctional. While this ideal is hardly ever commercially viable in printed dictionaries, it can be approximated in electronic dictionaries (cf. e.g. Bergenholtz/Bergenholtz (2011), for an exemplification).

The development of dictionaries, be they printed or electronic, is typically governed by lexicographic processes. These have in the past been exclusively geared towards the production of paper dictionaries, but since the advent of electronic dictionaries (or electronic versions of print dictionaries) they may also be more general (cf. Gouws, to appear), aimed at setting up a repository of lexicographical data that can be used in both an electronic and a print dictionary. We situate the discussion of the use of Natural Language Processing (NLP) tools and techniques in a scenario which is aimed at such a possible double or even multiple use of lexicographic data.

[1] For reasons of practicality, we use the masculine form throughout this paper, meant to denote both genders.

2.2 User Orientation in Electronic Dictionaries

In the above sense, we assume that lexicographers collect data to feed more than one dictionary; or, conversely, that not all collected data will show up in one given dictionary. Rather, most dictionary publishers create a broad repository of lexicographic data from which they will select appropriate items for individual dictionaries. This notion is close to the idea of a "mother dictionary" that feeds into several specific dictionaries, a concept introduced into the discussion by R. H. Gouws. Implicitly, the same idea is present in work by the proponents of the Function Theory: to avoid overloading the user, at query time, with (unnecessary) data, they require lexicographers to carefully select the data categories they want to present to the user for a given dictionary function.

While this requirement is very clear at an abstract, metalexicographic level, the way in which it can be satisfied in practice is described in the Function Theory in much less detail (cf., however, the tables given by Tarp 2008: 75-77).

In a scenario where electronic dictionaries are to be produced from an electronic data collection, providing users with data appropriate for a given need involves filtering the contents of the data collection. In the spirit the well-known distinction between lexicographic data description and lexicographic data presentation, filtering has the following aspects:

(1) selection of data categories relevant for a given dictionary function;

(2) selection of presentational properties appropriate for the targeted user public, in terms of the ordering of microstructural items, of their layout, metalanguage, presentational devices, and of the provision of appropriate access routes to the data.

Both of the above selections depend crucially on the following factors:

(1) the dictionary function;

(2) the pre-existing knowledge of users, in terms of the language (or languages) dealt with in the dictionary, as well as of general aspects of dictionary use (Tarp 2008) or of the use of online information tools;

(3) possibly the complexity of the language phenomena described in the targeted dictionary article(s).

The illustration in figure 1 schematically summarizes the relationship between data repository, filtering and user-oriented dictionary versions.

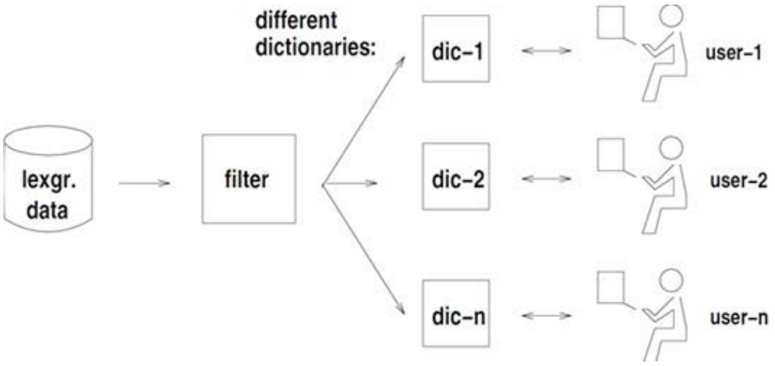


Figure 1: Scenario of the production of user oriented dictionaries: data repository – filtering – monofunctional dictionaries.

The definition of the above mentioned filtering criteria, as well as of the selected presentational devices is in principle part of the dictionary concept ("Wörterbuchplan", in Wiegand's and Gouws's terms); it is similar in nature to specifications of a piece of software; the selection of data categories is a contents-related specification, while the definition of presentational devices and of access routes is mainly a matter of the rendering of individual data categories for printed or on-screen presentation. The specification of access to the data is mainly conditioned by the lexicographer's assumptions about the user's pre-existing knowledge of the domain or language treated in the dictionary. For example, it is plausible to assume that users in need of collocational data, for text production, will know the base of the collocation (in the sense of Hausmann 2004), and will search for an appropriate collocate to express a given idea. Thus a sort of onomasiological access would be preferred: one that allows the user to search for expressions of ideas around a base concept, irrespective of whether these ideas are expressed by collocations, compounds or single words (cf. Giacomini 2012)

This difference in access is illustrated, in figure 2, below, for a printed dictionary (or a print-like electronic one) with sample data from the OCDSE (Oxford Collocations Dictionary for Students of English) for the lemma advance: for text production, the data are sorted according to OCDSE's principles (right side of fig. 2), i.e. per reading of the base, with subdivisions per syntactic model and semantic groups (cf. also Heid/Zimmermann 2012); for text reception (left side), a semi-integrated microstructure is suggested, with a section on the readings of the base, followed by an alphabetical listing of collocational adjectives (and, later in the article, but not shown in figure 2, of collocational verbs and nouns); in an electronic dictionary, reception-oriented access would be more flexible, i.e. from both elements of the collocation, as well as from the collocation as a whole.

Reception

- Readings
  (1) [military] forward movement
  (2) development
  (3) amount of money
- Typical adjectives
  - Allied etc. (cf. German etc) (1)
  - big (= considerable) (2)
  - cash (3)
  - considerable (= big) (2)
  - dramatic (2)
  - German (cf. Allied, etc) (1)
  - great (2)
  - important (1)
  - large (3)
  - notable (2)

Production

Reading 1: forward movement [military]

- ADJ + advance
  - [speed] rapid ~
  - [agent] German ~, Allied ~, etc.
- V + advance
  - [make] make an ~ on X
  The regiment made an advance on the enemy lines

Reading 2: development (often in the plural)

- ADJ + advance
  - [amount] considerable ~, big ~, substantial ~, dramatic ~, enormous ~, great ~, spectacular ~, tremendous
- V + Advance
  - [make] make ~ es (in/on) [plural]

Reading 3: amount of money

- ADJ + advance
  - [quantity] small ~, large ~ - [type] cash ~
- V + Advance
  - [provide] give so, an ~, pay so, an ~
  The university pays me an advance for thir business trip.

Figure 2: Microstructures for easy access to collocations: for text reception (left) vs. text production (right).

3 Data Representation and User Orientation

For an electronic dictionary scenario which involves a central data collection, monofunctional (or at least function-related) dictionaries and the appropriate filtering techniques, a cornerstone of successful implementation is a detailed classification of the available lexicographic data. If, for example, no distinction is made between collocations and idiomatic expressions, and if bases and collocates of collocations (in the above mentioned sense of Hausmann 2004) are not distinguished and marked up, it will be hard if not impossible to provide appropriately differentiated access to collocations vs. idioms. If, for example, both are classified as "multiword expressions", it will be hard to decide which items will go into a text production dictionary and which ones into a text reception dictionary. Most lexicographers would however agree that collocations are relevant for text production (but not needed – with the exception perhaps of what Grossmann/Tutin 2003 would call "opaque collocations", such as FR peur bleue – for text reception), while idiomatic expressions would need to be semantically explained in a text reception dictionary, but would rather not be described, for instance, in a learner-oriented text production dictionary.

If the classification of data categories is central, so is the functional markup of different data categories in the central repository used by a publisher. Some authors call this repository a "database" (e.g. Bergenholtz 2011). This may be technically adequate in the implementation described by Bergenholtz

(2011), but in the general case, this repository need not be a database in the technical sense; publishing houses may also use XML-based data models, or a representation within a content management system, or any other implementation: what counts is that different data categories are distinguished and identifiable. There are examples of publishing houses which use the fine-grained data classification of their data repository to "extract" dictionaries for certain functions and user-groups, without much need for adding new data. Such detailed data categorization and "markup" is also necessary if certain subsets of the lexicographical data available to a publisher are to be provided for the purposes of Natural Language Processing (NLP).

A note of caution may be in order here, with respect to the abovementioned example of collocations and idioms. Some lexicographers might rightly assume that users will not be able to distinguish between the two types of multiword expressions, and claim that they don't need to (cf. e.g. Tarp 2008). This is certainly an appropriate viewpoint, but it does not distinguish the internal representation if lexic(ographic)al data from the presentation of such data to the user. If the abovementioned assumptions about the differences between idioms and collocations, in terms of data selection and access, are correct, an optimal presentation of such data to the user will clearly have to rely on the distinction between the two types of multiword expressions, and on an appropriate markup of each multiword item contained in the data repository. In other words: an optimal presentation of dictionary contents to the user is (trivially) dependent on an adequate internal classification and representation if this contents.

3 Natural Language Processing Tools in Support of User Orientation

In our discussion, so far, Natural Language Processing (= NLP) tools have not been mentioned; a sensible level of user-friendliness can, as has been shown above, be reached without NLP technology, by adhering to good practice in data category classification and markup. A legitimate question is thus what the added value of computational linguistic technology is, in terms of a surplus of user-friendliness of the dictionary. Here, we understand user orientation in a wide sense; it is meant, here, to include aspects of enhanced usability of electronic dictionaries, such as improved access to lexicographic data, individualized support according to the user's pre-existing knowledge, or the availability of information (on demand) which goes beyond the amount of material encoded in the data repository underlying the dictionary, e.g. by means of the presentation of corpus data.

We will address this question in the following by first analyzing monolingual dictionaries for text production, then dictionaries for text reception. We will not discuss the use of NLP techniques for the provision of corpus data to the lexicographer, i.e. corpus analysis and query tools, such as, for instance, those embodied in the "Sketch Engine" (Kilgarriff et al. 2004). Such tools are essentially aimed at the lexicographer, and we will show in which way the end user may profit from other kinds of corpus analysis tools.

4.1 NLP Tools for Text Production Dictionaries

For text production, especially in a foreign language, a rich microstructure is necessary, for example one which explains to the user for each treatment unit which morphological forms it has, which syntactic patterns it follows, or which collocations it enters into. All these properties involve lists of options from which a user may want to select in a text production situation; while syntax and collocations are idiosyncratic (Hausmann 2004 talks of "coded combinatorics") and need thus be listed individually, morphological forms are often more regular and may be provided to the user by means of a morphological generator or of a list of inflection forms. The latter is limited and may require regular updates by the lexicographer, whereas a morphological generator may provide the advantage of comparatively easy ways of extending its coverage. In any case, offering users of a production dictionary on-demand access to inflection forms is certainly very useful.

4.1.1 Access to Corpus Data

Another possible use of NLP techniques in text production dictionaries has to do with the much discussed facilities that give the user access to corpus data, from a dictionary entry, (cf. e.g. Asmussen 2013; Heid et al. 2012; Tarp 2012). It has been claimed that links from the dictionary to corpus data or to the internet provide users with data about language in use which can serve as a model for the users' own text production. Such data thus serves the need for checking one's own formulation hypotheses against putatively "standard" usage.

Asmussen (2013) discusses the issues related with the realization of such links; using the German DWDS dictionary as an example, he shows that the mere coexistence of dictionary and corpus data in one portal is not sufficient to provide adequate service to users. Asmussen has examples of lemmas present only in one of the two sources of information, and he discusses the impossibility of linking corpus data, given today's technology, to readings of a dictionary entry, at least in a large-scale high-quality way. Asmussen's best example of the linking of dictionary and corpus data is from the domain of collocations. In fact, the portal of ordnet.dk provides direct access to usage examples for collocations, with example sentences retrieved from Korpus 2000, the current Danish corpus underlying the DDO dictionary (Den Danske Ordbog, cf. ordnet.dk). To activate the link (in the sense of an information-on-demand offer), the user has to press a button next to the collocation item in the dictionary. Technically, this activates a query in the corpus which is created from the (text form of) the collocation item (cf. Heid et al. 2012).

This facility could be further enhanced if the corpus were preprocessed at a more advanced level of linguistic analysis (e.g. by means of (flat) syntactic analysis), and if the query were made more sensitive to potentially ambiguous corpus sentences. For example a search that starts from the collocation give + resultater (EN "produce/lead to results") provides many relevant examples, e.g. gav betydelige resultater ("gave important results"), gav de ønskede resultater ("gave the intended results"); but it also provides gav 10.3 km/l til resultat ("gave 10.3 km/l as a result"), which is not an example of the searched collocation.

Other, similar facilities involve the retrieval of contexts for specialized terms in specialized texts, e.g. in the intranet of a company, or lists of cooccurrents of items (for the user to choose from) sorted by association strength, as they are provided within Verlinde's ILT tool (URL: https://idp.kuleuven.be/idp/view/login.htm).

On the basis of syntactically analyzed and annotated texts, the same device could also be offered for syntactic subcategorization. Parsed corpus data tend to identify subjects, objects, prepositional complements, verb-dependent clauses or infinitivals in each analyzed sentence; this is true for dependency parsing, which has reached, at least for several European languages, a degree of maturity makes its use in the intended context possible (cf. e.g. Bohnet 2010). Syntactic valency is, as mentioned above, an important property of lexical predicates (verbs, adjectives and nationalizations) that must be learned by foreign language learners. Illustrating valency in full sentences has the advantage of showing the user not only an abstract indication, but also concrete instanciations of it. This principle has been followed, very successfully, in the ELDIT dictionary (Elektronisches Lernerwörterbuch Deutsch-Italienisch, http://eldit.eurac.edu/), where the authors have provided the user with four complementary types of indications for the syntactic construction of verbs (cf. Abel 2002):

(I) a formula of the type "someone suggests something to someone";

(II) example sentences for each pattern;

(III) on-demand indications of the involved grammatical functions (e.g. "object" for "something" in the above example);

(IV) on-demand highlighting of the respective phrases in the example sentences, when the user points the mouse to an element of the formula (i): if, for example, the user points the mouse to "to someone" in the above example, not only the grammatical function (indirect object) is displayed, but also the respective stretch of the example sentence is highlighted.

In ELDIT, these devices are applied to prefabricated examples, i.e. a closed list of verbs and example sentences for these. By use of NLP tools, such a device could be made dynamic, i.e. provided on demand by the user, on the basis of a pre-analyzed (dependency-parsed) corpus and extraction tools for syntactic patterns. If the corpus is big enough and adequately annotated, also frequency data for individual valency constructions and lists of the most prominent fillers of valency relations could be provided. Finally, as Engelberg et al. (2012) have shown, the different possible valency constructions of verbs are used differently in different genres or text types: not all possible patterns are equally frequent in all kinds of texts; if notions like "genre" or "text type" are applied to the annotation of a corpus which is exploited to offer examples for valency patterns explained in a dictionary, such valency preferences by genre or text type can be made visible.

All this may appear futuristic to some readers; it is, however, only dependent on two conditions: an adequately detailed inventory of valency patterns in the dictionary, and good quality corpus parsing. The device would allow users to get real-text models of syntactic constructions, from which they could take inspiration for their own text production.

Such devices are in principle thinkable for all those linguistic properties of lexical items that can unambiguously be identified in a (syntactically parsed) corpus. With adjective+noun-collocations and, to a lesser extent, verb+object- and verb+subject-collocations, this is well possible; the same holds for syntactic valency, for the contextual use of terms from a specialized language; but it does not yet so for lexical semantic properties. Some partial results could be obtained if additional resources are used, e.g. WordNets that would support a search for word combinations and the pertaining sentences according to semantically defined sets of lexical items. Again, what is needed as a prerequisite, are appropriate classifications of the lexicographic data, mappable onto the classifications annotated in the corpus. In all cases, the combination of lexical items and targeted linguistic properties acts as search criteria for corpus data extraction.

Instead of lexical items and their linguistic properties, also pairs of translational equivalents from an electronic bilingual dictionary may be used as search constraints, in this case on parallel corpora; Verlinde's ILT tools provide access to the Europarl corpus (URL: http://www.statmt.org/europarl/), but they use only the source language item as a search criterion, in the hope of providing the user in this way a broad range of equivalence candidates; for very advanced users, and especially for those who are used to work with parallel corpora, this may provide indeed new insight. A more modest, though perhaps more focused (and thus easier to use) version of such search would be one that retrieves example sentence pairs for a given equivalent pair from the dictionary; it may then on demand also provide sentence pairs that do not contain the equivalents mentioned in the dictionary entry, as a complement of information.

4.1.2 Lexico-grammatical Guidance

With the above mentioned provision of example sentences for a given lexical property of an item from the dictionary, one problem mentioned by Asmussen (2013) still remains: for the dictionary user, the relationship between the text he is in the process of producing, and the sentences retrieved from the corpus, may still be rather indirect; the examples may illustrate the behaviour of the searched item, but they will still not necessarily provide an exact solution to the actual text production problem which the user is confronted with, as the other lexical items he intends to use are absent from the retrieved examples.

While the transfer between the corpus examples and the upcoming text of the user may be relatively simple for most European languages, it is much harder in languages with massive rule-based morphosyntactic variation, where lexical choice and grammatical choice interact in more complex ways. Examples of such situations are provided by the South African Sotho and Nguni languages. The morphosyntactic complexity of the noun class system of these languages, of their concordial and pronominal morphemes, as well as of their tense and mood systems interferes with issues of lexical choice that depend on semantic selection criteria. Examples are discussed, among others in this conference, by Bosch/Faaß (2014) and Prinsloo et al. (2014); they analyze possessive constructions in Zulu (type: the medicine of this doctor) and subjects, objects and relative clauses in Northern Sotho (type: the boy who helped the woman), respectively, from the viewpoint of English → Zulu or English → Sotho learner's dictionaries.

Both model the respective phenomena in an NLP tool that implements the morphosyntactic (agreement) rules of the language and interacts with a dictionary whose nominal entries are classified by noun classes and whose verbal entries can be inserted into the grammar of the constructions under analysis. Bosch/Faaß's (2014) system can operate in two modes: one that provides a translation from English to Zulu of the intended possessive construction, and one that in addition explains to the user which construction and agreement rules have been used. A next step could be a system that allows the user to produce his own solution and which then proposes modifications where necessary. Prinsloo's system is not yet implemented but intended to provide similar optional guidance: if, for example, the user plans to construct a Northern Sotho subject-verb-object sentence where the object is not expressed by a noun (phrase), but by a concord, the following situations may occur:

(I) the user may know the appropriate concord: the system will check the appropriateness and confirm it;

(II) the user may know the noun which he wants to express by the concord, but not the concord itself: the system will retrieve the noun class of the item (from its standard dictionary), identify the appropriate concord (from morphosyntactic tables) and propose the appropriate concord, possibly with additional information about the underlying grammatical facts;

(III) the user may only know the English equivalent of the nominal he intends to construct as an object: the system then retrieves the appropriate Sotho noun from a bilingual dictionary and then proceeds as explained under (ii).

In both cases, the objective is text production guidance for learners of the foreign language; the proposed solutions combine some amount of NLP with a well-structured dictionary. Such combined systems may be counted among online language learning systems, or among e-dictionaries. Irrespective of how they are classified, they combine lexicon and (partial) grammar data.

4.2 NLP Tools for Text Reception Dictionaries

While the function of NLP tools in the context of text production goes from a source of inspiration (through the selection of appropriate example sentences from corpora) to guidance in issues of lexico-grammatical choice, NLP tools function mostly as access tools in text reception dictionaries. Text reception starts from an existing text and aims at detecting its meaning and possibly other properties. Access support tools ease the user's retrieval of the right dictionary entry and the right indications within this entry. This kind of support starts with inflectional morphology and may involve word formation, syntax and possibly, at least to some extent, multi-word items and semantics. In all cases, the basic function of the tools is to analyse a word(form), possibly in the context of the sentence the user is reading, and to relate it to an entry or ideally even a reading in the dictionary.

For inflectional morphology, such devices work relatively well and are quite established: if the user enters a word form, the dictionary relates it to the appropriate base form and displays the entry of the pertaining lemma. Such interfaces exist in several online dictionaries, and they may either depend on large lists of inflected forms (related with the appropriate lemmas), or on morphological analysers.

However, similar devices may be used also for word formation: when Bergenholtz/Johnsen (2005) analyzed the log files of their Danish Internet Dictionary (Dansk Netordbog), they discovered that a considerable number of items searched by users, but not found in the dictionary, were word formation products. In Germanic languages, compounding is so productive that a standard monovolume dictionary cannot cover even a small portion of the items found with a non-trivial number of occurrences in a corpus. The same holds, at least to some extent, for derivation products: these also will likely not be covered in full in standard monovolume dictionaries. If the dictionary is meant for text reception, it may even reasonably adopt a policy of focussing on semantically non-transparent compounds and derivations, i.e. on those whose meaning needs to be explained beyond a simple recall of the morphological structure, e.g. because they have idiosyncratic meanings. In such a case, no space may be left for the treatment of transparent compounds.

A morphological word formation analyzer may be useful in such a situation, as it would be able to split a compound that is not part of the nomenclature of the dictionary into its components and guide the user to their entries in the dictionary. For derivation, even generic paraphrases may be given, as is the case in the DériF system for French (URL: http://www.cnrtl.fr/outils/DeriF/). Alternatively, a structural or morpheme decomposition hypothesis may be given, as in the canoo tools (URL: http://www.canoo.net/). In all cases, the user would get partial information in reply to his query, thus at least some guidance towards the analysis of the word formation product. Often, a recall of the morphological structure and links to the dictionary entries of the components may help users understand complex word.

While the above functions are in general seen independent from the context (see below for a discussion), any kind of tool intended for guidance of text readers towards an appropriate entry in the dictionary will inevitably be confronted with ambiguity and context-based disambiguation. This starts with categorial homographs (EN: can: modal verb or noun, cf. Bothma/Prinsloo 2013: 176) or forms which are homographous with items from other word classes (thought: participle or noun, ibid. 174) and goes all the way up to polysemy and reading distinctions. While the latter problems are in the general case still not solvable, categorial homographs of different types can be disambiguated fairly well on the basis of word class tagging and/or syntactic analysis. For the major European languages, tools for these functions are available, also as web services that would allow for an on-the-fly treatment, as suggested by Bothma/Prinsloo (2013: 187 ff.).

If syntactic (dependency) analysis is available, a variant of the search for examples discussed above, in section 3.1, could be applied; if a syntactic analysis of a sentence can be produced, at least the verb and its potential complements can be extracted from the analysis result and matched against the list of valency patterns offered by a dictionary. In many cases, this would reduce the number of readings which the user would need to consult in order to find the meaning of the verb used in the sentence he is reading. Obviously, not always all arguments of a predicate are explicitly mentioned in a sentence, which might increase the number of syntactic readings that have to be looked up by the user in such a situation.

Another type of contextual analysis has been available since the 1990s, already. It deals with (idiomatic) multiword expressions; the objective is to provide the user with the (part of a) dictionary entry that explains the multiword expression, when the user clicks on any of the words that make up the expression. Due to the high level of lexical specificity this device works quite well for idioms (cf. Seretan/Wehli 2013). In combination with a syntactic analysis of the text which the user is working on, this facility could be extended to collocations as well.

5 NLP Tools as a Part of Lexical Information Systems

In the two preceeding sections, we have listed a few simple devices based on NLP technology that could enhance the user friendliness of electronic dictionaries. A number of issues should however be discussed in this context which concern more general aspects of user interaction.

An important first aspect concerns the problem of the quality of the tool output; it can not be guaranteed that the linguistic analysis underlying the integrated tool is correct in all cases. The more NLP components are involved, the more possibilities are there for errors to occur in the processing chain. Thus a setup where, for example, the text being read by the user is syntactically analyzed and then searched, in order to find the appropriate dictionary entry for a given item, may provide fully correct results only in a certain percentage of cases (we would assume at least in three quarters of the cases).

The dictionary developer and the users should be aware of this situation, and such a device should be offered as experimental (cf. Tarp 2012 on this issue in the context of corpus data provision). It would need to be presented to the dictionary user as being fully automatic and not cross-checked by the lexicographer. A warning in the user interface would be appropriate in such a case (e.g. 'N.B.: the following results were automatically produced and have not been checked by lexicographers'). This type of warnings is regularly produced by the canoo morphology system when analyses are displayed which have not been cross-checked by canoo's lexicographers.

Furthermore, the results should be shown in a part of the interface that is clearly recognizable as a part of the dictionary or of the lexical information system. Early realizations, especially of tools that link dictionary and corpus, did not make clear enough to the user that the corpus data shown on screen were meant as an additional service of the dictionary (cf. Bank 2012 on an early version of the Base lexicale du français).

Another, related issue concerns the status of NLP-based services within a dictionary system. In our view, they should always be information-on-demand services: the user should have the possibility to explicitly decide in favour (or against) the use of the NLP-based service. In dictionaries with (personalized) user profiles, this choice may be an element of the user profile.

Finally the design of the overall integrated system should also be governed by the principles of user-oriented dictionary design: for each NLP component to be integrated, the lexicographer should assess which user need (and thus: dictionary function) it satisfies.

6 Conclusion

We have shown a few devices for the enhancement of text production and text reception dictionaries that are based on Natural Language Processing tools. While morphology systems or morphological tables are almost a standard component of current electronic dictionaries, this is much less so with tools for syntactic analysis, and technologies for semantic processing are still in an experimental phase, although some are very promising (cf. e.g. Cook 2014).

We have tried to show that a detailed classification of the data categories contained in an electronic dictionary (or in the data repository underlying it) is a major requirement for any work with NLP tools; this is due to the fact that these data categories (and a common "understanding" about them, between NLP tools and lexical repository) are the interface between the two components.

In our view, syntactic dependency - parsing has reached a degree - of maturity, at least for some European languages, which should allow for its experimental - and perhaps also productive - integration into lexical information systems of the kind discussed here. Prototypes of such systems should be built and tested with users.

If lexicographers join forces with NLP experts, and if they jointly produce integrated systems that present an added value, chances are good that users will accept these systems. Since people are particularly demanding with respect to the quality of language tools (and since dictionaries have a high reputation in this respect), offering integrated services as information-on-demand seems to be an adequate solution.

7 References

Abel, A. (2002). Darstellung der Verbvalenz in einem elektronischen Lernerwörterbuch Deutsch – Italienisch (ELDIT). Neue Medien, neue Ansätze. In: Braasch, A. et al. (eds.): EURALEX-2002 Proceedings, pp. 413-418.

Asmussen, J. (2013). Combined products: Dictionary and corpus. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (eds.) (2013). Dictionaries. An International Encyclopedia of Lexicography. Volume 5.4. pp. 1081-1090.

Bank, C. (2012). Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen. Information - Wissenschaft & Praxis. Volume 63/ 6. pp. 345–360. Accessed at: http://www.degruyter.com/view/j/iwp.2012.63.issue-6/iwp-2012-0069/iwp-2012-0069.xml?format=INT [28/05/2014].

Bergenholtz, H. (2011). Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): e-Lexicography: The Internet, Digital Initiatives and Lexicography, pp. 30-53.

Bergenholtz, H., Bergenholtz, I. (2011). A Dictionary Is a Tools, a Good Dictionary Is a Monofunctional Tool. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): e-Lexicography: The Internet, Digital Initiatives and Lexicography, pp. 187-207.

Bergenholtz, H., Johnson, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. In: Hermes, (34), pp. 117-141.

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.

Bothma, T. J. D., Prinsloo, D. J. (2013). Automated dictionary consultation for text reception: a critical evaluation of lexicographic guidance in Kindle e-dictionaries. Lexicographica: International Annual for Lexicography, Volume 29, pp. 165-198.

Bosch, S., Faaß, G. (2014). Towards an integrated e-dictionary application - the case of an English to Zulu dictionary of possessives. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).

Canoo. Accessed at: http://www.canoo.net/ [27/05/2014].

Cook, P., Rundell, M., Han Lau, J., Baldwin, T. (2014). Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).

DériF. Accessed at: http://www.cnrtl.fr/outils/DeriF/ [27/05/2014].

ELDIT – Elektronisches Wörterbuch Deutsch – Italienisch. Accessed at: http://eldit.eurac.edu/ [27/05/2014].

Engelberg, S., Koplenig, A., Proost, K., Winkler, E. (2012). Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations. Lexicographica: International Annual for Lexicography, Volume 28, pp. 13-48.

Fuertes-Olivera, P. A., Bergenholtz, H., Nielsen, S., Niño Amo, M. (2012). Classification in lexicography. The concept of collocation in the Accounting-Dictionaries. Lexicographica: International Annual for Lexicography, Volume 28, pp. 293-307.

Fuertes-Olivera, P. A., Bergenholtz, H. (eds.) (2011). e-Lexicography: The Internet, Digital Initiatives and Lexicography. London: Bloomsbury.

Giacomini, L. (2012). An onomasiological dictionary of collocations: mediostructural properties and search procedures. Lexicographica: International Annual for Lexicography, Volume 27, pp. 241-267.

Gossmann, F., Tutin, A. (eds.) (2003). Les collocations: analyse et traitement. Amsterdam : De Werelt.

Gouws, R. H. (2013). Aspekte des lexikographischen Prozesses in Print- und Onlinewörterbüchern. To appear. OPAL (IdS Mannheim).

Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In: Steyer, K. (ed.): Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York: De Gruyter, pp. 309-334.

Heid, U., Prinsloo, D. J., Bothma, T. J. D. (2012). Dictionary and corpus data in a common portal: state of the art and requirements for the future. Lexicographica: International Annual for Lexicography, Volume 28, pp. 269-289.

Heid, U., Zimmermann, J. T. (2012). Usability testing as a tool for e-dictionary design: collocations as a case in point. EURALEX-2012 Proceedings, Oslo, Norway, pp. 661-671.

ILT Tools. Accessed at: https://idp.kuleuven.be/idp/view/login.htm [27/05/2014].

Kilgarriff, P. R., Smrz, P., Tugwell, D. (2004). The Sketch Engine EURALEX-2004 Proceedings. Lorient, France, July: pp. 105-116.

Prinsloo, D. J., Bothma, T. J. D., Heid, U. (2014). User support in e-dictionaries for complex grammatical structures in the Bantu languages. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/ Bozen (EURAC).

Seretan, V., Wehrli, E. (eds.) (2013). Context-sensitive look-up in electronic dictionaries. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (2013). Dictionaries. An International Encyclopedia of Lexicography. Volume 5.4. pp. 1046-1053.

Trap-Jensen, L. (2013). Researching lexicographical practice. In: Jackson, H. (ed.): Bloomsbury Companion to Lexicography, pp. 35-47. London: Bloomsbury Academic.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Max Niemeyer.

Tarp, S. (2012). Online dictionaries: today and tomorrow. Lexicographica: International Annual for Lexicography, Volume 28, pp. 253-267.

Using Mobile Bilingual Dictionaries in an EFL Class

Carla Marello
Università di Torino
carla.marello@unito.it

Abstract

Are reference skills of native digital EFL students developed enough to take advantage of having a big bilingual dictionary on their smartphones? To carry out this research a class from an Italian technical high school was observed. Students aged 17 were split into three groups of users and were given different versions of the same bilingual dictionary, the Ragazzini Italian and English dictionary (Zanichelli Bologna 2013 edition). The first group was allowed to use the Android app; the second was given access to the online version on the web portal ubidictionary.zanichelli.it; the third group received paper copies of the dictionary.

Students were asked to answer some questions about their (un)familiarity with Italian monolingual and bilingual dictionaries. Then the three groups carried out the same activities during a two hour-English-lesson.

The case study reports similarities and differences in their performances, showing that linguistic proficiency proved more determinant than access to digital versions. Students were also asked to report on the main difficulties they had to overcome when looking words up in the dictionaries and were invited to suggest possible improvements in the way pieces of information were displayed on the mobile digital version.

Keywords: bilingual dictionary use; mobile dictionary app; advanced CLIL prerequisite

1 Bilingual Dictionaries in EFL Classes

In recent years whether English- L1 bilingual dictionaries should be banned from the EFL classroom has become an issue of debate (cf. Butzkamm 2009). This is the natural consequence of the reassessment of students' use of L1 to check the final results of a reading comprehension dealing with progressively more complex concepts and reasoning. In fact, an increasing number of students with B1 levels of English or higher are being asked to read and understand such texts. It is also the result of a dispute over binding defined as the 'cognitive and affective mental process of linking a meaning to a form' (Terrell 1986: 214) and not to a translation. Studies on language acquisition in learning situations that are at least partially guided have shown that interference from L1 (whether it is the real mother tongue and language used at school) and interference of a L2 studied previously and typologically similar to the foreign language being learned, involves more than the insertion of isolated words in a foreign language text. Above all, research on learning use of verbs in directed motion events in verb-framed languages such as Spanish or Italian by native speakers of satellite-framed languages such as English are revealing that meaning often has a form in L1 that students carry over into words from L2 (See Cadierno 2008). Quadripartition of a linguistic sign as formulated by Hjelmslev (1961) may be (and has been) neglected in studies on language teaching of beginners and intermediate level students, in which the semiotic triangle "meaning-form-reference" is conveniently considered sufficient. Its inadequacy for structures used to make complex references and needed even by beginners, for example the expression of modes, is bypassed by memorization of conversational routines. At advanced levels, however, learners deal with texts in which they are really confronted with the diverse mode with which natural languages form the substance of content, not only for isolated words, but for morphosyntactic structures as well. At that point, teaching English relies upon learner's dictionaries and their increasing effectiveness, resulting from studies based on learner corpora, usage notes and, more recently, on even more refined and frequent-mistake-based collocation dictionaries for learners of English.

Studies on teaching/learning languages other than English have never shown a rigid refusal to use bilingual dictionaries as a reference tool. The explanations can be found in the typological characteristics of the languages[1] and, therefore, in a teaching tradition that focuses (focused?) more on morphology and less on rote learning of word sequences. Studying them at intermediate and advanced levels has, even recently, been largely based on increasingly more effective bilingual lexicographic tools to describe verb patterns and rich in derivatives and phraseology. They are languages that linguists have dealt with extensively in terms of their collocations later because they provide less fixed, more easily split and variable collocations. Linguists truly appreciated the restrictions of the collocations in Romance languages by examining the concordance of the corpora, which were developed after the corpora of English. Italian lexicographers, for instance, recorded collocations first in bilingual and then in monolingual dictionaries, where today they are still found mainly in the examples, and therefore, do not receive the attention they deserve. In addition to the systematic characteristics a natural language has and the traditions tied to its didactic standards, the position a language holds in the global language market needs to be considered. The English only EFL classroom derives mostly from economic and political reasoning (see Cook 2010): a world market stimulates the creation of reference tools which can be sold everywhere. Attention to the user of a specific L1 is often entrusted to local publishing houses which "bilingualize" products by Anglo-American publishers for foreign learners of any mother tongue.

[1] We do not deal with this topic here because it would lead, among other things, to a discussion on macro- and microlexicographic structures in monolingual and bilingual dictionaries of languages morphologically poorer – like English – compared to the macro – and microlexicographic structures of morphologically richer languages, like the Romance and Slavic languages: richer due to their visible differentiation of parts of speech as well as word formation with prefixes and suffixes. See, however, below in § 4.1 the reference to the question in regards to the problems encountered by the students when using the dictionary.

If an English learner's dictionary is "bilingualized", however, the different ways the substance of content is organized is not always evident to the learner: the entry's organization into the various meanings and examples retains the style formulated in English by English speakers (see Marello 1998). A bilingual dictionary, however, set up using good monolingual microstructures is more likely to highlight the differences.

For languages used less internationally, often monolingual dictionaries for foreigners are not produced and bilingual dictionaries are recommendable tools, not only for translating, but also for understanding a text in L2. They help to appreciate the different structures and grasp connotations while giving comparative information that the monolingual dictionary, formulated with the native speaker in mind, deemed unnecessary to the user. In addition, bilingual dictionaries for languages with relatively developed markets and a consolidated lexicographical tradition have reached good levels in the last few decades, thanks to the progress of studies on meta-lexicography, the data available in large corpora, dedicated editing software that makes it possible to check word classes and to verify the good quality of reverse translation routes[2].

It remains to be seen whether the moment has arrived to allow bilingual dictionaries into an EFL class, even though English has very well developed monolingual lexicographical tools.

2 Bilingual Glosses enter EFL Learning through Mobile Devices

In the context of the 'connected' classroom online dictionaries and translation tools are readily available. Even if bilingual paper dictionaries are banned from the classroom, and use of online dictionaries is not practiced, teachers and researchers would have to admit that tablets and smartphones provide access to translation tools everywhere else and young people, in particular, use them frequently. Augustyn (2013) allowed her USA university students engaged in a first-year text-based approach to German to use whatever device they preferred to work with the electronic materials that were made available to them. She noticed that some learners rely entirely on translation because they choose to type every utterance they do not understand in L2, or want to produce in the L2, into a translation tool such as Google Translate (2013: 367-368). Augustyn concludes her paper with the following consideration: "The groundwork for bilingual practice and contrastive analysis for vocabulary acquisition has already been done by lexicographers. Maybe the convergence of a reassessment of translation in the context of SLA theory and foreign language pedagogy, on the one hand, and an i creasing dependence on online learning tools and digital media will introduce the language learner to the bilingual dictionary in digital format?" (Augustyn 2013:381). It is now time, then, for the foreign

[2] In Ragazzini 2013 for Engl. *endeavour* we find It. *sforzo, tentativo,* but in the Italian®English section we do not find *endeavour* as a translation of It. *sforzo* or It. *tentative*. Through this unsuccessful reverse translation the user might conclude that the two Italian words, though precious to roughly understand what *endeavour* means, perhaps should be avoided in a written translation of the passage where the English word appears.

language teacher, and especially for teachers of English which has the most information online, to decide whether the use of L1, at least for comprehension purposes, should continue to be avoided. With the forthcoming introduction of CLIL[3] in Italian secondary schools a decision should be made, because such a form of teaching implies extensive reading and the development of literacy skills in addition to communicative skills. The experiment described in this paper has been carried out to verify if the digital natives are able to use a digital bilingual dictionary covering a vast vocabulary and numerous translations of specialized meanings.

3 The Ideal Class for the Test

A 4th year class was chosen for the experiment from an Italian technical high school, specializing in electronics and made up of 17 year old boys, four of which were not native Italian speakers but had been living in Italy for many years. It was important that learning English had a practical value for the future (professional as well) of the students involved. It was also important that they had no experience consulting reference materials, but were able to judge the pros and cons of having a dictionary online or on their smartphone. The teacher said that she had never used bilingual or monolingual dictionaries for work in the classroom because she did not use translation to teach, but used reading comprehensions and the creation of concept maps in English. Her teaching method focuses on the overall content of the message more than on its form.

About two weeks before the test, the class was given a pre-test to verify the size of vocabulary in reception activities at the B1 level and the ability to produce a derivative with prefixes and suffixes[4].

The instructions for one exercise (see Appendix 1) said "Use the word in the box to form a word that fits in the text" and asked students to use *produce* or *equip* to form the words *product* and *equipment* needed in the context. Particularly noteworthy is the fact that the no one in the class wrote *useless* in the 6*th* blank, whose context was defined over two sentences: "The instruction booklets are always (6). They never help me at all". One of the answers given, the word *unuseful,* is particularly interesting because it is used much more rarely than *useless.*

The other exercise was a cloze test with eight blanks. The students could choose the correct word out of the four proposed. The three distractors given for each gap were semantically similar.

[3] Content Language Integrated Learning: teaching a subject in L2, usually English or French, will be implemented in Italian schools in the last year of secondary schools as of 2015; in high schools specialized in foreign language learning it will begin in the third year. It may be noted, however, that experiments with what is also called *dual learning* are widely implemented, especially in private schools, beginning in primary school.

[4] For the more interesting parts of the experiment for language learning purposes and reference skills see the contributions in progress by Elisa Corino and Elena Martra.

Based on the results of the pre-test, it was decided that on the day of the test rather simple tasks entailing reference skills would be given5. The teacher divided the class into three groups: 5 that had a smartphone which could download the bilingual Ragazzini dictionary application, 5 who would use a paper copy and the remaining 7 who would consult an online version6. Most of the students (12 of 17) stated they had studied English for over 8 years. Three of the 4 foreign mother tongue students said they had been studying English for less than 8 years. They all owned a dictionary of their mother tongue, except the Philippine student who said he owns a bilingual Philippine-Italian dictionary. From the data taken from the research, 15 students (of which 3 out of the 4 not native Italian speakers) admitted owning a bilingual dictionary; 9 said they used it only for school; 1 student (Italian student B), moreover, indirectly stated that he preferred the online dictionary since he admitted using a paper dictionary only when he did not have an internet connection.

Only 2 students (Philippine student C, Italian student Q) said they had no Italian-English bilingual dictionary; 2 of the 3 non native Italian students who said they owned one, said they did not use it.

4. The Test: a Description

The main objective of the experiment was

a) to verify if students today, digital natives, are able to use the electronic version of a bilingual dictionary for reception activities without being trained to do so and

b) if the online version or app really help them more than a paper dictionary.

We prepared five English sentences taken and slightly modified from examples contained in the glosses of the entry **to break** in the Ragazzini dictionary. One sentences was constructed in the same way for to serve and another one for the word **fast**. The verb to **break** was chosen because in Italian it gives the form in – **si** for the corresponding use of the English intransitive and has a vast phraseology.

The instructions asked students to do something completely new for them "Translate the following sentences." It also asked them to indicate which part of the gloss they used to identify the correct translation (See Appendix 2)

"Translate the following sentences."

"Which entry and which part of the dictionary entry did you look up to translate the words in bold?"

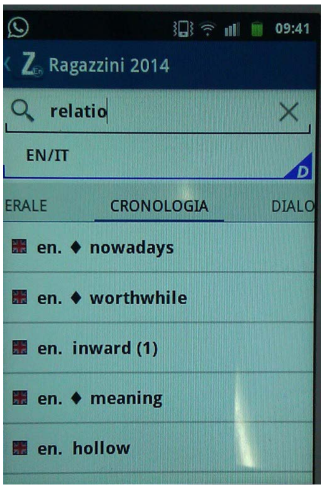
An example was provided of what had do be done.

In any case, it was necessary to explain to the class what an entry was.

[5] The students with the best results were C, L, N, O, Q . C and L are not native Italian speakers and most of their errors were spelling mistakes, for ex. the suffix *–ful* written with two ll.

[6] The publishing house Zanichelli generously provided the class with 5 apps, access to the online dictionary and paper copies of the dictionary.

Despite the bold character and underlining, students often disregarded the second part of the request; however, behind each student was an observer that took notes of the ways they consulted the dictionary. Only the app version, and not the one online, allowed the researcher or teacher to access a type of log file using the "chronology" of the words searched (see fig. 1)



09:41

Ragazzini 2014

relatio

EN/IT

ERALE CRONOLOGIA DIALO

en. ♦ nowadays

en. ♦ worthwhile

en. inward (1)

en. ♦ meaning

en. hollow

Figure 1: Search Chronology in Ragazzini 2014 app.

The observers were university and Ph. D students from the Torino University Department of Foreign Languages and Literature and Modern Culture trained to observe a series of behaviours (see the observation protocol given to the observers in Italian and reproduced in appendix 3 in an English translation). After the experiment the observers were invited to write a report on their experience and it became clear that many of the tested students were not really able to understand what a headword was and go straight to the entry with the right part of speech. The students' inability to choose the right English entry in the presence of homonymic headwords is widely reported in literature with students of different ages and mother tongues.

Here are two significant excerpts from the observers' reports:

> The student was not able to lemmatize and looked up words as they were written in the text (for ex. He looked up "cheaper", then "cheape"; "broke" etc).
>
> What probably struck me most was that he checked the lexical entry of the noun "break" as well as the verb "to break" many many times, going back and forth from one to the other.

4.1 Two “Simple” Tasks

Two of the sentences to translate which contained different meanings of the verb *break* were:

The engine broke when he tried to speed up and I need some small change, can you break a 5€ note?

These examples were chosen because they are functionally closer to the type of English students use in a technical high school course and, therefore, easy to translate. This conviction was also based on the fact that the meaning of the verb in the sentences was indicated as a level A2 in the Dictionary of the English Vocabulary Profile.

> **To Break [omission]**
>
> **NOT WORK**
>
> A2 [I or T] If you break a machine, object, etc., or if it breaks, it stops working because it is damaged.
>
> **Dictionary example:**
>
> I think I’ve broken your camera.

In reality, the first problem was that students looked for **broke** in the search window and found broke /brəʊk/ 🔈

> A pass. di **to break**
>
> B a.*(fam.)* senza soldi; in bolletta; spiantato; rovinato; fallito: **I’m broke,** non ho un soldo; sono in bolletta; **<u>to go</u> broke,** andare in rovina; fallire; (fam.) **<u>flat</u> broke,** completamente al verde ● (fam.) **to go for broke,** rischiare il tutto per tutto.

A second problem occurred when they were told to look for **to break:** 5 students out of 17 forgot that it was a past tense and translated it in the present. When indicating the part of the lexicographic entry in which they found the correct translation, student I says to broke v.t. 3, student B more correctly indicates v.i. 1[7], while M and R indicate to break v.t.1.

[7] In fact, in Italian the translation is *si ruppe* or *si è rotto,* therefore, it is an intransitive pronominal verb form. Those students who indicated the intransitive verb 1 paid attention to the meaning *rompere; infrangere; spezzare* and then translated with the *si* needed.

Student N mistook *engine for engineer and translated engineer broke it when he tried to make it go faster*[8] : such a mistake might be explained by the fact that he was using the online version, where if you type *engine* in the search window some other words also appear listed below the word entered. They are words that follow in the alphabetical macrostructure: in this case the words that appear are *engined* and engineer.

For *I need some small change, can you* ***break*** *a 5€ note?* We were counting on the fact that in the first part of the sentence there was the very frequent *I need* (A1) and the noun phrase *small change,* which we assumed the students would know: in the English Vocabulary Profile the meaning of the noun change is B1 and contains the same phrase *small change*

Change [omission]

COINS

> B1 [U] money which is coins rather than notes
>
> **Dictionary examples:**
>
> She gave me £5 in change.
>
> My dad always used to carry a lot of loose/small change in his pocket.

Unfortunately a good 12 out of 17 students took *small change* to mean ‘piccolo cambio’, mistaking it for an exchange of currencies; this did not invalidate the correct translation “puoi cambiarmi 5€?” in the second part of the sentence.

Only two translated it perfectly: student G and R who is Romanian.

Ho bisogno di spiccioli, puoi cambiarmi una banconota da 5€?

Mi servono delle monete, potresti cambiarmi una banconota da 5€?

And another two got very close (student O and Q who we had already noted as two of the best in the pre-test) translating with:

Ho bisogno di **pezzi piccoli** (literally. small pieces), puoi cambiare un 5€?

Ho bisogno di un **taglio più piccolo** (literally. lower denomination), potrebbe cambiare una banconota da 5€?

[8] The various translations ‘accendere’, ‘velocizzare’, ‘ andare avanti ‘of *to speed up* will not be discussed here – the electronic versions often show the headword of phrasal verbs with labels of specialised fields such as *autom(obilismo)*
■ speed up
A v. t. + avv.1 *(autom., ecc.)* accelerare 2 sveltire; velocizzare: **to speed up production,** sveltire la produzione
B v. i. + avv.1 affrettarsi; affrettare il passo 2 *(autom., ecc.)* accelerare.

Student H translated word for word

Io ho bisogno di una piccola variazione, potrei interrompere per 5€ facendo attenzione

Word for word: I need a small variation, I might interrupt for 5€ paying attention

This translation is strikingly similar to the one obtained using the site[9] http://translate.google.it; however, it is clear that a speck of human imagination played its part in constructing something different in the second part[10].

*I need some small change, can you **break** a 5€ note?*

Ho bisogno di qualche piccola modifica, si può rompere una nota di 5 €?

Word for word I need some small adjustments, is it possible to tear a note of 5 €?

The two students that translated well used different formats, the Italian G the app and the Romanian R a paper dictionary and said, correctly, that they took information from meaning 5 of the transitive verb:

> **to break** [omission]
>
> 5 cambiare *(una banconota, spec. pagando qc. e ricevendo un resto);* spicciolare: to break a £50 note, *cambiare un biglietto da 50 sterline*

The students that translated *piccolo cambio* managed, however, to get the translation of the second part of the sentence and they either did not indicate anything or they indicated coherently meaning 4, since they had translated *puoi dividere in valute da 5€? Puoi dividere una banconota da 5 €?*

> **to break** [omission]
>
> 4 suddividere; dividere; frazionare: **to break a word into syllables,** dividere una parola in sillabe

Those who translated *piccolo cambio* used the bilingual dictionary the way you do when, you do not have an overall ideal of the meaning of the sentence so you form a plausible hypothesis and make the rest fit around it.

The Ragazzini bilingual dictionary gave every imaginable form of support possible: under *change* noun meaning 3 is dedicated to the equivalent 'spiccioli' with the expressions *loose* and *small change* translated as *spiccioli;* there is even a USAGE NOTE that, if read carefully, would have helped the student to avoid confusing *change* and *exchange.*

[9] Consulted 19 April 2014

[10] *You* is disregarded and continued as an *I; fare attenzione* is the first equivalent of the verb *to note* in Ragazzini 2013

> NOTA D'USO
>
> **change** o **exchange?** Quando si parla del tasso di cambio tra valute non si usa il sostantivo *change,* ma *exchange: What's the exchange rate of the euro against the dollar?* qual è il tasso di cambio tra euro e dollaro? (non *What's the change of euros and dollars?). Change* viene usato in relazione al denaro per indicare il "resto" o gli spiccioli in moneta: *Did the cashier give you the right change?* il cassiere ti ha dato il resto giusto?; *Do you have any change at all?* hai della moneta?
>
> USAGE NOTE
>
> **change** or **exchange**? When you talk about the exchange rate of currencies you do not use change but exchange: *What's the exchange rate of the euro against the dollar?* qual è il tasso di cambio tra euro e dollaro? (not *What's the change of euros and dollars?).* Change is used for money to indicate what "remains" or coins: *Did the cashier give you the right change?* il cassiere ti ha dato il resto giusto?; *Do you have any change at all?* hai della moneta?

Moreover, it should be noted that the instructions also contributed to the problem since they led students to start with the word **in bold**; we are convinced that had we put *small change* in bold the error would not have occurred, because the processing left to right would have led to the identification of the expression first.

On the other hand, it would be wrong to draw the conclusion from the results of our "easy" tasks that the bilingual dictionaries are not helpful or worse, lead to mistakes: users normally use them for texts on topics they are familiar with. We were interested in whether students could go from the example to the meaning and we wanted them to look up the word **break** many times, also demonstrating that they had identified the correct part of speech called for by the various sentences.

It has been correctly observed that the task of associating meanings in a dictionary to the contexts of use is a complex task and in some ways metalinguistically unnatural. Still more difficult and often impossible to decide is the task of matching concrete examples to the appropriate specific meanings given in a dictionary and this is not because a dictionary is badly arranged, but because it is possible for circumstances to exist in which a combination of families of meanings is applied collectively in a context of use and not necessarily just one isolated meaning. (Chiari 2012: 116)

To translate isolated sentences is unnatural, but the task was important to evidence the problems which occur when consulting a dictionary. The second part of the test, in fact, was based on a much more natural understanding of a rather easy text, and the same students carried out the task correctly, using the dictionary very little or not at all, as we expected, having encouraged them to use it only when needed.

4.2 A Difficult Task

Students were given a clearly more difficult task when asked to translate the sentence

*In a **fast break**, a team attempts to move the ball up court and into scoring position as quickly as possible*

because it was a matter of a) – understanding how to translate technical terminology that was not entered as a multi-word headword, and b) – clarifying that under **fast** and under **break** not much would be found. Even if the sentence was itself a definition of *fast break,* in the Ragazzini 2013 under **fast (2) adjective**[11] there is nothing very useful, under **break** noun you need to go to the end of the list of meanings to find something useful, but not decisive, because just the equivalents of *break* are given.

> **Break [omission]**
>
> 23 *(calcio, ecc.)* incursione; penetrazione; discesa
>
> 24 *(basket)* break ; sfondamento; vantaggio (o svantaggio) incolmabile .

Real answers are given by the *full text*[12] search mode of both words: when the expression is searched, two entries come up, in English **to stop** and in Italian **contropiede.** Clicking on **to stop** the app automatically brings it up exactly where it appears in the phraseology of the very long entry of the word,

> *(basket, calcio, ecc.)* **to stop a fast break,** stoppare un contropiede

fast break is highlighted in yellow in the app version, framed in red online. Whoever knows something about sports understands that *contropiede* is the correct translation; clicking on **contropiede,** noun, masculine, the first meaning comes up:

> *(sport:calcio)* counter-attack, fast break

[11] The lexicographic norms always list the homonym that has part of speech noun first, followed by the adjective and by the verb.
In this case **fast (1)** is the noun that we translate as *digiuno*

[12] It should be noted that the only suggestion given to students before the test was "Remember that in the app and online versions in addition to the word search there is a full text search". In this specific case an observer gave the student misleading advice suggesting that he look for the "words in bold" separately, while the student had initially looked for the two words.

With the equivalent *fast break* highlighted in yellow in the app version and framed in red online, the observers recorded that

> The exercise that wasted the most time was the one containing the expression "fast break", as well as the exercise containing the idiom "break a leg" (I believe they were the exercises that took the student the longest time, about 16 and 34 minutes).

Only three students, F, L, Q out of 17 were able to translate *in un contropiede* and they all were using the online site. One did not indicate the part of the entry that helped him and the other two indicated the useful parts of the entry *fast break v.i. 1* fast *break 4* and that, if referred to the verb **to break,** they did not make sense. As for the other students, three did not translate, six opted for 'pausa veloce' ("fast pause"), one for 'azione veloce' ("fast action") and one for the most sensible 'sfondamento veloce' ("fast smashing")[13], indicating meaning 24 as his source of information. It can be concluded that a third of the class attributed the meaning 'pausa' to *break*, familiar to them because the word is used in Italian as a loan from English with this meaning and the meaning in tennis.

In this exercise having a paper dictionary was a handicap because only with the full text search can the equivalent be found and as De Schryver (2003: 146) observed only the implementation of fully integrated hypermedia access structures makes electronic dictionaries really different from their paper counterparts.[14]

5 Full Text Search is not like a "Google Search"

The full text search is not like a "google search", as emerged from the students' translations of the sentence

*It **served** him **right** to fail the exam: he had never studied hard*

When a few of the 7 that had translated well indicate "to serve 10" as part of the dictionary lexicographic article that helped them to translate, they really meant to indicate the phraseology which follows meaning 10.

> **to serve** /sɜːv/ 🔈
>
> v. t. e i. [omission]
>
> 10 (naut.: della marea) essere favorevole

[13] Also the translation of *to move the ball up court* caused significant problems, but we won't discuss them here.

[14] In regards to hypermedia access structures, it should be noted that in the online version every word of the gloss can be clicked on, bringing up the entry that word belongs to, but in the app version this mode was not implemented, probably to use less memory space in the smartphone.

> ● (mil.) **to serve as an officer,** prestare servizio come ufficiale; **to serve as a reminder [as a spoon],** servire da promemoria [da cucchiaio]; **to serve at table,** servire ai tavoli; **to serve behind the counter,** servire (o stare) al banco (in un negozio, ecc.); (mil.) **to serve a gun,** servire un pezzo; caricare un cannone; (fig. fam.) **to serve sb. hand and foot,** servire q. di barba e di capelli [...]; **to serve a purpose,** servire a uno scopo; **to serve sb.'s purpose,** servire a q.; andare bene (lo stesso): *I haven't got a screwdriver, but a knife will serve my purpose,* non ho un cacciavite, ma un coltello va bene lo stesso; **to serve sb. right,** trattare q. come si merita; (impers.) meritarsi: *It served him right to lose his job: he was always taking time off for no reason,* il licenziamento se l'è meritato: faceva sempre assenze ingiustificate (omissis)

What can be noticed is how they make the example fit: *il licenziamento becomes il fallimento dell'esame* in a good three students.
Il fallimento dell'esame se l'è meritato: non ha mai studiato
In this way the Italian in the translation is not wrong but rather unlikely. Another four, the same group L, N, O, Q as before, plus A, F, L, choose a more appropriate Italian translation *Si è meritato/ se lo è meritato di fallire l'esame; non ha mai studiato (duramente, bene, abbastanza)*
Student A translates *it served* in the present but is not the only one to forget it is in the past.
Si merita di non passare l'esame: perché non ha mai studiato
At the same time, however, he is the only one who does not fall into the trap *to fail the exam = fallire l'esame* , a translation that the students did not check in the Ragazzini that gives under **exam** **" to fail an exam,** non passare un esame" and under

> **to fail** v.t.
> [omissis]
> 2 non superare (un esame); essere respinto (o bocciato) in: to fail one's driving test, non superare l'esame di guida; *I failed maths,* sono stato bocciato (o mi hanno bocciato) in matematica

Reading quickly through the other ten translations of to serve and then right, it becomes clear that the translation task was not so banal. We cite them here:

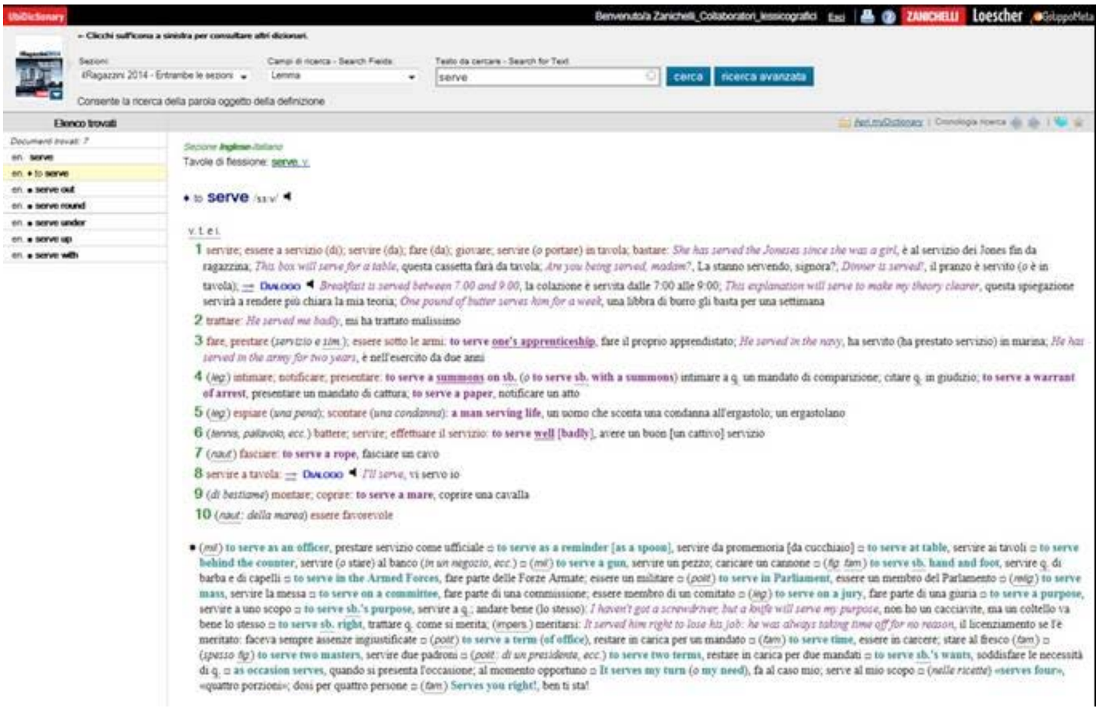
Esso ha suggerito correttamente a lui per fallire l'esame: lui non ha mai studiato duramente.
L'ha aiutato in modo corretto per non essere bocciato all'esame: non ha studiato molto.
Ha fatto il suo esame giusto ma lo ha sbagliato: non aveva studiato bene.
Ha servito di lei la cosa giusta per farlo fallire all'esame, non ha mai studiato tanto.
L'ha aiutato correttamente per fare l'esame, lui non ha mai studiato tanto.
Gli bastava a lui giusto a fallire l'esame: lui non aveva mai studiato intensamente
Gli sta bene di fallire l'esame: non ha mai studiato tanto.

The last translation uses the final indication in the Ragazzini phraseology:

> (fam.) **Serves you right!,** *ben ti sta!*

which is – very inappropriately – detached from **to serve sb. right.** And it is translated by a student using the paper version, acting almost as a confirmation of the experiment by Tono (2011), which concludes that we read the first and the last parts of a gloss.[15] The user of the app does not skim through other parts once he has found the example.

Student O used the paper version and the other six, the one online. In this case the app seems to penalize the students, if the search starts with the entry **to serve,** because the visualization on the limited display implies having enough patience to skim through the gloss many times to get to the phraseology; on the other hand, the entire entry fits on the computer screen (see fig. 2).



UbiDictionary
Benvenuto/a Zanichelli_Collaboratori_lessicografici Esci
ZANICHELLI
Loescher
GruppoMeta
← Clicchi sull'icona a sinistra per consultare altri dizionari.
Sezioni
Campi di ricerca - Search Fields
Testo da cercare - Search for Text
cerca
ricerca avanzata
Consente la ricerca della parola oggetto della definizione
Elenco trovati
Apri myDictionary | Cronologia ricerca
Documenti trovati: 7
en. serve
en. ♦ to serve
en. ■ serve out
en. ■ serve round
en. ■ serve under
en. ■ serve up
en. ■ serve with

Sezioni: ilRagazzini 2014 - Entrambe le sezioni

Campi di ricerca - Search Fields: Lemma

Testo da cercare - Search for Text: serve

Sezione Inglese-Italiano

Tavole di flessione: serve, v.

♦ to **serve** /sɜːv/

v. t. e i.

1 servire; essere a servizio (di); servire (da); fare (da); giovare; servire (o portare) in tavola; bastare: *She has served the Joneses since she was a girl*, è al servizio dei Jones fin da ragazzina; *This box will serve for a table*, questa cassetta farà da tavola; *Are you being served, madam?*, La stanno servendo, signora?; *Dinner is served!*, il pranzo è servito (o è in tavola); → DIALOGO *Breakfast is served between 7.00 and 9.00*, la colazione è servita dalle 7:00 alle 9:00; *This explanation will serve to make my theory clearer*, questa spiegazione servirà a rendere più chiara la mia teoria; *One pound of butter serves him for a week*, una libbra di burro gli basta per una settimana

2 trattare: *He served me badly*, mi ha trattato malissimo

3 fare, prestare (*servizio e sim.*); essere sotto le armi: **to serve one's apprenticeship**, fare il proprio apprendistato; *He served in the navy*, ha servito (ha prestato servizio) in marina; *He has served in the army for two years*, è nell'esercito da due anni

4 (*leg.*) intimare, notificare, presentare: **to serve a summons on sb.** (o **to serve sb. with a summons**) intimare a q. un mandato di comparizione; citare q. in giudizio; **to serve a warrant of arrest**, presentare un mandato di cattura; **to serve a paper**, notificare un atto

5 (*leg.*) espiare (*una pena*); scontare (*una condanna*): **a man serving life**, un uomo che sconta una condanna all'ergastolo; un ergastolano

6 (*tennis, pallavolo, ecc.*) battere; servire; effettuare il servizio: **to serve well [badly]**, avere un buon [un cattivo] servizio

7 (*naut.*) fasciare: **to serve a rope**, fasciare un cavo

8 servire a tavola: → DIALOGO *I'll serve*, vi servo io

9 (*di bestiame*) montare; coprire: **to serve a mare**, coprire una cavalla

10 (*naut.: della marea*) essere favorevole

■ (*mil.*) **to serve as an officer**, prestare servizio come ufficiale □ **to serve as a reminder [as a spoon]**, servire da promemoria [da cucchiaio] □ **to serve at table**, servire ai tavoli □ **to serve behind the counter**, servire (o stare) al banco (*in un negozio, ecc.*) □ (*mil.*) **to serve a gun**, servire un pezzo; caricare un cannone □ (*fig. fam.*) **to serve sb. hand and foot**, servire q. di barba e di capelli □ **to serve in the Armed Forces**, fare parte delle Forze Armate; essere un militare □ (*polit.*) **to serve in Parliament**, essere un membro del Parlamento □ (*relig.*) **to serve mass**, servire la messa □ **to serve on a committee**, fare parte di una commissione; essere membro di un comitato □ (*leg.*) **to serve on a jury**, fare parte di una giuria □ **to serve a purpose**, servire a uno scopo □ **to serve sb.'s purpose**, servire a q.; andare bene (lo stesso): *I haven't got a screwdriver, but a knife will serve my purpose*, non ho un cacciavite, ma un coltello va bene lo stesso □ **to serve sb. right**, trattare q. come si merita; (*impers.*) meritarsi: *It served him right to lose his job: he was always taking time off for no reason*, il licenziamento se l'è meritato: faceva sempre assenze ingiustificate □ (*polit.*) **to serve a term (of office)**, restare in carica per un mandato □ (*fam.*) **to serve time**, essere in carcere; stare al fresco (*fam.*) □ (*spesso fig.*) **to serve two masters**, servire due padroni □ (*polit.: di un presidente, ecc.*) **to serve two terms**, restare in carica per due mandati □ **to serve sb.'s wants**, soddisfare le necessità di q. □ **as occasion serves**, quando si presenta l'occasione; al momento opportuno □ **It serves my turn** (o **my need**), fa al caso mio; serve al mio scopo □ (*nelle ricette*) «**serves four**», «quattro porzioni»; dosi per quattro persone □ (*fam.*) **Serves you right!**, ben ti sta!

Figure 2: Entry to serve Ragazzini 2014 on line.

If the student had looked for the full text he wouldn't have found anything because it is one of the cases in which the user needs to substitute the variable part with placeholder pronouns. In other words, he needs to know how to go from *to serve him right* > **to serve sb. right.** Only this precise mode of set-

[15] An observer reported that the his student began to realize that he had to read the whole lexicographic article and not just the beginning and the end of the big articles only after three exercises. Three weeks later, a post-test revealed that the lesson had been learned; students continued to read the entire entry even though it was a comprehension test and not translation.

ting up the full text search will bring up to serve, but it does not place it at exactly the right point of the article, nor is there the yellow highlighting or red framing that there were for **fast break.**

The same problem emerged while trying to do the first exercise in the second part of the test reproduced below.[16]

> Massive and unreliable, the first computers of thirty years ago are as dead as the dinosaur. **Today, computers which are 30,000 times smaller and 10,000 times cheaper can beat them hollow**
>
> What does **the part in bold mean?** It means
>
> a) today computers have less hollow space in their cases
>
> b) modern computers can beat easily thirty year old computers
>
> c) thirty year old computers are as dead as the dinosaur.
>
> d) thirty year old computers can still beat modern computers
>
> Was context enough to understand the meaning? □ Yes □ No
>
> Which dictionary item helped you in answering the question?

In this exercise the bold print was much longer than the idiomatic expression, therefore the student should have identified the limits of the idiomatic expression and substitute them with the placeholder sb.; looking for **beat sb. hollow** in Full text brought up both **to beat** and **hollow** and it is found in their phraseology.

In the second part of the text, the task was not translating but choosing the right answer from four distractors; a clearly easier task and in fact a good 12 out of 17 gave answer b; three chose a) " today computers have less hollow space in their cases" and another two chose d). Only one student answered no to the first question, but identified the right answer. Seven said that consulting the entry **hollow** helped them, two found help in **beat**

Seven did not indicate anything, perhaps because they answered Yes to the question "Was context enough to understand the meaning? " and did not use their dictionaries.

The fact that in an electronic dictionary or apps typing in sequences of idioms with variables produces the response "no results found" rather upset the students who were expecting it to work the same way Google[17] does, where *them* automatically leads back to *somebody* and typing in *beat them hollow* brings up as the first site http://idioms.thefreedictionary.com/beat+hollow

> **beat somebody hollow** (British & Australian)

[16] The passage used in the first multiple choice exercise was taken from a B1 level English text formed by 432 words.

[17] Typing *beat them hollow* in http://translate.google.it/#en/it/beat%20them%20hollow (19/04/2014) brought up the translation *batterli cava,* literally *beat them empty* adj. feminine singular or n. *quarry*

> to defeat someone easily and by a large amount We played my brother's school at football and beat them hollow.
>
> See also: beat, hollow
>
> Cambridge Idioms Dictionary, 2nd ed. Copyright © Cambridge University Press 2006. Reproduced with permission.

Essentially students would like something more than an electronic dictionary; they would like what Tarp (2008:123) called a leximat "a lexicographical tool consisting of a search engine with access to a database and/or the internet, enabling users with a specific type of communicative or cognitive need to gain access via active or passive searching to lexicographical data".

Here we have not dealt with tasks of consultation tied to production activities, but a leximat would be even more welcome when producing a text in L2.

6 Conclusions

The experiment has shown that without knowing how to consult a dictionary the user does not take full advantage of the electronic dictionary and, on the other hand, that a knowledgeable user is able to get the information needed from a paper dictionary just as well.

Compared to searching an electronic dictionary, consulting a paper version is slower, but penalizes the user only when looking up idioms or multiword units.

Regardless of the lexicographic tradition in lemmatization, the first 3000 most frequent English words are characterised by a peculiar morphological poverty in part of speech markers which makes homonymy abundant in the macrostructure. Students should at least be aware of this before using a handheld dictionary.

To deal with the polysemy, when opening long lexicographic article in the app an index, a menu of the microstructure -including signposts- should be adopted, similar to the ones already found in English learners dictionaries (see Medal), Latin and Greek dictionaries (see Montanari) and in the VanDale and Sansoni bilingual dictionaries published in the 1980s (see Marello 1989: 77-98)

Specialized field labels abound in Italian monolingual and bilingual lexicography but are not used enough by unskilled users, whether they are in paper or electronic versions.

Searches using jolly characters were not tested. Jolly characters are available only in the electronic dictionaries for purchase and not in the ones free on the Internet. Our observers, however, reported that students used a similar type of search, not based on morpheme boundaries and without jolly characters, taking advantage of the list of alphabetically near headwords which comes up on the electronic dictionary when the first letter are typed.[18]

[18] For example, when typing *common* commonable*, commonage, commonality, commonalty, commoner, commonhold, commonly, commonness, commonplace, commonplaceness, commonweal, commonwealth* come up on the screen.

Observers reported that students gradually understand how an electronic dictionary works and they do not need a lot of training: they need practice, as noted previously, with the homonymy in the core English lexicon, also because in the so-called collaborative bottom-up dictionaries, produced by volunteers, like Wictionary and Wikizionario, homographs are not separated. As Chiari observes (2012: 108-109) this difference significantly changes the very idea that Wikizionario gives of a headword, which ends up matching the form keyed in and is strongly influenced by the user's need for a keypad search, so that it even includes some inflected forms. For now dictionaries are not considering the option "perhaps you were looking for..." but only give cross references similar to the one seen for broke in § 4.1. This is another "defect" our students accuse the dictionaries of, whether, apps or even more rightly so, online on computers, because it represents a valid aid especially for languages with complex spelling like English.

If students had been trained beforehand and given a common basis of elementary knowledge of dictionary use, the difference in the time of execution for online users and app users might have emerged. The difference between users of the paper version and of electronic versions are only indirectly revealing in our experiment, because 2 out of 5 students with paper dictionaries were non native speakers of Italian and another 2 were among the best students in the class. The latter two finished before the others both the test section requiring dictionary use and the comprehension in which they used the dictionary considerably less than the others while achieving a very good score. Students' proficiency influenced both score and speed in completing the test, more than the lexicographic tool.[19]

The convenience of having a dictionary on a smartphone is sufficient to justify allowing its use in the classroom: an expert user practices his/her ability to use a dictionary and a weaker one carries out a (meta)linguistic activity that, in any case, has important repercussions on his/her knowledge and ability to search more complex databases. However, interviewed on whether or not they would purchase or suggest purchasing the app for a bilingual dictionary to use on their own smartphones, the students in the class seemed perplexed since they were basically convinced that the translation programs and dictionaries free online were good enough for their extra-scholastic needs. As for their scholastic needs, so far they have been limited, but that could change with the introduction of teaching methods based on Content Language Integrated Learning. We hope to repeat the experiment in types of secondary schools where teachers are preparing students of the same age for university studies, and we dare say that having a good bilingual dictionary at hand - on a smartphone or tablet – and knowing how to use it might increase these students' understanding of important details in texts as well.

[19] Their English teacher probably meant to help the two non native speakers giving them the paper dictionary and knew that for the best students a paper version would not be a handicap. See the contributions in progress by Elisa Corino and Elena Martra on these aspects. Of course, if we had performed the test in a lab we might have selected the coupling tool-testee differently, in order to obtain more telling results, but when an experiment is performed in a true class, the research has to respect the teacher's educational concern.

7 References

Augustyn P. (2013). No dictionaries in the classroom: translation equivalents and vocabulary acquisition. In *International Journal of Lexicography,* 26 (3), pp. 362–385

Bergenholtz, H. Johnsen, M. (2007). Log files can and should be prepared for a functionalistic approach. In *Lexikos*, No.17, pp. 1-21.

Butzkamm, W. (2009). *The Bilingual Reform: A Paradigm Shift in Foreign Language Teaching.* Tübingen: Gunter Narr Verlag.

Cadierno, T. (2008). Learning to talk about motion in a second language, in P. Robinson, N. Ellis, *Handbook of Cognitive Linguistics and Second Language Acquisition,* London: Routledge, pp. 239-275.

Chiari I. (2012). Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto. In *Bollettino di Italianistica. Per Tullio De Mauro,* II, pp. 94-125.

Cook, G. (2010). *Translation in Language Teaching: An Argument for Reassessment.* Oxford: Oxford University Press.

de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-dictionary Age. In *International Journal Lexicography,* 16 (3), pp. 143–199.

Hjelmslev L. (1961[2]). *Prolegomena to a Theory of Language.* Madison:The University of Wisconsin Press

Laufer, B. and N. Girsai. (2008). Form-focused Instruction in Second Language Vocabulary Acquisition: A Case for Contrastive Analysis and Translation. In *Applied Linguistics,* 39 (4), pp. 694–716.

Macmillan English Dictionary Online. Accessed at: http://www.macmillandictionary.com [19/04/2014].

Marello C. (1989). *Dizionari bilingui con schede sui dizionari italiani per francese, inglese, spagnolo, tedesco.* Bologna: Zanichelli

Marello C. (1998). Hornby's Bilingualised Dictionaries. In *International Journal of Lexicography,* Special Issue 1998, pp. 292-314

Montanari F. (1995, 2013[3]) GI - *Vocabolario della lingua* greca.Torino: Loescher

Il Ragazzini (2013, 2014) *Dizionario Inglese-Italiano Italian-English Dictionary* Bologna: Zanichelli; 2014 online version and app for iPhone e iPad

Slobin, D. (1997). *The crosslinguistic study of Language Acquisition,* vol.5: Expanding the contexts, cap.1.3: The cross-Typological Approach. London: Erlbaum, pp. 11-34.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge.* Tübingen: Niemeyer

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

Acknowledgements

I would like to express our appreciation to the class IV E of the Avogadro technical high school, their English teacher, Elena Martra and the principal Tommaso Deluca, as well as to the publishing house, Zanichelli. We would also like to thank Elisa Corino for her assistance in preparing the tests and conducting the project, Andrea Belluccia for the tabulation, the observers for their observations and Kathy Metzger for the translation.

Appendix 1

Use the word in the box to form a word that fits in the text

How does the video work?

When I was young, I always dreamed of becoming a famous (1) When I was at school I decided to study engineering, and then become a millionaire by inventing a wonderful new (2) which would make the world a better place. Unfortunately, I wasn't very good at technical subjects. Any time I operate any kind of (3) , something terrible happens. Machines which use (4) , such as computers or televisions, always seem to give me a (5) shock. The instruction booklets are always (6) They never help me at all. Nowadays you need to have specialised knowledge just to turn on the video. To my great embarassment it is always a child of six who helps me out of my (7) .

1 SCIENCE ....
2 PRODUCE ....
3 EQUIP..
4 ELECTRIC ....
5 POWER ....
6 USE..
7 DIFFICULT ....

From: Michael Vince, First Certificate Language Practice, Heinemann, Oxford, 1996, pag. 235

Appendix 2

Translate the following sentences. **Which entry and which part of the dictionary entry** did you look up to translate the words in bold?

*Ex. Her bones **break** easily*

Le sue ossa si rompono facilmente

Entry and part of the dictionary entry   to break B v.i. 1

*The engine **broke** when he tried to speed up*

Entry and part of the dictionary entry

*The news **broke** and everybody knew the truth*

Entry and part of the dictionary entry

*Well-wishers typically say "**Break a leg**" to actors and musicians before they go on stage*

Entry and part of the dictionary entry

I need some small change, can you ***break*** a 5€ note?

Entry and part of the dictionary entry

In a ***fast break***, a team attempts to move the ball up court and into scoring position as quickly as possible

Entry and part of the dictionary entry

It ***served*** him ***right*** to fail the exam: he had never studied hard

Entry and part of the dictionary entry

You can put it in the washing machine, it's a ***fast***-colour T-shirt

Entry and part of the dictionary entry

Appendix 3

Observation Protocol for 5 March 2014

Objectives of the Experiment

(1) Establish the class's ability to use dictionaries

(2) Establish whether students are aware of the pros and cons of a paper dictionary compared to the electronic versions

(3) Establish whether students are aware of the pros and cons of an online dictionary compared to an app(lication).

(4) Exercises were prepared which explicitly required the use of a dictionary to verify if students were able to trace a statement back to the schematization of microstructures in the dictionary (profile

of the gloss in a bilingual dictionary) or to use the examples to find the correct translations (semantic equivalents).

(5) Reading comprehension exercises were prepared which could be done without a dictionary, if students believed they knew how to answer.

Your job is to observe how students work

(1) The time they spend on each exercise

(2) If they look up words other than the ones in bold in the exercise.

Ex. in *We have to accept the microchip, or face the alternative of leaving off the free world market* they look up **face** as well as **leave off**

(3) If they check the translations (equivalents) by looking them up in the section in which they are entries.

N.B. online all the words can be clicked on the smartphone no.

(4) If the students with electronic versions use the **full text search**

(5) If they know how to lemmatize (broke → break) or they get help from the online references

(6) If they go to the entry with **the right part of speech** without hesitating

(7) If, for idioms, they realize they are idiomatic or go to the phraseology section to look for it.

You need pen and paper to write down what they look up and the order they look things up for those that have paper copies. **Stop those with online access and smartphones BEFORE they close the session.** In this way we can take a picture of the list of entries looked up.

Meanings, Ideologies, and Learners’ Dictionaries

Rosamund Moon
University of Birmingham
R.E.Moon@bham.ac.uk

Abstract

This paper looks at the treatment of ideologically loaded items in monolingual learners’dictionaries of English, and issues in the lexicographical description of their meanings. It begins by considering non-denotative meaning and the question of evidence; then considers a selection of entries relating to ethnocentricity, gender and sexuality, and age. Entries are drawn mainly from the standard British big five; the 1948 edition of Hornby; and two crowd-sourced dictionaries.

Keywords: ageism; critical lexicography; culture; ethnocentricity; gender; ideology; learners’ dictionaries; meaning; sexism

1 Introduction

My talk at Euralex 2014 is concerned with the presentation of meanings in learners’ dictionaries of English: in particular, the meanings of words which denote, represent, or reflect politicized concepts and phenomena – ideologically loaded items, totemic and socioculturally significant. Such words have been the frequent focus of linguistic investigations more widely, for example in corpus-led studies from a discourse analytic perspective, or sociological and cultural studies (Raymond Williams’ discussions of “keywords” are a case in point). In relation to lexicography, ideology is where dictionaries collide with the social world: it brings in impolite and polite aspects of language, taboo items, evaluative orientation, connotation, and cultural allusion; the sublexicons, of course, of semantic fields such as politics, religion, ethnicity, sexuality, and so on; and above all the role of lexis, an unstable and mutable role, in naming and othering.

Ideologically positioned meaning is central to the concerns of critical lexicography, and particularly important with respect to learners’ dictionaries because of their positioning as global texts for a pluralist multicultural usership. It is a topic that regularly surfaces at Euralex congresses[1] – though historically, perhaps not as often as might be expected – and in lexicographical journals; it is covered, at least tangentially, in lexicographical manuals and metalexicographical monographs (Svensén 1993;

[1] Euralex papers include those by Coffey 2010; Harteveld & van Niekirk 1996; Iversen 2012; Schutz 2002; Swanepoel 2010; van der Meer 2008; Veisbergs 2000, 2002, 2004; Whitcut 1983; and several due at Euralex 2014: see the Euralex database of past proceedings at http://www.euralex.org/publications/ for these and other presentations on the topic.

Atkins & Rundell 2008: 422ff and passim; Landau 2001: 228ff and passim; Béjoint 2010), and more directly as the subject of a 1995 festschrift for Zgusta (eds. Kachru and Kahane). In returning to the topic today, I do not offer solutions, but there are, I think, a number of points that are still worth making in a 2014 context.

his written version of my talk provides an overview of topics to be discussed. In section 2, I look at non-denotative aspects of word meaning, the contribution of corpus evidence, and methodological issues; sections 3, 4, and 5 present and review a sample of words and dictionary explanations relevant to discussion of, respectively, ethnocentricity, gender and sexuality, and age/ageism. As a basis for observations, I draw largely on four current British monolingual learners' dictionaries of English (from Cambridge, Longman, Oxford, and Macmillan), specifically their free online versions as accessed in April 2014: by default, these four are the online learners' dictionaries referred to below. The fifth British learners' dictionary, *Cobuild*, is represented instead by its second print edition of 1995: partly because the current online version has restricted accessibility; partly because the 1995 text was based on an early version of the Bank of English corpus (BoE),[2] which I draw on below; and partly because of my own editorial involvement with the 1995 text (and as an editor of the first edition of 1987). Other dictionaries cited include the 1948 version of Hornby's first dictionary (ALD1); and the collaborative or crowd-sourced online texts *Simple English Wiktionary*[3] and *Urban Dictionary.*[4] It goes without saying that online dictionaries are complex multidimensional, multimodal texts, and although I focus mainly on linguistic explanations, examples, and labelling in my discussion, there are other parts of their entries where ideology is displayed and where othering may be performed.

2 Meanings: Culture, Connotation, Evaluation – and Evidence

In later sections, I will look at some words that are obvious sites for projection of contested and ideologically-bound attitudes. But many of the issues of lexicographical description overlap with more general issues of how and whether to represent non-denotative aspects of meaning, including connotation, evaluation, and culture. To demonstrate this, I want to look at the word *cardigan*, an item with a very clear concrete meaning and reference in the real world. It can be defined or explained straightforwardly, as in the current online entry in Oxford:[5]

[2] BoE is a 450-million word corpus, created by COBUILD at the University of Birmingham. 71% of its texts are British English, 21% North American, and 8% Australian, mainly drawn from the period 1990-2003; 86% of its texts are written, and the remaining 14% consists of transcribed spoken interaction and radio broadcasts.

[3] A simplified text, affiliated with Wiktionary, constructed with something of a controlled defining vocabulary, and claiming almost 22,000 definitions or "entries" in May 2014.

[4] In Fuertes Olivera's terminology, these are classifiable as "collective free multiple-language Internet reference works", in distinction to "institutional Internet reference works", which would include publisher-produced texts (2009: 103).

[5] To facilitate comparisons, here and below, I have standardized typography and layout irrespective of the original: headwords and phrases are in bold; where included, examples are italicized and set out on new lines; usage labels are italicized in parentheses.

(1) a knitted jacket made of wool, usually with no collar and fastened with buttons at the front

where it is accompanied by a photographic illustration (and adverts from clothing companies). Compare Macmillan's online entry, which combines a broadly comparable descriptive and denotative explanation with a "cultural note":

(2) a jacket knitted from wool, that you fasten at the front with buttons or a zipCultural note: cardigan

Cardigans are usually thought of as an old-fashioned, rather boring piece of clothing, worn mainly by older people.

We might dispute the explanations of cardigan (they can be made of fibres other than wool; not all cardigans have fastenings), as well as Macmillan's stereotyping comment (fashions change; babies, toddlers, schoolchildren wear cardigans), but there is an important point that even innocuous and simple items generate associations which reflect attitude and, in this case, ageism. Evidence for these associations can be detected in corpus data, as in this small sample from BoE, in particular the fifth, sixth, and seventh tokens:

```
this short green knitted dress and cardigan yesterday to promote the Irish
      Isabella was wearing a bulky cardigan with horizontal orange and red.
Next. Not long ago a pure cashmere cardigan alone would have cost about £300.
    cardigan Invest in our classic cardigan and you'll wonder what you ever
   maths teacher who favours comfy cardigans and looks like the perfect grandad
 it. There are plenty of ladies in cardigans and old gentlemen in ties. It
teeth; it meant shuffling around in cardigans and ranting about the youth of
up the sleeves of her grey knitted cardigan and got to work. <p> Robina went
      for the day, or under a long cardigan for evening. This month we are
was in the 80s, she wore a woollen cardigan buttoned to the neck and
```

Compare, too, two of the examples given in OED for the abbreviated form cardy (and discussion of middle-aged below):

(3) 1969 *Guardian* 3 Nov. 7/2 Grey gentlemen in shrunken cardies.

981 *Daily Tel[egraph]* 29 Aug. 11/2 A flock of over-50s wearing pastel cardis and floppy hats.

Words that are more obviously politicized offer particular scope for critical lexicography: examination of differences between explanations in dictionaries, unpacking of attitudes projected towards the concept under definition, and dominant ethos of the lexicographers – or publishing/cultural context. For example, the four current online learners' dictionaries broadly agree on what materialism is (in the non-philosophical sense):[6]

(Macmillan)

the belief that money and possessions are the most important aspects of human existence

[6] Cf. Williams 1988: 197-201; Bennett et al. 2005: 209-211.

(5) (Oxford)

(*usually disapproving)* the belief that money, possessions and physical comforts are more important than spiritual values *the greed and materialism of modern society*

but while Oxford adds a usage label to reflect sociocultural attitudes towards materialism and an example that indirectly condemns by collocation, Macmillan offers no evaluation at all of materialism as a modus vivendi.

Distinctions in attitude towards a concept can be subtle, as in these entries for *equality*[7] in the online dictionaries:

(6) (Cambridge)

the right of different groups of people to have a similar social position and receive the same treatment:

equality between the sexes

racial equality

the government department responsible for equalities

(7) (Longman)

a situation in which people have the same rights, advantages etc

equality of All people have the right to equality of opportunity.

equality with Women have yet to achieve full equality with men in the workplace.

quality between equality between men and women

racial/sexual equality The government must promote racial equality.

(8) (Oxford)

the fact of being equal in rights, status, advantages, etc:

racial/social/sexual equality

equality of opportunity

the principle of equality before the law (= the law treats everyone the same)

Don't you believe in equality between men and women?

The entries are broadly similar, all using examples to indicate arenas in which equality is an issue, particularly gender and race. But comparison of the genus words in explanations raises the question of what equality actually is – *fact* implies that it is a principle achieved; *situation* is non-committal; only Cambridge's *right* suggestions that it is more of an ideal than real or implemented in practice, something hinted at in the following small sample from BoE:

[7] Cf. Williams 1988: 117-119; Bennett et al. 2005: 109-111.

```
         in an atmosphere of ease, equality and immense opportunities for the
          the Commission for Racial Equality and other ethnic minority and
  is further evident that political equality, co-existing with an increasing
      relationship characterized by equality, disagreement and conflict are
aim of that document was to achieve equality for women by 2000. <p> The
      as a means of achieving human equality have been taken up by most of the
   movement to greater fairness and equality in schools seemed to go naturally
to ensure that the new regime makes equality of opportunity, in terms of access
I wish to take part. He believed in equality then and I believe in it now.
   in line with feminist notions of equality. Yet feminist therapists, most of
```

Top collocates of *equality* in BoE provide further data concerning its textual and contextual environments (see groupings below), and further hints of the elusive nature of equality:[8]

(9) (arenas) gay, gender, homosexual, lesbian, men, race, racial, sexes, sexual, women
(ideals and concepts) democracy, fairness, freedom, justice, liberty, rights
(pressure groups and committees) campaign, commission, council, struggle
(general items) commitment, economic, issues, law, legal, opportunity, political, principle*, social, society, treatment

Similarly, to return to *materialism*, collocational evidence helps identify both semantic/philosophical contexts of occurrence and negative attitudes towards it, as in this from BoE:

(10) (nouns) society, idealism, science, spirituality, greed, philosophy, hedonism, marxism, values, atheism, theory, feminism, selfishness, corruption, consumerism, ideology, capitalism, utilitarianism, pragmatism
(prenominal adjectives) historical, dialectical, scientific, cultural, Western, crass, rampant, modern, new, gross, Marxist, atheistic, vulgar, subversive, secular

(further supported by lower-frequency collocates such as *self-interest, self-centredness, emptiness, godless, mindless).* At the same time, such evidence is a reflex of the discourse world of the corpus: a normative view that perhaps reveals more about prevailing attitudes in the language-owning culture at the time of data capture than about materialism itself. So there is a particular dilemma facing lexicographers attempting to deal with contested and ideologically loaded words: to balance a description of what data suggests about meaning with how in a postmodern inclusive society, the relevant concept "ought" to be regarded and represented. The English word *civilized* is another case in point: cf. Moon 1989: 88-90, and see discussion below.

[8] For discussion of the role of collocation and co-text in relation to meaning, see for example Hanks 2013; Leech 1981: 16ff; Louw 1993; Sinclair 1991, 2004.

3 Ethnocentricity

It goes without saying that racist terms have to be labelled clearly as offensive in learners' dictionaries. However, the problem of ethnocentricity extends much further than labelling, something that has been explored at length by Benson (2001) and in Ogilvie (2013), as well as in multiple critical linguistic/lexicographical papers, including Krishnamurthy's examination of the words *ethnic, tribal, racial* in corpus and dictionary data (1996) and, for example, Hornscheidt's with respect to colonialist words in Danish dictionaries. In my own 1989 paper on ideology and lexicography, I was particularly concerned with the word civilized, its meanings, and dictionary representations of meaning. Here is the relevant entry from *ALD1*:

(11) civilize

1 bring out from a savage and ignorant state; give teaching in art, science, culture, good government, good customs and manners.

2 improve and educate.

Many a rough man has been civilized by his wife.

A product of its time and purpose, this now raises all sorts of questions: the paternalism of bring *out*; the subtext of *savage*; the meaning of *good*; and so on (similarly in the second extended sense, the use of *improve* and the sexism of its example). Fifty years later, and drawing on BoE corpus evidence, this is *Cobuild2's* explanation for **civilized**:

(12) 1 If you describe a society as **civilized**, you mean that it is advanced and has sensible laws and customs.

I believed that in civilized countries, torture had ended long ago.

≠ barbaric

2 If you describe a person or their behaviour as **civilized**, you mean that they are polite and reasonable.

I wrote to my ex-wife. She was very civilized about it.

Advanced? sensible? is *barbaric* really an appropriate antonym? and with the example in the second sense, do the implicatures make it seem as sexist as ALD1's? The next entries are from two online learners' dictionaries (extended senses have been omitted here):

(13) (Macmillan)

1 a civilized country, society etc has developed an advanced culture and institutions

A civilized society does not solve conflicts in a way that causes so much suffering.

(14) (Oxford)

1 well-organized socially with a very developed culture and way of life
the civilized world
rising crime in our so-called civilized societies
civilized peoples
2 having laws and customs that are fair and morally acceptable
No civilized country should allow such terrible injustices.

While examples demonstrate something of what's implied by *advanced, developed* and indeed *civilized* itself, the overall meaning is unclear, almost insiderist (only someone from a "civilized" society would identify with the description – a circularity of meaning that is as problematic as the circularity of inspecting entries for *advanced, barbaric, cultured, developed, primitive, savage, uncivilized* etc. in order to locate what *civilized* might mean). Meanwhile, corpus data reflects, inevitably, an anglocentric view of the world:

```
    to think of human beings as all civilised but they're not. Some remain
     terrorism. We British, and all civilised countries, should back America
    a world where a man who, in any civilised country, would - even though his
 fortunate brethren the benefits of civilized culture. Though more successful
ask itself the big question: If the civilised European can allow the
to a world seeking reassurance that civilised governments and legislatures
  that no country could call itself civilised if the sick are refused medical
 bullet, prohibited for use between civilised nations but sanctioned for big-
 seeds and causes of conflict among civilised nations will speedily appear. Of
them, they too, can usher in a more civilised order. The Chinese took on
       idealism can cause seemingly civilised people to misuse, even destroy,
       human rights violations in a civilised secular democracy. These are not
      that an important hallmark of civilised societies is the extent to which
    by such horror close at hand, a civilised society has a choice. It can act,
           taxation system. <p> Any civilised society should provide education
  one every civilised nation, every civilised state, including the great
  European Communism to be open and civilised. The phenomenon of Eurocommunism
         syndrome is one of the <f> civilized world's most common diseases. It
   too close. `We warn you that the civilised world objects to your aggressive
   I think that the free world, the civilised world, understands that this
```

This presents a near-insoluble problem: should an explanation in a learners' dictionary present this kind of traditionalist anglocentric monocultural world-view evident in corpus data, thus promoting a particular ideological stance? Should an apologist usage note be added? Should there be instead a broader, multicultural, universalist, non-elitist explanation, even at the expense of misrepresenting of what the English word is actually used to mean, what mindset it reflects?

With *ethnocentric* itself, Macmillan is clearest in indicating that it is a derogatory label for inegalitarian perspectives:

(15) showing a failure to recognize that other people's cultures are also important and valuable

while Longman's usage comment could be misinterpreted (exactly what is being disapproved of?):

(16) based on the idea that your own race, nation, group etc is better than any other – used in order to show disapproval:

and Oxford fails to indicate that ethnocentricity is evaluated negatively at all:

(17) based on the ideas and beliefs of one particular culture and using these to judge other cultures

Of the many items in the English lexicon which have potential for ethnocentric and racist usage, an important subset includes *foreign(er)* and other words which contribute to the othering of non-natives and non-nationals – cf. a critical linguistic study by Gabrielatos and Baker (2008), who are concerned with the discursive representation of refugees and asylum seekers in news media (and incidentally in definitions). The following brief discussion looks at *alien, asylum seeker, illegal immigrant/alien, migrant, refugee* in recent/current learners' dictionaries. Explanations are broadly comparable, but what's especially interesting is the selection of examples, particularly where these, in the decontextualized world of dictionary text, seem hortatory or imply moralistic value judgements. For example, these are *Cobuild2's* entries:

(18) migrant

1 A **migrant** is a person who moves from one place to another, especially in order to find work.
The government divides asylum-seekers into economic migrants and genuine refugees.
... migrant workers following harvest northwards.

(19) s.v. **illegal** 2

Illegal immigrants or workers have travelled into a country or are working without official permission. > Illegal immigrants or workers are sometimes referred to as **illegals**.
... a clothing factory where many other illegals also worked.

Examples in the following entries for *refugee* show typical collocations with *flee, flow, stream* etc. – collocates which have been shown elsewhere to contribute to the negative discursive construction of refugees as a social group (Gabrielatos & Baker 2008: 22ff; Semino 2008: 87ff):

(20) (Longman)

someone who has been forced to leave their country, especially during a war, or for political or religious reasons:
Refugees were streaming across the border.
refugee camps

(21) (Oxford)

a person who has been forced to leave their country or home, because there is a war or for political, religious or social reasons
a steady flow of refugees from the war zone
political/economic refugees
a refugee camp

(22) (Cambridge)

a person who has escaped from their own country for political, religious, or economic reasons or because of a war:

Thousands of refugees fled across the border.

Compare too the subtexts of examples in these:

(23) **asylum seeker** (Cambridge)

someone who leaves their own country for their safety, often for political reasons or because of war, and who travels to another country hoping that the government will protect them and allow them to live there:

genuine/bogus asylum seekers

(24) **alien** (Longman)

1 someone who is not a legal citizen of the country they are living or working in:

illegal aliens entering the country.

(25) **illegal** (Longman)

(American English spoken) an illegal immigrant:

Illegals are still slipping through in unacceptable numbers.

Also relevant, though this cannot be discussed in detail here, are the subtexts, the intertextual implicatures, which are created through the links to thesaurus entries or lists of semantically related words that are triggered automatically by searches for specific words. For example, Cambridge's entry for *refugee* displays a set "Runaways and refugees", listing items *boat people, deserter, displaced person, escapee, evacuee, fugitive, political asylum, refugee camp, transit camp.* Of course, these features are intended for vocabulary extension work, but many such examples of "interesting" juxtapositions can be found in online learners' dictionaries: these, by association, reinforce both othering and sociocultural evaluations.

4 Gender and Sexuality

The asymmetries of gendered nouns in English, together with the gendered collocational/semantic preferences of adjectives, have been widely discussed. With respect to dictionaries, discussion has tended to focus on sexism in general, asymmetric definitions of paired male/female terms, and representation of men and women in examples: see, for example, papers by Graham (1975), Whitcut (1984), Landau (1985), Barnickel (1999), Connor Martin (2005), etc., particularly with reference to orthodox (= androcentric) dictionaries; Russell (2012) examines feminist dictionaries which provide something of a counterdiscourse.

Where pairs of English words for (human) males and females are concerned, the lexicographical challenge is to balance two conflicting ideas. First, men and women in the UK, as in so many other nations, now have equal status legally and legislation protecting their rights. Second, the continuing disparities in practice between the lives of men and women, along with biological/physiological distinctions, are reflected in lexis and language use and in attitudes communicated through language. Thus decisions taken when designing and constructing entries for paired terms cannot just be linguistic decisions: they must inevitably be ideological as well.

A pair such as *boy* and *girl* demonstrate the issues. Their primary and simplest senses – non-adult male/female, son/daughter – are clear counterparts; however, their symmetry changes when they are used to refer to adults (cf. discussion by Caldas-Coulthard & Moon 2010; Holmes & Sigley 2001; Sigley & Holmes 2002). In particular, *girl* continues to be applied to young women, especially in their late teens and twenties, whereas *boy* is more likely to be replaced by another term: *young man* or informal *lad, guy* etc. Both *boys* and *girls* are used informally of groups of adult male/female friends, and groups of male workers (soldiers, police, fire fighters, sometimes factory operatives, etc.) or female workers (typically in low-status occupations). While these are infantilizing usages, they are also ambivalent: showing affection and solidarity if used by speakers who are part of the group concerned, or who position themselves as part of the group; but often paternalistic, condescending, or demeaning if used by outsiders or those with higher status.

It seems reasonable now to expect that dictionary entries for *boy* and *girl* would have parallel explanations for primary senses, then present information about the various usages that relate to adults, including register and potential for offence. However, there are some surprising asymmetries and inconsistencies. Those in *ALD1* in 1948 could be predicted:[9]

(26) **boy**

1 a male child up to the age of 17 or 18

2 a son (of any age)3 a male servant

(27) girl

1 a female child of any age; a daughter

2 a female child not yet grown up; one who is not yet married3 a maidservant

4 a girl or woman working in a shop, office, etc.5 (*colloq., vulg.*) a sweetheart

But there are also curious asymmetries in *Cobuild2*, written in the 1990s by a strongly pro-feminist team (as was the 1987 first edition):

(28) **boy**

1 A **boy** is a child who will grow up to be a man.

2 You can refer to a young man as a **boy**, especially when talking about relationships between boys and girls.

3 Someone's **boy** is their son; an informal use4 You can refer to a man as a **boy**, especially when you are talking about him in an affectionate way.

[9] Here and below, for reasons of space I have mostly omitted examples given in dictionary entries for boy and girl, though these too are interesting and very relevant to examinations of ideological stance and sexism.

5 You can use **boy** when giving instructions to a horse or dog.

(29) girl

1 A **girl** is a female child.

2 You can refer to someone's daughter as a **girl**.

3 Young women are often referred to as **girls**. Some people find this use offensive.

4 Some people refer to a man's girlfriend as his **girl**; an informal use.

Missing altogether is the use of *boys/girls* to refer to friendship groups, and the offensive, mainly old-fashioned American, use of *boy* to address an inferior. There seems no clear reason for the different wordings of *boy* 1, 3, and *girl* 1, 2; nor for the inclusion of sense 5 of *boy* (or conversely exclusion of a parallel vocative for female horses and dogs).

While current online versions of Longman, Macmillan, and Oxford have more symmetrical entries and explanations for at least the primary senses of *boy* and *girl*, Cambridge does not:

(30) **boy**

a male child or, more generally, a male of any age

Their little boy (= their young son) is very sick.

the boys a group of male friends(also **our boys**) an approving way of speaking about your country's soldiers

(31) girl

a female child or young woman, especially one still at school:

a daughter:

a woman worker, especially when seen as one of a group:

a group of female friends

Macmillan is representative of the other three in its parity and careful labelling (though it is still partially asymmetric); it also adds a usage note at **girl**:

(32) **boy**

1 a male child

a. a son

2 a young man

3 a man of any age, especially when you are talking about where he comes from

a. (*American, offensive*) an extremely offensive word used for talking to a black man, especially in the past

4 **the boys** (*informal*) a group of men who are friends

(*British*) the members of a sports team

5 used when speaking to a male dog or horse

6 a boy or man of any age who has a particular job

(33) **girl**

1 a female child

a. a daughter

2 a female adult, especially a young one. This use is considered offensive by many women

a. **girls** used for talking to or about a group of women, especially by women who are the same age or older. This is often considered offensive when used by men

b. (*old-fashioned*) a young woman who works as a servant or in a shop, office etc

3 a female animal, especially a pet

PHRASES

my girl (*British spoken*) used by some people when talking to a girl or woman who is younger than they are, especially to show that they are angry. This is usually considered offensive

someone's girl (*old-fashioned*) someone's girlfriend

Words that may cause offence: girl

People sometimes say **girl** to refer to a young adult woman, but this use may cause offence. Avoid using **girl** if it would seem wrong to use **boy** about a young man of the same age. Do not use **girl** about an adult woman.

Though such uses of *girl* to refer to adult women have been problematized and contested – as has *lady*[10] – not all anglophone adult women feel so strongly about the words, and may even prefer to be called a girl, or lady, rather than woman, according to situational context.

It is interesting to compare entries for **boy** and **girl** in these mediated publishers' texts with those in crowd-sourced *Simple Wiktionary*, which are indeed simple and mainly asymmetric. Here, examples are included as reminders of the stereotyping potential with such words:

(34) **boy**

1 (countable) A **boy** is a male child.

He had a pretty wife and two little ones: a boy and a girl.

My oldest son was a Boy Scout in England.

The boys basketball team won five games in a row.

Two teenage boys died in the crash.

A 12-year-old boy stands at the window and watches two men outside.

(35) **girl**

1 A **girl** is a female child.

Many girls like to play with dolls.

I have two children: a boy and a girl.

2 (informal) A female person of any age (even a woman).

The girls are going out tonight, do you want to come?

I really love that girl.

3 (informal) A female animal.

My cat is a girl.

She is a girl cat.

[10] See discussion by Lakoff (1975) and subsequent feminist linguists, who point out that lady trivializes even while apparently showing politeness and respect.

Where items referring to sexual behaviour and sexual preferences are concerned, comparisons between historical and current dictionaries show the extent of social change. For example, in all of the big five British learners' dictionaries, entries for *gay* give priority to the sense "homosexual", labelling as old-fashioned its sense "happy, cheerful"; the derogatory use of *gay* "stupid, absurd, inadequate", if included at all, is labelled as offensive. *Bisexual, heterosexual, homosexual, lesbian, same-sex, transgender* etc. are all routinely covered. Particularly interesting in April 2014 – at the time of writing, less than a month after same-sex marriage was legalized in the UK – is how far online versions of learners' dictionaries reflect this in entries for *husband, wife, marriage, married, marry*. For example, Longman and Oxford have primarily heteronormative, though symmetrical, explanations for husband and wife, while Cambridge (British English version only) has non-specific explanations but heteronormative examples:

(36) **husband**

the man that you are married to:

I've never met Fiona's husband.

(37) **wife**

the woman that you are married to:

I met Greg's wife for the first time.

She's his third wife (= she is the third woman he has been married to).

Macmillan's entries are also symmetrical and non-specific and include non-heteronormative examples (a change from its print edition of 2007, *MED2*):

(38) **husband**

a male partner in a marriage

Carole's husband died last year.

She isn't looking for a husband.

He may be separated from his husband and deported back to Venezuela.

(39) **wife**

a female partner in a marriage

I'd better phone my wife and tell her I'll be late.

a reception for the wives of the ambassadors

In April she became the proud parent of twins with her wife Alex.

The four online learners' dictionaries provide mostly non-specific explanations for *married, marry* and *marriage*, though often imply heteronormativity through choices of examples which indicate mixed-sex couples. However, Macmillan is explicit in extending its explanation (another change from *MED2*);

(40) **marriage**

the relationship between two people who are husband and wife, or a similar relationship between people of the same sex

a long and happy marriage

Too many marriages end in divorce.

by marriage: *I'm related to Bill by marriage* (= he is a relative of my husband or wife).

Compare the cultural information in entries and examples in *Simple Wiktionary*, which are almost entirely heteronormative:

(41) **married**

1. A man and a woman are **married** if they are husband and wife to each other. Usually when two people are **married** they live in the same house and they often have children. Two people have a special day to become **married**.

I don't need to meet more young men – I'm already married.

(42) **marry**

1 When two people **marry** they become husband and wife; that is, they become married. In many countries this is a legal agreement. In some cultures **marrying** is a part of the religion. **Marrying** is often done with a wedding (a special day for those people to marry).

I cannot believe he ***married*** *her when there are nicer girls out there.*

There are many other items which could be used to test how far dictionary texts represent attitudes towards sexuality and acknowledgement of changing paradigms, as realized in lexis and therefore in need of definition – not least *gender* itself, where all of the big four online learners' dictionaries currently offer purely male-female binary explanations of *gender*, though gender is now widely considered a social and cultural construction with non-binary variations. Oxford's explanation mentions the first of these points; of the dictionaries examined, but only *Simple Wiktionary* mentions both:

(43) (Longman)

the fact of being male or female

(44) (Oxford)

the fact of being male or female, especially when considered with reference to social and cultural differences, not differences in

biology

(45) (*Simple Wiktionary*)

1 A living thing's **gender** is its sex: male (man, boy), female (woman, girl), both, or neither.

3 (*psychology*) Someone's gender is whether they behave like a boy or girl. This is called masculine or feminine, and not the same as male or female. [*sic*]

We can expect dictionaries eventually to adapt to new norms more fully; at the same time, we have to acknowledge that some of these new norms are not universally accepted by any means, and may seem abnormal, even abhorrent, to some sectors of the global usership. Thus the changing mores and attitudes of the culture within which dictionaries are written – specifically here the community of native-speakers of British English – may be at odds with the mores and attitudes of the markets and readerships to which the texts are presented: an interesting tension between language, lexicography, and receiver. Is it possible that culture-specific splinter dictionaries, with different ideological perspectives, may develop, for example for/in cultures where homosexuality is illegal or stigmatized, or where women do not have equal rights? Should this matter? and who has the right to say it matters?

5 Representing Age

The last area I want to consider is age and ageism, as represented in dictionaries: the subject of a case study and part of an ongoing research collaboration into discourses of ageing. Particular sites for potential ageism are those adjectives and nouns which reference age directly, such as *young, old, teenager, codger*, though many other items embed or entail notions of age indirectly, including adjectives with age-related semantic preferences: see discussion in Moon 2014.

This written version of my talk looks at just two items to demonstrate the issues in an area that has been, metalexicographically, underexplored. The first is *middle-aged* which, like *young, elderly, old*, is a generalized indicator of life stage. Since age labels carry evaluations – young is good, old is bad – any discussion of when a label begins or ceases to be appropriate is evaluatively loaded and ideologically weighted. Some dictionaries set time parameters for *middle age*, others are more vague; *Cobuild2* offers both strategies:

(46) **Middle age** is the period in your life when you are no longer young but have not yet become old. Middle age is usually considered to take place between the ages of 40 and 60.
Men tend to put on weight in middle age.

These next entries and explanations are from current online dictionaries:

(47) **middle age** (Cambridge)
the period of your life, usually considered to be from about 45 to 60 years old, when you are no longer young, but are not yet old:
Once you reach middle age, you have to be sensible with your health.

(48) **middle age** (Longman)
the period of your life between the ages of about 40 and 60, when you are no longer young but are not yet old:
Men who smoke are more likely to have heart attacks in middle age.

(49) **middle-aged** (Longman)
1 between the ages of about 40 and 60:
a middle-aged businessman.

(50) **middle-aged** (Macmillan)
1 no longer young but not yet old:
He seems prematurely middle-aged

Oxford agrees with Cambridge as to age range, but while Longman agrees with *Cobuild2*, it does not agree with itself, since its online word focus feature for **old**, which appears automatically at **midd-**

le-aged, explains it as “aged between about 50 and 60 years old”. Examples, where included, sometimes have a hortatory flavour, or reference dullness and decline.

Dullness is more directly represented in subsidiary senses for *middle-aged* in learners’ dictionaries. The connotations of *middle-aged*, and its overall negative evaluation, in these senses are reflected in the following small selection of BoE corpus lines:

```
n’roll even, start to sound so middle-aged? <p> We’ll ignore the occasional
    with a son’s girlfriend. A middle-aged Conservative MP, the essence of
 tartly. Exactly: the balding, middle-aged fanclub is here early, packing DAT
      at his side, was a plump middle-aged lady in a brown jumper and navy-blue
half the time!” <p> Meanwhile, middle-aged locals shuffle through the scene,
<p> Wilson was a middle-sized, middle-aged man in a grey, herringbone suit. His
      a conversation between a middle-aged man and his wife about insurance.
       baldheads or of selfish middle-aged people.” Thomas Wentworth Higginson
    climbed out. An overweight middle-aged salesman trying to look slimmer in a
  into an outer office where a middle-aged woman sat at a secretarial desk
```

For example:

(51) (*Cobuild2*)

2 If you describe someone’s activities or interests as **middle-aged**, you are critical of them because you think they are typical of a middle-aged person, for example by being conventional or old-fashioned.

Her novels are middle-aged and boring.

(52) (Cambridge)

(*disapproving*) too careful and not showing the enthusiasm, energy, or style of someone young:

What a conventional, middle-aged attitude he has to life!

(53) (Longman)

someone who seems middle-aged seems rather dull and does not do exciting or dangerous things:

Living with Henry had made her feel middle-aged.

(54) (Macmillan)

2 used for suggesting that someone's behaviour, clothes, etc. are boring and typical of middle-aged people:

They are in their twenties, but have very middle-aged views.

(55) (Oxford)

3 (disapproving) (of a person's attitudes or behaviour) rather boring and old-fashioned.

He has a very middle-aged attitude to life.

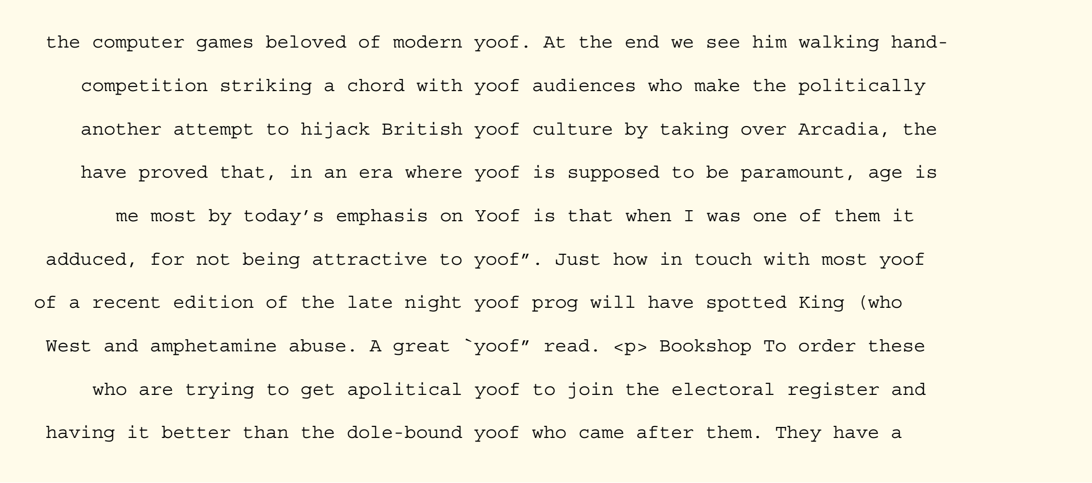
Perhaps Macmillan's example for its first sense seems to be semantically closer to its second sense; Longman's example provides a context of use, but nothing to distinguish *middle-aged* from *depressed, fulfilled, secure, young, happy...*, unless dullness is to be inferred from the name Henry (an unreasonable expectation). Compare too an entry in crowd-sourced *Urban Dictionary*, which in explaining its age reference also rationalizes its connotations:

(56) **middle aged**

i. a period between early adulthood and old age, anywhere from 30 to 65 years old.

ii. Something most people will not admit to being. (It sucks to be older than 29...)

Many other items, used to identify whole age groups or individuals within age groups, are also evaluatively charged and communicate attitude. *Youth* itself is ambivalent: sometimes a focus for nostalgia, the ideal, a life force; sometimes a focus for disapproval. Both evaluation and youth culture are bound up in its respelling *yoof*, my final example here, as in these BoE lines from journalistic media:



```
 the computer games beloved of modern yoof. At the end we see him walking hand-
    competition striking a chord with yoof audiences who make the politically
    another attempt to hijack British yoof culture by taking over Arcadia, the
    have proved that, in an era where yoof is supposed to be paramount, age is
       me most by today's emphasis on Yoof is that when I was one of them it
 adduced, for not being attractive to yoof". Just how in touch with most yoof
of a recent edition of the late night yoof prog will have spotted King (who
 West and amphetamine abuse. A great `yoof" read. <p> Bookshop To order these
     who are trying to get apolitical yoof to join the electoral register and
 having it better than the dole-bound yoof who came after them. They have a
```

The big four online learners' dictionaries all label *yoof* as informal and humorous:[11]

[11] Cf. discussion of youth/youth in Bennett et al. 2005: 380-382; Thorne 2009: 343-5.

(57) (Macmillan)
(*British, very informal, humorous*) young people. This word is used especially on television, in the newspapers etc, as a humorous way of spelling the word 'youth'

(58) (Oxford)
(*British English, informal, humorous*) a non-standard spelling of 'youth', used to refer to young people as a group, especially as the group that particular types of entertainment, magazines, etc. are designed for

But humour is only part of the pragmatics of its usage, which also seems to involve contempt and trivialization – condescension is evident in the corpus lines above. The explanation in *Urban Dictionary* is more expansive and explicit:

(59) **yoof**
2 Cynical description for a style of marketing or programming created by establishment or corporate interests that seeks to identify with the under-21's and thereby sucker them into parting with their cash or individuality with its promise of street credibility or non-conformity. Media vehicles or brands that tell kids what to do by creating an ersatz peer group for them which they then feel they have to conform to.

As with *middle-aged*, this alternative lexicography affords insights into attitudes and connotation that are beyond the mandates and controls of conventional dictionaries: see Smith (2011) for discussion of *Urban Dictionary* and its significance as a lexicographical text.

6 To Conclude

I have looked at only a small selection of entries and, as I warned at the outset, I have offered no solutions. My intention was instead to emphasize the problems that persist, are perhaps insoluble, in the lexicographical treatment of a disparate range of items where ideology and institutionalized attitudes come into play. I have focused almost entirely on monolingual learners' dictionaries, but definitions and analyses in inventory/concise dictionaries, bilinguals, dictionaries for children or school students, are no less problematic, as critical lexicography has repeatedly found. There is massive potential in online dictionaries, including collaborative crowd-sourced dictionaries, for radicalism and inventiveness in entry design and for more effective representations of meaning – including the meanings of ideologically loaded words; but there is also potential for the filtering of world views, re-presentation rather than representation, in ways that may not seem desirable to us, here with our western perspectives and our own filtered views.

In his seminal paper on dictionary definitions, or explanations, Hanks says:

> In the last resort, perhaps, all meanings are displaced, since all meanings rely on constructive intrepretation by the hearer/reader, as well as by the utterer. If this is true, there is no such thing as literal

> meaning, and a dictionary explanation is no more than a compromise with the impossible, a desperate attempt to state the unstateable. (1987: 135)

Yet desperate attempts go on, and meaning remains at the heart of any dictionary, of overwhelming importance. Over thirty years ago, Béjoint drew attention to his survey finding that "87% of the students [advanced learners of English] placed meaning among the three most often sought-after pieces of information" in a dictionary (1981: 215), substantially more than any other information type; at the same time, the highest-ranked cause of failed look-ups (29%) was "unsatisfactory definitions"(1981: 217), almost one in three.[12] Has the situation changed that much? Certainly with respect to ideologically loaded words, the difficulties of producing satisfactory entries are compounded by the complexities and instability of their meanings, the questions of stance and audience, and the balancing of what words mean with what they can be said to mean or be allowed to mean: moreover, the very process of composing entries for such words is essentially an ideological act. I may not have offered solutions, but I hope that at least I have demonstrated something of the nature, and seriousness, of the issues.

7 References

7.1 Dictionaries Cited

(ALD1) Hornby, A.S., Gatenby, E.V., Wakefield, H. (1948). *A Learner's Dictionary of Current English.* Oxford: Oxford University Press. (Previously published as *Idiomatic and Syntactic English Dictionary*, Tokyo: Kaitakusha, 1942).

(Cambridge) *Cambridge Learner's Dictionary Online*. Accessed at: http://dictionary.cambridge.org/ [April 2014].

(Cobuild2) Sinclair, J., Bullon, S. (eds.) *Collins Cobuild English Dictionary* (1995, 2nd edition). London and Glasgow: HarperCollins.

(Longman) *Longman Dictionary of Contemporary English Online*. Accessed at: http://www.ldoceonline.com/ [April 2014].

(Macmillan) *Macmillan English Dictionary Online*. Accessed at: http://www.macmillandictionary.com/ [April 2014].

(MED2) Rundell, M. (ed.) (2007, 2nd edition). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

(OED) *Oxford English Dictionary* Accessed at: http://www.oed.com [April 2014].

(Oxford) *Oxford Advanced Learner's Dictionary Online*. Accessed at: http://www.oxfordlearnersdictionaries.com/ [April 2014].

(Simple Wiktionary) *Simple English Wiktionary*. Accessed at: http://simple.wiktionary.org/wiki/Main_Page [April-May 2014].

(Urbandictionary) *urban dictionary* Accessed at: http://www.urbandictionary.com/ [April 2014].

[12] Other later surveys have still higher figures.

7.2 Other Literature

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

Barnickel, K.-D. (1999). Political correctness in learners' dictionaries. In Th. Herbst, K. Popp (eds.) *The Perfect Learners' Dictionary?*. Tübingen: Niemeyer, pp. 161-174.

Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: a study of language needs and reference skills. In *Applied Linguistics*, 2(3), pp. 207-222.

Béjoint, H. (2010). *The Lexicography of English*. Oxford: Oxford University Press.

Bennett, T., Grossberg, L., Morris, M. (2005). *New Keywords: A Revised Vocabulary of Culture and Society*. Oxford: Blackwell.

Benson, P. (2001). *Ethnocentrism and the English Dictionary.* London: Routledge.

Caldas-Coulthard, C.R., Moon, R. (2010). Curvy, hunky, kinky: using corpora as tools in critical analysis. In *Discourse and Society*, 21(2), pp. 1-35.

Connor Martin, K. (2005). Gendered aspects of lexicographic labeling. In *Dictionaries*, 26, pp. 160-173.

Fuertes Olivera, P.A. (2009). The function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective free multiple-language internet dictionary. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) *Lexicography at a Crossroads: Dictionaries and Encylopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp. 99-134.

Gabrielatos, C., Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. In *Journal of English Linguistics*, 36(1), pp. 5-38.

Graham, A. (1975). The making of a nonsexist dictionary. In B. Thorne, N. Henley (eds.) *Language and Sex: Difference and Dominance*. Rowley, Massachusetts: Newbury House, pp. 57-63.

Hanks, P. (1987). Definitions and explanations. In J. Sinclair (ed.) *Looking Up*. London and Glasgow: Collins, pp. 116-136.

Hanks, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge, Mass.: MIT Press.

Holmes, J., Sigley, R. (2001). What's a word like *girl* doing in a place like this?. In P. Peters, P. Collins, A. Smith (eds.) *New Frontiers of Corpus Linguistics*. Amsterdam: Rodopi, pp. 247-263.

Hornscheidt, A.L. (2011). Postcolonial continuities in Danish monolingual dictionaries: Towards a critical postcolonial linguistics. In E.A Anchimbe, S.A. Mforteh (eds.) *Postcolonial Linguistic Voices: Identity Choices and Representations.* Berlin: De Gruyter Mouton, pp. 265-298.

Kachru B.B., Kahane, H. (eds.) (1995). *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*. Tübingen: Niemeyer.

Krishnamurthy, R. (1996). Ethnic, racial and tribal: the language of racism?. In C. Caldas-Coulthard, M. Coulthard (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge, pp. 129-149.

Lakoff, Robin (1975). *Language and Woman's Place*. New York: Harper Colophon Books.

Landau, S.I. (1985). The expression of changing social values in dictionaries. In *Dictionaries*, 5, pp. 261-269.

Landau, S.I. (2001, 2nd edition). *Dictionaries: The Art and Craft of Lexicography*.

Leech, G. (1981, 2nd edition). *Semantics*. Harmondsworth: Penguin.

Louw, B. (1993). Irony in the text or insincerity in the writer? - the diagnostic potential of semantic prosodies. In M. Baker, G. Francis, E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair.* Amsterdam: John Benjamins, pp. 157-176.

Moon, R. (1989). Objective or objectionable? Ideological aspects of dictionaries. In M. Knowles, K. Malmkjær (eds.) *Language and Ideology* (*ELR Journal*, 3). Birmingham: English Language Research, University of Birmingham, pp. 59-94.

Moon, R. (2014). From gorgeous to grumpy: adjectives, age and gender. In *Gender and Language*, 8(1), pp. 5-41.

Ogilvie, S. (2013). *Words of the World*. Cambridge: Cambridge University Press.

Russell, L.R. (2012). This is what a dictionary looks like: the lexicographical contribution of feminist dictionaries. In *International Journal of Lexicography*, 25(1), pp. 1-29.

Semino, E. (2008). *Metaphor in Discourse*. Cambridge: Cambridge University Press.

Sigley, R., Holmes, J. (2002). Looking at *girls* in corpora of English. In *Journal of English Linguistics*, 30(2), pp. 138-157.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. (2004). *Trust the Text*. London: Routledge.

Smith, R.E. (2011). Urban dictionary: youth slanguage and the redefining of definition. In *English Today*, 27(4), pp. 43-48.

Svensén, B. (1993) *Practical Lexicography: Principles and Methods of Dictionary-making*. Oxford: Oxford University Press.

Thorne, T. (2009). *The 100 Words that Made the English.* London: Abacus.

Whitcut, J. (1984) Sexism in dictionaries. In. R.R.K. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 141-144.

Williams, R. (1988, expanded edition). *Keywords: A Vocabulary of Culture and Society*. London: Fontana.

The Dictionary-Making Process

The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary

Tressy Arts
Oxford Arabic Dictionary
tressy.arts@gmail.com

Abstract

In this presentation, we will illustrate the process of making our brand-new Arabic-English/English-Arabic dictionary, which is due for publication in print and online in August 2014. It is intended for speakers of both English and Arabic. It contains over 26,000 entries on each side, including many up-to-the-minute words and expressions. Collocations and examples are an important feature. The dictionary has been compiled using dictionary writing software that enables editors to work and communicate with one another regardless of their location. It will be available in both print and online.

We show the entire process of making an Arabic dictionary, from finding a reliable framework in both languages, to developing a unique online functionality. We show the difficulties lexicographers face when compiling an Arabic dictionary, and the ways in which we dealt with those. In addition, the Oxford Arabic Dictionary has quite a few features that are entirely new to Arabic dictionaries, and we illustrate how we went about developing those.

Keywords: Arabic; bilingual; dictionary-making process

1 Arabic Lexicography

Arabic lexicography has a centuries-old tradition, starting with the *Kitāb al-'Ayn*, a large monolingual dictionary, in the eighth century. This impressive legacy is both a blessing and a curse, since conservatism rules, and the mediaeval tomes like the 13th century *Lisān al-'Arab* still count as *the* standard in lexicography. Modern monolingual dictionaries often do little more than repeat what has been said before, regardless of whether the senses, examples, or even the headwords mentioned are still in actual use. Many Arabic lexicographers, linguists and laymen see Classical Arabic, the language of the Koran and the pre-Islamic poetry, as the perfect standard, and any deviations from that are seen as corrupting the language. Therefore many loanwords and words derived from colloquial Arabic are not included in modern dictionaries, despite them being commonplace in Standard Arabic texts. Haywood says in his "Arabic Lexicography": "The lexicographers helped to keep the written language static, and to aid the understanding of it, as the spoken dialects diverged more and more from it."

(Haywood 1960: 116). There are several Arabic Language Academies which aim to find proper Arabic words (that is, words which confirm to the Arabic root-and-pattern system for word forming, rather than words which are simply English words written in Arabic letters) for new concepts, but they are slow-moving and don't always communicate clearly, so that loanwords for new concepts often are commonplace long before an approved version is released. Sometimes the two continue to exist side by side (*tilifūn* and *hātif* (telephone)), sometimes the loanword remains strong (*tilifizyūn* (television)), sometimes the approved version gains prominence (*ḥāsūb* (computer) is more common than *kumbyūtir* these days).

Still, it is rare to find modern words in monolingual dictionaries, either loanwords, or ones formed according to Arabic morphology. One will be hard put to find *ḥāsūb* (computer) in most, and a common modern word like *nazzala* (to download) I haven't been able to find in any, not even online.

Bilingual lexicography with Arabic as source or target language also has a long history. Already in the 11th century al-Zamak̠šarī published an Arabic-Persian dictionary, and al-Kāšġarī a Turkish-Arabic one. For Europeans, the first bilingual dictionary was Golius' *Lexicon Arabico-Latinum*, published in Leiden in 1653. The first Arabic-English one was produced by Edward William Lane in the late 19th century.

Looking at current-day Arabic-English/English-Arabic lexicography, however, one cannot help but be a bit disappointed with what's there, especially considering the status of both languages as major world languages. The standard Arabic-English work, Hans Wehr's *Dictionary of Modern Written Arabic*, was translated from German and not much updated since it was first published in 1961. Though groundbreaking at the time, it is nowhere near comparable to modern bilingual dictionaries for most other languages. Obviously the word list is outdated, but also the presentation of the entries is not really what one would expect from a modern dictionary: long lists of possible English translations are given for most Arabic words, without any guidance for the user on which translation to choose in which context; no word senses are distinguished. No examples are given and only very few collocations, many of which don't actually exist in modern Arabic. Yet, this is still the dictionary that everyone translating from Arabic into English will use.

The situation for English-Arabic is even direr. Arguably the best option is still Oxford's English-Arabic (hand-written!) dictionary from 1972 (cf. Benzehra 2012); the most popular work is *al-Mawrid* (Baalbaki & Baalbaki 2013), which is updated regularly, but has certain major flaws. The English used seems mostly derived from very old texts, and its policy for including new words in updated editions is not clear: the most recent edition, from 2013, has "fuscous" and "silvern", featured new entries "naughties" [sic] and "smirt", but not "blog" or "text message". Examples given seem to be taken from very old English dictionaries/texts ("swallows that affect chimneys", "the hes would quarrel and fight with the females"). As to the Arabic translations, many of the senses given cannot be attested to exist in modern English, senses are ordered from oldest to most modern, leading to the most common sense often being the last one, and the editors even create new Arabic "words" to serve as translations:

> compilers of bilingual dictionaries are not only entitled to coin words which may or may not gain currency, but that coinage becomes an essential duty of theirs, especially with a language like Arabic, where a huge number of terms, particularly scientific ones, are lacking. (Baalbaki 2004: 68)

This statement is disputable: Arabic does not lack terms for scientific or any other areas, but some of the terms will be loanwords or multi-word constructions, so it depends on what is understood by "Arabic terms". We believe that it's more useful for an English-Arabic dictionary to give those loanwords or multi-word phrases, than to coin new words which are useful for neither decoding or encoding users.

Other European languages fare slightly better, but not much, with the exception of Dutch, which has some good and modern dictionaries, published at the beginning of this century (*Woordenboek Nederlands-Arabisch/Arabisch-Nederlands* (Hoogland et al. 2003) and *Leerwoordenboek Arabisch* (Van Mol & Bergman 2001)).

There are plenty of online Arabic dictionaries, but the accuracy there usually leaves much to be desired.

So in developing a large bilingual Arabic-English/English-Arabic dictionary that was corpus-based and could live up to modern expectations, our team was truly taking on a unique endeavour. Not being able to rely on any previous works, either bilingual or monolingual, all progress had to be made by painstaking research.

2 The Arabic Language

The lack of reliable predecessors is far from the only complicating factor in Arabic lexicography. To highlight a few more:

2.1 Diglossia

In the Arabic world a diglossia exists, with the "high" variant, Modern Standard Arabic, being the accepted language for any written and official spoken discourse. Meanwhile, a multitude of "dialects" or "colloquials" are the languages people actually speak. These are officially known as dialects, but are often mutually unintelligible, and can be considered different languages on purely linguistic grounds. It is hard to say how many exist, as there is a dialect continuum, but in many countries the dialect of the capital has the major status (so if you pick up a text book or travel guide in "Egyptian Arabic", this will be the dialect of Cairo, etc.). The "dialects" have no agreed orthography, though they are used more and more in written communication, mostly on the internet, and even a few novels have appeared.

Since we wanted to make a broadly useful dictionary, not focusing on one or a few dialects, and since it is hard to write dialects because of the lack of standard orthography, we chose to use (almost) exclusively Standard Arabic for the Arabic in the dictionary. Even for English expressions which are very informal or chatty, we have tried to give Standard Arabic equivalents, or if needed descriptions. Both languages have level markers so the user is alerted if there is a difference in level.

Then we had to devise a strategy for deciding which words are indeed Modern Standard Arabic, and should be included in the dictionary. We could not simply include all written words in general texts, since these days many dialect words are written. We could also not rely on dictionaries to verify existence of words, since many common modern words are not listed in monolingual or bilingual dictionaries. So the criterion we decided on was that words which are used without quotation marks in a reasonable number of non-specialist, otherwise Standard Arabic texts, could themselves be considered Standard Arabic, and therefore deserved a place in the dictionary. On the other hand, if a word is only found in dictionaries, we did not include it.

2.2 No Native Speakers

It is standard practice these days in writing bilingual dictionaries to have them compiled by native speakers of the target language. However, there are no native speakers of Modern Standard Arabic; children in Arabic countries grow up with one of the dialects or even a non-Arabic language (Berber, Kurdish) as their mother tongue. Standard Arabic is learnt, like a foreign language, at school and in the mosque. Research has shown that indeed Arabic people process Standard Arabic in the brain in the same way as foreign languages (Ibrahim & Aharon-Peretz 2005). So although the editors hired were highly educated linguists, they lacked that native speaker sense that lexicographers for most other languages have.

2.3 Geographical Spread

In addition, the Arab world has a very large geographical spread, and regional language preferences exist even in Standard Arabic, so editors need to constantly be wary of using words and phrases that are limited to their home country. Very little research has been done in this area, partly because of the prevailing ideal of Arabic being one uniform language.

If we take all these factors together (no reliable antecedents, no native speaker sense, uncharted geographical differences), we can understand that it's very hard for the "native speaker"[1] editors to feel secure in their translation of a word or phrase, necessitating elaborate checking in a corpus and on the internet, as well as discussion with native speakers from other locations, to find the right terms. This meant that the project took much longer, and was more expensive, than originally envisaged.

[1] We will use this to refer to native speakers of one of the Arabic dialects with good knowledge of Standard Arabic, for want of a better term.

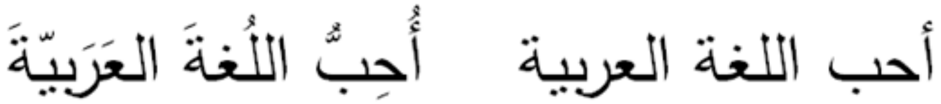
3 The Bases

For the Arabic-English side, the data of Hoogland's Arabic-Dutch *Woordenboek* mentioned above, which had got very good press, was licensed.

For the English-Arabic side, we used an English framework that had been developed for use as a basis for Oxford unabridged bilingual dictionaries, expanded with words that are especially relevant for Arabic, like the English names of the Islamic months.

4 Vocalization

The Arabic script is a consonant script. Vowels are indicated by diacritic signs over and under the letter (see figure 1), but are not commonly written, so a word like *al-maḡrib* (Morocco) will be written as *al-mḡrb*. In addition, the three cases Arabic has are most often only indicated by vowels, and hence not visible in most texts.



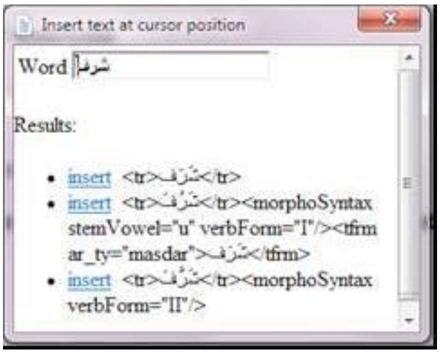
أحب اللغة العربية   أُحِبُّ اللُّغَةَ العَرَبِيَّةَ

Figure 1: Vocalized (left) and unvocalized (right) Arabic.

For learners of Arabic it is useful to be able to find all vowels both in the Arabic-English side (if one doesn't know a word, one probably doesn't know how to pronounce it), and in the English-Arabic side (when one finds a new Arabic word, it's useful to find the vowels). Similarly, case endings are useful to understand the syntax and word combinations. So in order to be most useful for non-Arab users, as well as clearest for Arab users, we wanted all Arabic on both sides of the dictionary, headwords, translations, examples and descriptions, to have full vocalization, including the case endings. This placed an extra burden on our editors, since most people are not used to writing vocalized Arabic, and there is no standard vocalization system, so we had to devise our own rules (most were taken over from the system used in the Hoogland dictionaries). It also made the use of an adapted font necessary, since most Arabic fonts are not designed with the vowels in mind, and they can "disappear" in certain letter combinations.

For the English-Arabic side, not only did we want the translations to be written in the above-mentioned vocalization system, we also wanted to provide the unpredictable grammatical information for single-word Arabic translations (the plurals of nouns and adjectives, and conjugational information and infinitives of many verbs are unpredictable in Arabic, so it's useful when this is provided with a word functioning as a translation). This information was already present for all the headwords in the Arabic framework, so a "translation picker" tool was developed within Oxford University Press: the

English-Arabic translator could enter an unvowelled word in this tool in the dictionary database, and was presented with all possible vocalized headwords in the Arabic data, with their grammatical information (see figure 2).



Word شرف

Results:

- insert <tr>شَرَف</tr>
- insert <tr>شَرُفَ</tr><morphoSyntax stemVowel="u" verbForm="I"/><tfrm ar_ty="masdar">شَرَف</tfrm>
- insert <tr>شَرَّفَ</tr><morphoSyntax verbForm="II"/>

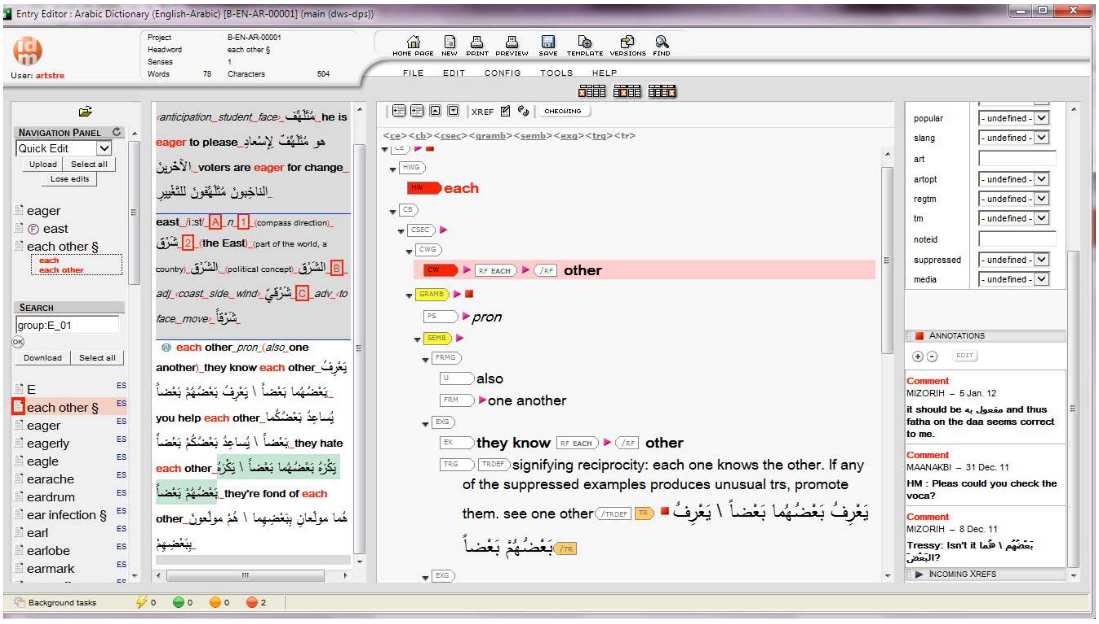
Figure 2: The Translation Picker in use.

By simply selecting the one they wanted, a correctly vocalized word with all relevant grammatical information was entered as translation.

This had an added advantage: if the translator could not find the word they wanted as translation in the Translation Picker, it meant that that word was not in the Arabic lemma list for the Arabic-English dictionary. The translator then made a note for the chief editor, who would decide if the Arabic word in question was a valuable addition to the Arabic-nglish lemma list, and add it to the latter if it was.

5 The Database Used

The dictionary writing software used was DPS (Digital Publishing System) produced by French company IDM, a database system specialized in creating and developing dictionaries, which is used by Oxford University Press and many other publishers. Its application, the Entry Editor, lets users connect to the central database in Oxford and download and upload entries to work on. It lets users edit entries uploaded by others, keeps track of who made which changes, lets you revert to any earlier uploaded version, and allows users to communicate with one another through a system of searchable annotations that can be made on every level (entry, sense, translation, etc.), a very valuable tool for a dictionary where much discussion of terms needs to take place. Via its search engine, editors could make highly specific searches, for example every translation that has a “region: Egypt” attached, and for the project managers its workflow manager allowed us to distribute the workload and keep track of progress.


Entry Editor : Arabic Dictionary (English-Arabic) [B-EN-AR-00001] (main (dws-dps))

User: artstre

Project		B-EN-AR-00001	
Headword		each other §	
Senses		1	
Words	78	Characters	504

Navigation Panel: Quick Edit

eager
east
each other §
each
each other

Search: group:E_01

E ES
each other § ES
eager ES
eagerly ES
eagle ES
earache ES
eardrum ES
ear infection § ES
earl ES
earlobe ES
earmark ES

‹anticipation_student_face› مُتَلَهِّف_he is eager to please_هو مُتَلَهِّفٌ لِإسْعادِ الآخَرِينَ|_voters are eager for change_ النَاخِبونَ مُتَلَهِّفونَ للتَّغْيِيرِ_

east_/i:st/_A_n_1_(compass direction)_ شَرْق_2_(the East)_(part of the world, a country)_الشَّرْق_(political concept)_الشَّرْق_B_ adj_‹coast_side_wind›_شَرْقِيّ_C_adv_‹to face_move›_شَرْقاً_

each other_pron_(also_one another)_they know each other_يَعْرِفُ بَعْضُهُما بَعْضاً \ يَعْرِفُ بَعْضُهُمْ بَعْضاً_ you help each other_يُساعِدُ بَعْضُكُما بَعْضاً \ يُساعِدُ بَعْضُكُمْ بَعْضاً_they hate each other_يَكْرَهُ بَعْضُهُما بَعْضاً \ يَكْرَهُ بَعْضُهُمْ بَعْضاً_they're fond of each other_هُما مولَعانِ بِبَعْضِهِما \ هُمْ مولَعونَ بِبَعْضِهِمْ_

<ce><cb><csec><gramb><semb><exg><trg><tr>

HWG
HW each
CB
CSEC
CWG
CW RF EACH /RF other
GRAMB
PS pron
SEMB
FRMG
U also
FRM one another
EXG
EX they know RF EACH /RF other
TRG TRDEF signifying reciprocity: each one knows the other. If any of the suppressed examples produces unusual trs, promote them. see one other /TRDEF TR يَعْرِفُ بَعْضُهُما بَعْضاً \ يَعْرِفُ بَعْضُهُمْ بَعْضاً /TR
EXG

popular	- undefined -
slang	- undefined -
art	
artopt	- undefined -
regtm	- undefined -
tm	- undefined -
noteid	
suppressed	- undefined -
media	- undefined -

Annotations

Comment
MIZORIH – 5 Jan. 12
it should be مفعول به and thus fatha on the daa seems correct to me.

Comment
MAANAKBI – 31 Dec. 11
HM : Pleas could you check the voca?

Comment
MIZORIH – 8 Dec. 11
Tressy: Isn't it هُما \ بَعْضُهُمْ البَعْضَ?

Incoming Xrefs

Background tasks 0 0 0 2

Figure 3: The Entry Editor showing entries in a group, preview, tag tree, and attributes and annotations.

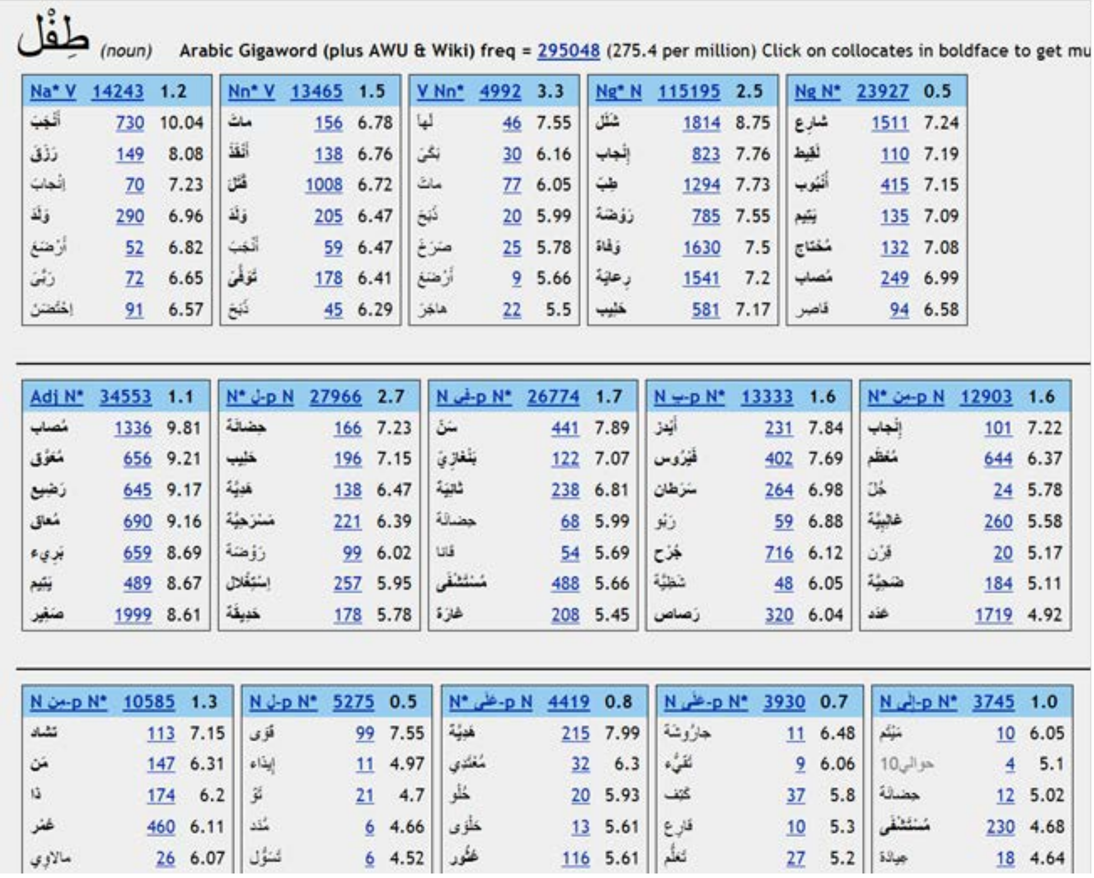
6 The Corpus Used

The Oxford English Corpus contains 2.5 billion words of modern English from all over the globe, and as such was a very valuable resource for our editors. A unique new resource, however, was the Oxford Arabic Corpus, made searchable with the Sketch Engine software, developed by Lexical Computing Limited of Brighton. This enabled us to do truly unprecedented work.

The Oxford Arabic Corpus comprises the Arabic Gigaword Corpus Fourth Edition from Linguistic Data Consortium: 840 million words of news text from nine publications covering the period 1996-2008, plus 10 million words of fiction from the Arabic Writers Union of Damascus, and 30 million words from Arabic Wikipedia.

After the corpus was assembled, the raw data was processed using MADA (Habash, Rambow & Roth 2009). MADA uses the Buckwalter morphological analyzer (by Linguistic Data Consortium) to provide alternative analyses (part-of-speech, vowelled form, and lemma) of each input token, after which a Support Vector Machine classifier ranks the competing analyses in context. The highest ranked analysis is loaded into the Sketch Engine corpus software (Kilgariff et al. 2004), allowing items to be searched and concordanced by word form or lemma, and collocate phrases to be displayed by part of speech structure and grammatical class, as illustrated in Figure 4.

Here we see the Word Sketch for the word *ṭifl* (child). The abbreviation with the asterisk is the search term, and the sequence is from right to left, so the top left column is verbs (V) followed by the search term (N for noun) in the accusative (a). The top collocate here, with 730 results, is *anjaba* (to give birth to). The results give us collocates in the form of verbs, nouns in so-called genitive constructs, adjectives, and prepositions with other nouns.



طِفْل (noun) Arabic Gigaword (plus AWU & Wiki) freq = 295048 (275.4 per million) Click on collocates in boldface to get mu

Na* V	14243	1.2
أنجب	730	10.04
رزق	149	8.08
إنجاب	70	7.23
ولد	290	6.96
أرضع	52	6.82
ربى	72	6.65
إحتضن	91	6.57

Nn* V	13465	1.5
مات	156	6.78
أنقذ	138	6.76
قتل	1008	6.72
ولد	205	6.47
أنجب	59	6.47
توفى	178	6.41
ذبح	45	6.29

V Nn*	4992	3.3
لها	46	7.55
بكى	30	6.16
مات	77	6.05
ذبح	20	5.99
صرخ	25	5.78
أرضع	9	5.66
هاجر	22	5.5

Ng* N	115195	2.5
شلل	1814	8.75
إنجاب	823	7.76
طب	1294	7.73
روضة	785	7.55
وفاة	1630	7.5
رعاية	1541	7.2
حليب	581	7.17

Ng N*	23927	0.5
شارع	1511	7.24
لقيط	110	7.19
أنبوب	415	7.15
يتيم	135	7.09
مختاج	132	7.08
مصاب	249	6.99
قاصر	94	6.58

Adj N*	34553	1.1
مصاب	1336	9.81
معوق	656	9.21
رضيع	645	9.17
معاق	690	9.16
بريء	659	8.69
يتيم	489	8.67
صغير	1999	8.61

N* ل-p N	27966	2.7
حضانة	166	7.23
حليب	196	7.15
هدية	138	6.47
مسرحية	221	6.39
روضة	99	6.02
إستغلال	257	5.95
حديقة	178	5.78

N في-p N*	26774	1.7
سن	441	7.89
بنغازي	122	7.07
ثانية	238	6.81
حضانة	68	5.99
قانا	54	5.69
مستشفى	488	5.66
غارة	208	5.45

N ب-p N*	13333	1.6
أيدز	231	7.84
فيروس	402	7.69
سرطان	264	6.98
ربو	59	6.88
جرح	716	6.12
شظية	48	6.05
رصاص	320	6.04

N* من-p N	12903	1.6
إنجاب	101	7.22
معظم	644	6.37
حل	24	5.78
غالبية	260	5.58
قرن	20	5.17
ضحية	184	5.11
عدد	1719	4.92

N من-p N*	10585	1.3
تشاد	113	7.15
من	147	6.31
ذا	174	6.2
عمر	460	6.11
مالاوي	26	6.07

N ل-p N*	5275	0.5
قوى	99	7.55
إيذاء	11	4.97
تو	21	4.7
مدد	6	4.66
تسول	6	4.52

N* على-p N	4419	0.8
هدية	215	7.99
معتدي	32	6.3
حلو	20	5.93
حلوى	13	5.61
عثور	116	5.61

N على-p N*	3930	0.7
جاروشة	11	6.48
تقيء	9	6.06
كتف	37	5.8
قارع	10	5.3
تعلم	27	5.2

N إلى-p N*	3745	1.0
منيم	10	6.05
حوالي10	4	5.1
حضانة	12	5.02
مستشفى	230	4.68
عيادة	18	4.64

Figure 4: Word Sketch for ṭifl (child).

This allowed us to fine-tune translations, add relevant collocations and examples, and discover new words and senses.

Using the corpus wasn't without its pitfalls though. Correctly tagging Arabic is much more complex than English or Italian. Due to the complexity of Arabic's morphology and the surface-form ambiguity, mistagging in the corpus is inevitable – *wīlz* (Wales) is consistently analysed as a form of *lazza* (to be compressed), because *wīlz* is not in the lemma list and *wa-yaluzzu*, a conjugated form of *lazza*, has the same surface form: WiYLZ and WaYaLuZzu (the capitals are the letters visible in the surface form). Also, because the case endings aren't usually visible, and Arabic syntax has a VSO structure, differentiation between verb + nominative noun and verb + accusative noun is nearly impossible. In Figure 4, the biggest result in the verb + nominative noun table is *qatala* (to kill). We can be reasonably certain that in most of those cases the child is the unfortunate object, rather than the subject, and so is in fact an accusative noun.

Editors therefore needed a sharp critical eye when analysing the results.

7 Finding New Words and Examples in the Corpus

The English framework was developed by a skilled team of Oxford lexicographers, and continued to be expanded, so we could reliably assume that all relevant words and phrases were included. However, for the Arabic, this was a different story: the Arabic-Dutch dictionary was developed in the nineties, when Arabic corpus linguistics was still very primitive. The editors could only search the corpus on surface forms, which is remarkably difficult in a language where a simple verb like "to eat" can have no fewer than 46 surface forms, most of which also can mean "to feed". They found ways to work around those problems though (Hoogland 2004), and the chief editor calculated that the dictionary covered 99.95% of (non-lemmatized) word forms in Modern Arabic texts (Hoogland 2003). However, with the new technology at our disposal in the form of the Oxford Arabic Corpus and Google, we could certainly see many possibilities for improving and expanding the Arabic framework.

We started on the level of the vocabulary, comparing the lemma list of the corpus to the Arabic-Dutch lemma list, and added any lemmas we thought warranted inclusion, like *rawwasa* (to sharpen; to supply with a header). However, this only discovered words that were already listed in the Buckwalter lexicon. Because of Arabic's complex morphology, it's often not possible to automatically distil a lemma from a string if the lemma is unknown, so many surface forms were still unidentified. Mohammed Attia, one of our translators and a computational linguist, had developed a system to distil potential lemmas from the corpus by checking frequency and compliance with Arabic morphological patterns (Attia et al. 2014). These potential lemmas were again compared to our lemma list, and thus we managed to find genuine new words, many of which not only weren't listed in the Arabic-Dutch data, but had not been previously listed in any Arabic dictionaries, like *tamāhā* (to be congruent) and *ṭawā'ifī* (sectarian).

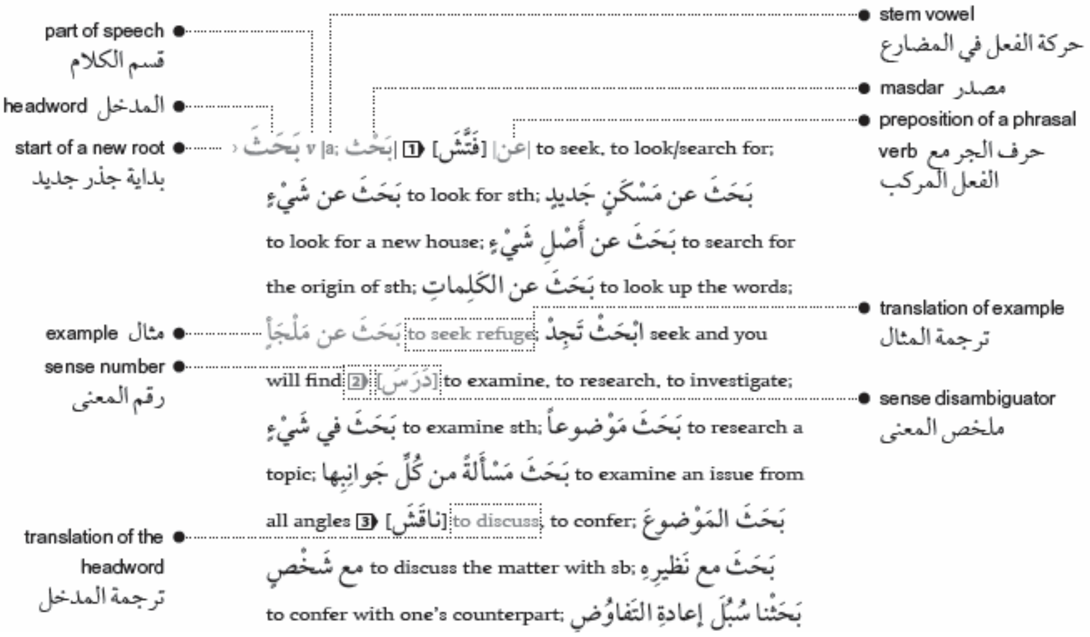
So altogether we had four ways to expand the Arabic lemma list to contain more, currently relevant, entries: listing Arabic translations for English words that weren't in the Arabic data, comparing the lemma list of the Gigaword corpus with the Hoogland lemma list, using Attia's potential lemma extractor, and old-fashioned handwork: critical reading of web sites and being alert to any newly developing words and collocations, like *al-rabī' al-'arabī* (Arab spring) and *taġrīda* (tweet). In these ways we managed to expand the lemma list by over 2,000 entries. At the same time, we pruned entries that were deemed obsolete or archaic, thus removing about four hundred entries from the original Arabic data. This latter was not a priority however, since, from a user's point of view, obsolete entries are not in the way: if a user doesn't encounter the word, they are not going to look it up; so we focused on improving the relevant entries rather than on pruning the less relevant ones. All in all we raised the number of Arabic entries from 24,682 to 26,316.

On the level of individual entries, the editors used the corpus, as well as Google, to check the existing examples for relevance, to find new or better examples and collocations, and on occasion find entirely new senses for existing words (Arts & McNeil 2013). Especially on this microstructural level, the Arabic-English dictionary has been greatly expanded compared to its Arabic-Dutch predecessor.

8 Microstructure and Translations

On both sides of the dictionary, entries are divided into one or more senses, which have disambiguators in the source language indicating the meanings. Many senses have examples to illustrate typical uses of the headword in that sense. Idioms are given separately, outside the scope of the senses. Several types of translations are given:

- the direct translation, which is an equivalent of the headword in that sense and can be used as its translation in most contexts, or which is the equivalent of the example given, e.g. "computer" and *ḥāsūb.*
- the approximate translation, which is a nearly equivalent translation, or a translation in certain contexts (which is then specified), e.g. *Allāhu yuḵallīka*, literally "God bless you", which is used to thank someone, gets the approximate translation "thank you". Approximate translations are indicated with an approximation sign (≈).
- the translation or approximate translation followed by an explanation in brackets. Often the translation is useful for the encoding user, and the explanation for the decoding user, with the explanation further specifying the translation, e.g. "muezzin (*mosque official who recites the call to prayer*)". The explanation is in brackets in the English-Arabic and in brackets and italics in the Arabic-English.
- the definition: where a headword or example doesn't have an equivalent in the target language (for English e.g. "haggis", for Arabic for example many Islamic terms), a description of what it means is given in the target language. In the English-Arabic these definitions are in square brackets, in the Arabic-English in italics.



› بَحَثَ v |a; بَحْث| 1 [فَتَّشَ] |عن| to seek, to look/search for;
بَحَثَ عن شَيْءٍ to look for sth; بَحَثَ عن مَسْكَنٍ جَدِيدٍ
to look for a new house; بَحَثَ عن أَصْلِ شَيْءٍ to search for
the origin of sth; بَحَثَ عن الكَلِماتِ to look up the words;
بَحَث عن مَلْجَأٍ to seek refuge; اِبْحَثْ تَجِدْ seek and you
will find 2 [دَرَسَ] to examine, to research, to investigate;
بَحَثَ في شَيْءٍ to examine sth; بَحَثَ مَوْضوعاً to research a
topic; بَحَثَ مَسْأَلَةً من كُلِّ جَوانِبِها to examine an issue from
all angles 3 [ناقَشَ] to discuss, to confer; بَحَثَ المَوْضوعَ
مع شَخْصٍ to discuss the matter with sb; بَحَثَ مع نَظيرِهِ
to confer with one's counterpart; بَحَثْنا سُبُلَ إعادةِ التَفاوُضِ

Figure 5: An Arabic-English entry with some aspects of its structure illustrated.

8.1 Arabic-English Translations

The existing Hoogland Arabic-Dutch dictionary was translated by Dutch Arabists with a good command of English. They made sure to keep the translation relation between the Arabic and the English in mind, rather than translating from Dutch to English, but the existing Dutch translations were too valuable a tool to disregard. Also, most of the editors had previously worked on the Arabic-Dutch dictionary, so they were familiar with Arabic lexicography and with this specific dictionary.

Then all entries were reviewed by native English-speaking Arabists, correcting the English where necessary and further improving the entries.

Since the original was made in a time when corpus research was hardly possible, the existing entries and examples were checked in the corpus and on Google: are the senses and examples representative of current Arabic, or are they so-called "dictionary words/examples", copied from monolingual dictionaries that are not in actual use any more? The latter were weeded out and where needed we replaced the examples by new ones distilled from modern language found in the corpus and Google. We checked if all senses were attested, and checked senses we couldn't find evidence for with native speakers. Sometimes the corpus showed senses that were not yet listed, and we added those. During the entire process, we could ask native speakers' or other editors' opinion via annotations.

See figure 5 for an example of an Arabic entry and its structure.

8.2 Arabic-English Translations

The English-Arabic side of the dictionary was formed by having the English framework translated into Arabic by translators and lexicographers from several Arabic countries: Algeria, Tunisia, Egypt, Palestine, Lebanon, and Iraq. For certain fields, namely Medicine, Technology, and Law and Business, bilingual experts were found to advise on the terminology. The English framework contains field markers for many entries, so exporting all entries in a specific field was easily done.

Since, as I stated above, no one is a true native speaker of Arabic, it was important for the translators and reviewers to be able to check the translations not just against their own language sense, but also against a corpus. Even a 900-million-word corpus is not quite reliable enough to state that a certain word sense/construction is never used, so for verification of usage of Arabic translations of English terms, Google proved invaluable. Using quotation marks makes it possible to search for exact constructions, enabling translators and reviewers to verify that the translations are actually in use in modern Arabic.

Annotations were used to communicate with other native speakers verifying translations or asking for suggestions ("in Iraq we say this, but do you also use this in North Africa?").

All entries were, after translation, reviewed one or more times by revising editors, at least one of which, for every entry, was a native speaker. Often the reviewer would discuss with the original translator to find the best solution for tricky aspects.

See figure 6 for an example of an English-Arabic entry and its structure.

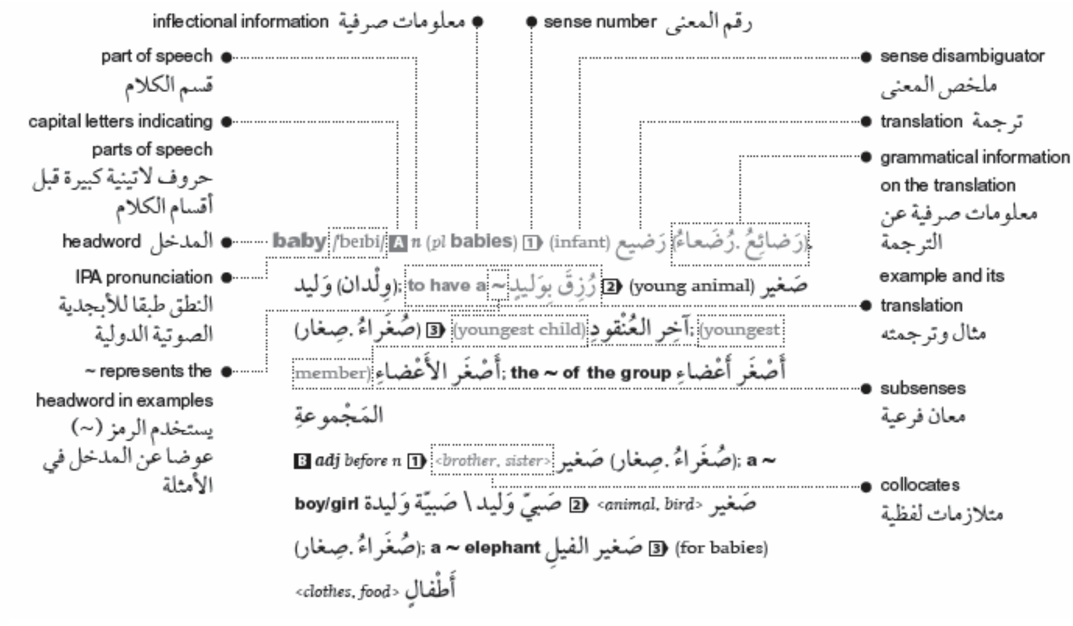

inflectional information معلومات صرفية
sense number رقم المعنى
part of speech قسم الكلام
sense disambiguator ملخص المعنى
capital letters indicating parts of speech حروف لاتينية كبيرة قبل أقسام الكلام
translation ترجمة
grammatical information on the translation معلومات صرفية عن الترجمة
headword المدخل
IPA pronunciation النطق طبقا للأبجدية الصوتية الدولية
example and its translation مثال وترجمته
~ represents the headword in examples يستخدم الرمز (~) عوضا عن المدخل في الأمثلة
subsenses معان فرعية
collocates متلازمات لفظية
baby /ˈbeɪbi/ A n (pl babies) 1 (infant) رَضيع (رَضائِعُ. رُضَعاءُ);
وَليد (وِلْدان); to have a ~ رُزِقَ بِوَليدٍ 2 (young animal) صَغير
(صُغَراءُ. صِغار) 3 (youngest child) آخِرِ العُنْقودِ; (youngest
member) أَصْغَرُ الأَعْضاءِ; the ~ of the group أَصْغَرُ أَعْضاءِ
المَجْموعةِ
B adj before n 1 <brother, sister> صَغير (صُغَراءُ. صِغار); a ~
boy/girl صَبِيّ وَليد \ صَبِيّة وَليدة 2 <animal, bird> صَغير
(صُغَراءُ. صِغار); a ~ elephant صَغير الفيلِ 3 (for babies)
<clothes, food> أَطْفالٍ

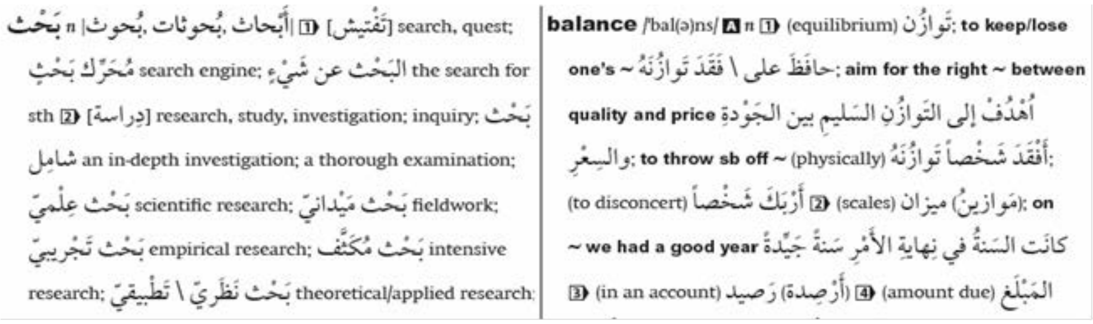
Figure 6: An English-Arabic entry with some of its structure illustrated.

9 Technical Challenges

A number of technical challenges had to be overcome in getting the dictionary writing software to work in harmony with the Arabic. The first challenge was the simple fact that Arabic is written from right to left and the letters within a word are joined up. This causes problems whenever Arabic and computers come into contact with each other, and made the news when Arabic and Persian visitors to the London Olympics were told the equivalent of N O D N O L O T E M O C L E W.

We had to make sure that the software could deal with English and Arabic, and even with English words inside Arabic tags, and vice versa, which gave the OUP dictionary technology staff no end of work. One of the largest problems turned out to be any numbers in examples, which in Arabic go in the same direction as in English, but which consistently were turned back to front in Arabic tags, since we had given the instruction that Arabic tags needed to output from right to left! The dictionary technologists then developed "bidirectional-override" tags, in which data that needed to output in a different direction than the default direction could be entered. This worked inasmuch as the numbers came out correctly... however, the rest of the data switched place, so the start of the sentence jumped to after the number and the back to before it. After months of tinkering we eventually had to instruct the typesetters to set all numerals back to front.

An additional complication is that our text is not just Arabic, but half Arabic and half English. Due to the large volume of text, we had to have English and Arabic constantly interplay, rather than having them in separate columns (see Figure 7).



بَحْث n |بُحوث, بُحوثات, أَبْحاث| 1▸ [تَفْتيش] search, quest;
مُحَرِّك بَحْث search engine; البَحْث عن شَيْء the search for
sth 2▸ [دِراسة] research, study, investigation; inquiry; بَحْث
شامِل an in-depth investigation; a thorough examination;
بَحْث عِلْمِيّ scientific research; بَحْث مَيْدانِيّ fieldwork;
بَحْث تَجْرِيبِيّ empirical research; بَحْث مُكَثَّف intensive
research; بَحْث نَظَرِيّ \ تَطْبِيقِيّ theoretical/applied research;

balance /ˈbal(ə)ns/ A n 1▸ (equilibrium) تَوازُن; **to keep/lose**
one's ~ حافَظَ على \ فَقَدَ تَوازُنَهُ; **aim for the right ~ between**
quality and price اُهْدُفْ إلى التَوازُنِ السَليمِ بين الجَوْدةِ
والسِعْرِ; **to throw sb off ~** (physically) أَفْقَدَ شَخْصاً تَوازُنَهُ;
(to disconcert) أَرْبَكَ شَخْصاً 2▸ (scales) ميزان (مَوازِينُ); **on**
~ we had a good year كانَت السَنَةُ في نِهايَةِ الأَمْرِ سَنَةً جَيِّدَةً
3▸ (in an account) رَصيد (أَرْصِدة) 4▸ (amount due) المَبْلَغ

Figure 7: Both sides of the dictionary have text running on.

The direction of the dictionary is left-to-right, so fitting right-to-left phrases in there was a major challenge, not in the least because some information that is to the right (e.g. the start of an Arabic example sentence) needs to go on the top line in the case of a line break, whereas other information on the right (e.g. the grammatical information with an Arabic translation) needs to go on the bottom line in case of a line break. It took many tries and elaborate diagrams until we had all the kinks sorted out.

For some users it may seem a bit odd at first to have to "jump" when reading a translation, but we have found that one gets used to it very quickly, and indeed this is the way the entries are presented in most dictionaries. The alternative, writing the English on the left and the Arabic on the right, leaves too much white space and isn't feasible for a print dictionary of this size.

10 Finding an Arabic Word in the Print Dictionary

An important factor of Arabic is that it is root-based, that is, every word has a root of (usually three) consonants carrying the basic meaning of a word (e.g. *ktb* with basic meaning "writing", which is modified by adding vowels and affixes, making *kātib* (writer), *kitāb* (book), *kataba* (he wrote), *kitāba* (writing), *maktaba* (library; bookshop), etc. Arabic dictionaries, with the exception of learners' dictionaries, are usually ordered by these roots, rather than in "proper" alphabetical order. We chose to do this as well, since the advantage of having all words of one root in one outweighs the difficulty a beginner may have in finding the root of a word. This means the Arabic-English side has a kind of double ordering, first the roots in alphabetical order, then a logical order for the words within one root.

Loanwords do not have an Arabic root. For them we listed each written letter as a root letter, and fitted them in into the root system like that.

11 Finding an Arabic Word in the Electronic Dictionary

Though all words are fully vocalized, we cannot expect the user of the electronic dictionary to enter a word with all its vowels, especially not if it's a word they don't know – not knowing a word means you could at best make an educated guess at its vowels, and there is no standard system of vocalization. So one of the challenges for the electronic dictionary was to make it so that the user is able to enter the unvowelled or partially vowelled form of a word, and be redirected to the entry or entries that correspond to that form.

But that is not the biggest challenge. I mentioned before the many possible surface forms of one lemma. Also, words are often joined together: the article *al-* is prefixed to the noun or adjective, object and possessive pronouns are suffixed, and some grammatical words like *wa-* (and) and *li-* (for) are joined to the word following them as well. All this can be combined, so for example *wali'uxtihi* (and for his sister) is one string. This can make strings of letters highly ambiguous – a common string like *l'nh* can be interpreted in at least seven different ways! We want the user to be able to enter any string they don't understand, without having to distinguish the different morphemes themselves, which can be a challenge for even quite advanced Arabic learners. For this, we again use the Buckwalter Arabic Morphological Analyzer integrated with our own headword list. Thus the strings are analysed into the appropriate morpheme(s) and the user will be redirected to the relevant entry. For example if a user enters the string وكتبه (and his books), they will be redirected to كِتاب (book), with the information that it's preceded by وَ (and) and followed by هُ (his). In case of multiple possible analyses, like كتب, which can mean "he wrote", "writing", or "books", it gives the possible entries with a summary of the part of speech and meaning, and allows the user to choose which entry to display fully.

12 Conclusion

The world of Arabic lexicography is a very challenging field, where little can be relied on, and expected and unexpected pitfalls abound. Despite, or maybe even because of this, it is a fascinating area, where truly significant achievements can be made. With English and Arabic being world languages, and being two of the six official languages of the UN, it is amazing that so few resources exist for translation between the two, with no reasonably modern ones.

We feel that this dictionary fills that hole, and with the opportunity to constantly update that online dictionaries offer, will continue to fill the gap for many years to come.

We hope this look behind the scenes has been interesting for lexicographers and other linguists.

13 References

Arts, T. & McNeil, K. (2013). Corpus-based lexicography in a language with a long lexicographical tradition: The case of Arabic. In *Proceedings of WACL'2, Second Workshop on Arabic Corpus Linguistics, 22 July 2013.* Lancaster University, UK.

Arts, T. et al. (Forthcoming 2014). *Oxford Arabic Dictionary*. Oxford: Oxford University Press.

Attia, M., Pecina, P., Toral, A. & Van Genabith, J. (2014). A corpus-based finite-state morphological toolkit for contemporary Arabic. In *Journal of Logic and Computation*, 24 (2), pp. 455-472.

Baalbaki, R.M. (2004). Coinage in Modern English-Arabic Lexicography. In *Zeitschrift für Arabische Linguistik*, 43, pp. 67-71.

Baalbaki, M. & Baalbaki, R.M. (2013) *Al-Mawrid Al-Hadeeth*. Beirut: Dar El-Ilm Lilmalayin.

Benzehra, R. (2012). Modern English-Arabic Lexicography: Issues and Challenges. In *Dictionaries: Journal of the Dictionary Society of North America*, 33,pp. 83-102.

Doniach, N.S. (ed.) (1972). *Oxford English-Arabic Dictionary*. Oxford: Oxford University Press.

Habash, N., Rambow, O. & Roth, R. (2009). MADA + TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR).* Cairo, Egypt.

Haywood, J.A. (1965). *Arabic Lexicography*. Leiden: Brill.

Hoogland, J. (2003). Coverage. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: http://wba.ruhosting.nl/Content1/1.4_Coverage.htm [10/04/2014].

Hoogland, J. (2004). Working Methods. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: http://wba.ruhosting.nl/Content1/1.4_Working_Methods.htm [10/04/2014].

Hoogland, J. et al. (2003). *Woordenboek Arabisch-Nederlands*. Amsterdam: Bulaaq.

Ibrahim, R. & Aharon-Peretz J. (2005). Is Literary Arabic a Second Language for Native Arab Speakers?: Evidence from Semantic Priming Study. In *Journal of Psycholinguistic Research*, 34 (1), pp. 51-70.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D.A. (2004). The Sketch Engine. In *EURALEX Lorient Proceedings*. Lorient, France.

Van Mol, M. & Bergman, K. (2001). Leerwoordenboek Arabisch. Amsterdam: Bulaaq.

Wehr, H. (1979) *A Dictionary of Modern Written Arabic*. Rev. ed. Urbana: Spoken Language Services.

Simple and Effective User Interface for the Dictionary Writing System

Kamil Barbierik, Zuzana Děngeová, Martina Holcová Habrová, Vladimír Jarý,
Tomáš Liška, Michaela Lišková, Miroslav Virius
Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i.
kamil.barbierik@fjfi.cvut.cz, dengeova@ujc.cas.cz, holcova@ujc.cas.cz,
vladimir.jary@foxcom.cz, tomas.liska@foxcom.cz,
liskova@ujc.cas.cz, miroslav.virius@fjfi.cvut.cz

Abstract

A new monolingual dictionary of the contemporary Czech language is being prepared by the Institute of Czech language at the Academy of Sciences of the Czech Republic, v.v.i. A dictionary writing software is being developed as part of a grant supported by the Ministry of Culture of the Czech Republic within the National and Cultural Identity (NAKI) applied research program. We will present the overall architecture of the software and then focus on its user interface and two modules: the referencing system and a new module – the editorial tool (that was promised in (Barbierik 2013)).

Keywords: dictionary writing system; DWS; cross-reference module; editorial tool; lexicography

1 Introduction

Since 2012, the Department of Contemporary Lexicology and Lexicography within the Institute of Czech Language at the Academy of Sciences of the Czech Republic, v. v. i., has been preparing a new monolingual dictionary of contemporary Czech.

Its working title is Akademický slovník současné češtiny (The Academic Dictionary of Contemporary Czech). It is a medium-sized dictionary with an expected number of 120,000–150,000 lexical units.

To aid this project, a new Dictionary Writing System (DWS) is developed. More information about the project can be found in (Kochová 2014). A detailed specification of the requirements from the lexicographer's point of view can be found in the article A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System (Barbierik 2013).

We introduce basic functionality of our DWS with emphasis on the user interface in this paper. Further, we focus mainly on the editorial tool which will be described in more detail.

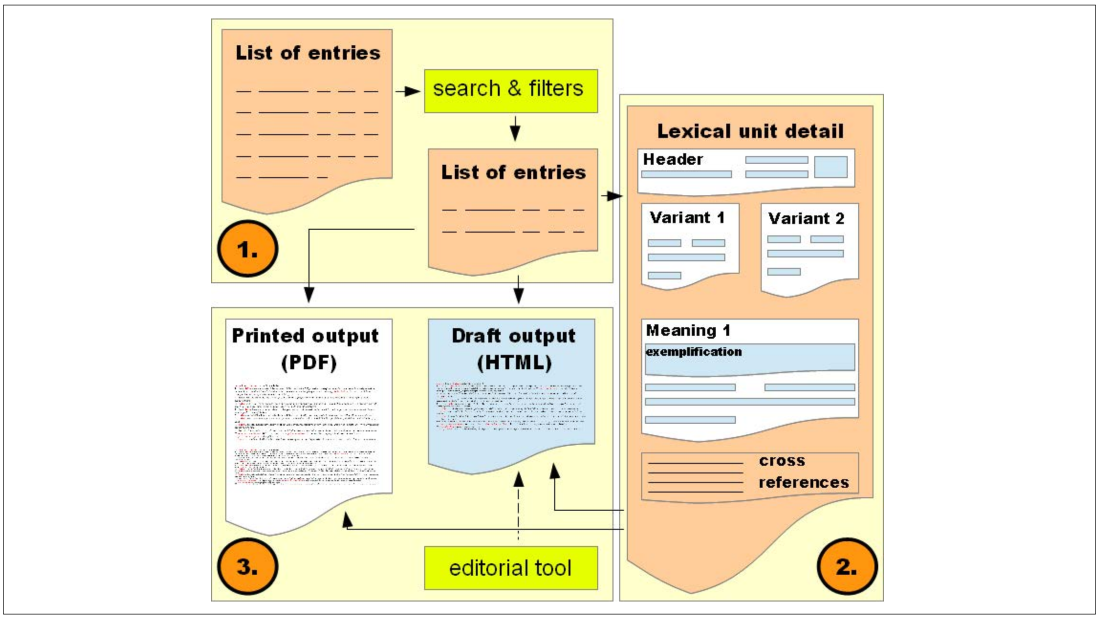


Figure 1: Basic scheme of our Dictionary Writing System.

2 Existing Dictionary Writing Systems

There are several commercial Dictionary Writing Systems (DWS) available (e.g. TshwaneLex (2013), IDM DPS (2013),

iLEX (2013)) as well as open-source systems (e.g. the Matapuna Dictionary Writing System (2013)). The DEB II (2013),

(DEBDict 2013), dictionary editor and browser, is available for the Czech language.

The lexicographic team at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i., considered three options: to buy an existing commercial DWS, to use one of the open-source systems or to develop their own system. One significant criterion used for the DWS selection was the amount of necessary adjustments due to specifics of the compilation of the dictionary and the time allocated to this task. Another criterion was the DWS price. After evaluation of the DWSs available, we decided to develop our own DWS that will fully respect the significant specifics of the compilation of the dictionary (Abel 2012: 1-23; Atkins 2008). The greatest advantage of this decision, which the lexicographic team benefits from, is the fact that any request for the user interface, some process modification or a new handy feature implementation etc., can be processed and implemented almost immediately.

3 Basic Functionality of Our DWS

From the common user point of view, the software is divided into three parts as it is shown on Fig. 1: the list of entries, the lexical unit detail and the output. The numbers of these main parts are used in the titles of the following subsections.

3.1 List of Entries (part 1)

After successfully logging into the system, the user inputs the list of entries which mainly represents the macrostructure of the dictionary. It is a list of lexical units with some basic information that is important for linguists at this stage.

Some helpful functions are available for better orientation in the long list of entries. The most important one is probably the quick search engine together with a set of predefined filters. The quick search allows users to search for entries by selecting any field of the microstructure of lexical unit (through which the user intends to search) and entering a search query. The search query can contain wildcards granting users better control over the search results. In addition, the query field is automatically updating its mode according to the type of information the user is searching for to make the search process easier for the user. For instance, if the user is trying to find information in fields where only a few values are available (e.g. type of lemma), the query field updates itself to select box. Thus, the user does not have to guess what values are available within the selected field. To avoid typing errors, an auto-complete function is implemented when searching in fields that contain short texts (like lemma).

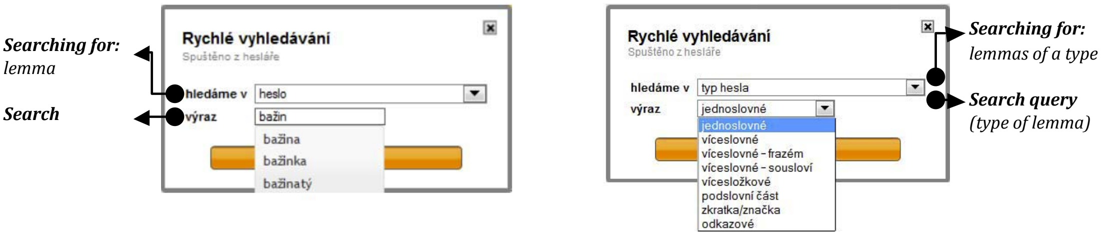


Figure 2: The "Quick search" function with auto-complete (left) when searching for lemma and select box when searching lemmas of certain type (right).

For example, the user is able to filter out, with combination of the quick search with predefined filters, the following entries:

- entries which I (the logged in user) founded and begin with the defined letter,
- entries of a specified word class created in some time interval,
- entries from manual selection containing some phrase in any of its exemplification, etc.

3.2 Detailed View of The Lexical Unit – Editing Module (part 2)

When the lexicographer finds the entry, he or she may continue to the detailed view of it. The detailed view of the entry contains all the information mainly from microstructure of the lexical unit that is available in a well-arranged and sophistically structured way. Additionally, the user can edit any information required in this view. To make the editing process more effective, different input fields are used according to the type of information it needs to gather. The whole editing form is designed so that the linguists editing the lexical unit information do not need to learn any special markup language or have any advanced computer literacy skills.

This editing form is organized into 4 sections (see Fig. 1 – Part: Lexical unit detail):

(1) Header

(2) Section of variants

(3) Meanings

(4) Cross-references

Header section. General information about the lexical unit can be found in the header section. It contains the entry status indicator which shows the progress of the work on this entry. Also, the output status can be set in this section which indicates in which output (electronic or paper) the entry will be presented. Furthermore, it contains information about the time of creation and about the last editing of the entry. The header section also contains information on the responsible user as well as a field very where the lexicographers may leave a note concerning the entry.

Section of variants. The section of variants may contain one or more variants of the lemma with all required microstructure elements. The variants of variant lemmas are often equivalent in majority of values, thus the function "Add variant as a copy of the last one" was implemented. This function creates a new variant and copies all the values from the last existing variant to it. Thus, only a few values have to be edited in the new variant. Consequently, creating new similar variants is much more efficient.

Section of meanings. The section of meanings consists of one or more panels, where the meaning of the word is described together with other related information. The section is organized as a set of panels. Each panel contains a large form, where the information about the meaning can be edited. The user can change the order of the panels; this will affect the order in the dictionary printed or electronic output. Meanings are numbered and when reordered, the numbering of meanings is automatically updated. The panel containing the form, with information related to the meaning, can be minimized or maximized according to which panel the user intends to work with. It helps for better orientation, whereas some lemmas may have quite a lot of meanings. The quick navigation is also helpful when a word has a lot of meanings.

This allows navigating directly to the meaning with the certain number, without scrolling the page.

Implementation of the editing form. As we mentioned at the beginning of this section, the whole detailed view is basically a well arranged set of fields of different types. These field types were chosen according to the types of information contained in lexical unit microstructure. For the gathering of short textual information (mostly comments, but also pronunciation for instance, synonyms, etc.) we have used simple one row text input field. Multiline longer texts are collected using textboxes. Often it is necessary to format the input text in some way. Special textboxes with rich text functions were implemented for this purpose. Such fields are used to store the exemplifications or the meaning explanations. Probably the most complex input fields are administrated select boxes which provide lexicographers with finite number of options prepared by the administrative user. This prevents editors from committing typing errors and unifies the values in certain places in microstructure through the whole dictionary. But it does not limit them thanks to the option of adding their own entry if it is necessary. The statistics of these entries are collected, and if some value is used too frequently, the administrative user may integrate it to the select box and "standardize it" very easily. Due to limited range we cannot provide an adequately descriptive picture of the editing form. For more information about the editing form, please refer to our poster "Simple and effective user interface of our new DWS".

3.3 The Output Module (part 3)

It is possible to evoke the output view of one or more lexical units from the detail of the lexical unit as well as from the list of entries. The output module takes the information collected using the editing form described above and utilizes some complex and very strict formatting rules on them to form the output. Thus, the user has a great possibility to preview the entry (or more entries at once) in its printed form and to see how it will exactly look like in the printed dictionary.
Two outputs are available in our DWS system: printed and electronic output. The printed output is not editable and it is presented in PDF format ready to be printed on the paper. The electronic or draft output is presented in HTML form. Even this output is not editable, but it is possible to implement additional interactive functions for it. Thanks to HTML format the user can interact with it using a web browser. One of the interesting functions we designed and implemented in this output is an editorial tool. It allows the lexicographers to fine tune the output or to correct mistakes or inconsistencies in cooperation with other lexicographers or editors.

4 Recently Implemented Features

4.1 Cross-References Module

A very necessary feature of the system, especially from the linguistic point of view, is the ability to define relations between entries. Relations are defined between two dictionary entries; one of the entries is considered to be the main or "master" entry, the second is the "slave". The system always allows creating the connection from both sides. This means that the user may define the relation if he is editing the slave as well as the master entry.

There are different types of relations from the linguistic point of view: run-on entries, references between one-word and multi-word lexical units and linked entries. From the user point of view, each type of relation needs a slightly different approach, but from the system perspective it is always just a relation between two entries supplemented with some information that is important from the linguistic point of view. When the user is at the detailed view of the lexical unit, he can always define all available types of cross-references to another entry. The window of the referencing module is evoked by clicking on the buttons at certain sections of the editing form. These buttons are placed according to the element of the microstructure from which the user references the other entry. For instance, run-on entry may be referenced from the whole entry (the button is at the end of the form) or from any of its meanings or exemplifications (buttons are under the corresponding input fields). The referenced units are then displayed at correct places according to this information when the output is compiled.

Other type of referenced entry is the linked entry. It can be referenced from the whole lexical unit or from the particular meaning of the entry to other foreign entry or its meaning. Thus, the button for bringing up the referencing tool popup is always at the end of the editing form.

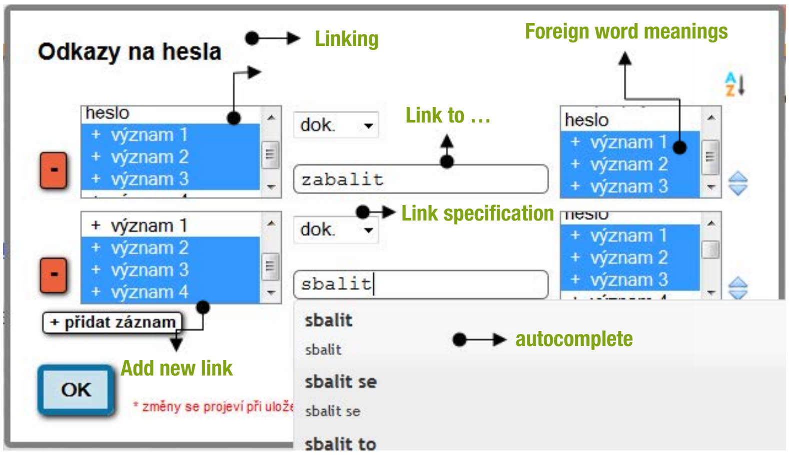
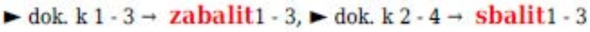


Figure 3: The cross-reference dialog box for linking words.

The interface for referencing is very simple and contains some clever functions to help the user to reference and to manage entries effectively.

Fig. 3 above is a snapshot of popup of referencing module evoked from lemma "balit" and it allows creating references (of type linked entry) to foreign entries and their meanings. As can be seen, references to multiple entries may be defined at once. The reference is created by putting the desired referenced lemma in the middle text input field. These inputs are provided by auto-complete functionality to make the process easier for the user. After writing in the foreign lemma, all its numbered meanings are loaded to the right select-box. Thus, the user knows how many meanings the foreign entry contains and he can comfortably choose the desired ones. Additional information to each reference can be added using a select box. When more references are defined (two in our case), it is reasonable to have an ability to sort the entries. It is possible to do it manually using little arrows next to each referenced entry, or to sort it alphabetically by the program using the AZ button in the top right corner. Links, if any, are according to the formatting rules for creating the monolingual dictionary attached to the end of each entry and our example will produce the output shown in Fig. 4 at the end of the word "balit" definition in the printed output.



► dok. k 1 - 3 → **zabalit**1 - 3, ► dok. k 2 - 4 → **sbalit**1 - 3

Figure 4: The printed output of linked words to the word "balit".

Editorial Tool

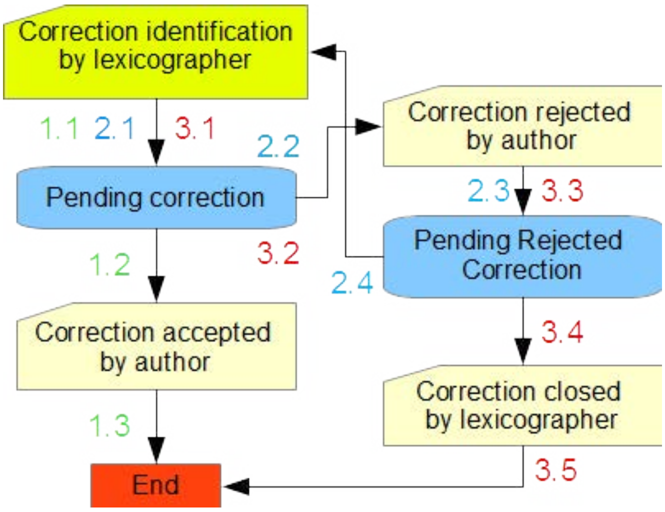
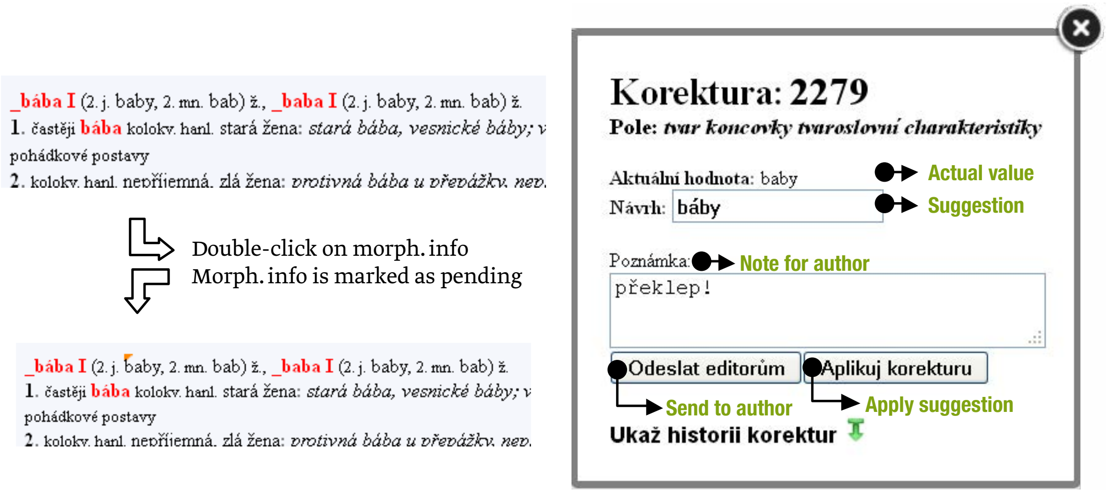

Correction identification by lexicographer
1.1
2.1
3.1
2.2
Correction rejected by author
Pending correction
2.3
3.3
Pending Rejected Correction
1.2
3.2
2.4
Correction accepted by author
3.4
Correction closed by lexicographer
1.3
3.5
End

Figure 5: The editorial process.

In order to produce and maintain a high quality dictionary, our DWS system implements an editorial tool, which is, as mentioned above, connected to the electronic HTML output. This editorial tool has been projected to replace the standard editorial process, where some portion of the submitted entries is printed on the paper, sent to the other lexicographers or editors, and received back reviewed. By implementing this feature directly in the DWS system, we save a vast amount of paper, as well as the time and money for the transaction agenda. The whole dictionary, each and every entry of it, is always ready to be reviewed without printing or posting anything.

The editorial process on one entry is captured schematically on Fig. 5. When the author (lexicographer) of the entry submits it to the system, the reviewer is able to see it and to review it using the draft (electronic) output. This draft output is very similar to the final printed output, so he or she is revising the entry almost as it was printed on the paper.



_**bába I** (2. j. baby, 2. mn. bab) ž., _**baba I** (2. j. baby, 2. mn. bab) ž.
1. častěji **bába** kolokv. hanl. stará žena: *stará bába, vesnické báby; v*
pohádkové postavy
2. kolokv. hanl. nepříjemná, zlá žena: *protivná bába u přepážky, nep.*

Double-click on morph. info
Morph. info is marked as pending

_**bába I** (2. j. baby, 2. mn. bab) ž., _**baba I** (2. j. baby, 2. mn. bab) ž.
1. častěji **bába** kolokv. hanl. stará žena: *stará bába, vesnické báby; v*
pohádkové postavy
2. kolokv. hanl. nepříjemná, zlá žena: *protivná bába u přepážky, nep.*

Korektura: 2279

Pole: ***tvar koncovky tvaroslovní charakteristiky***

Aktuální hodnota: baby

Návrh: báby

Poznámka:

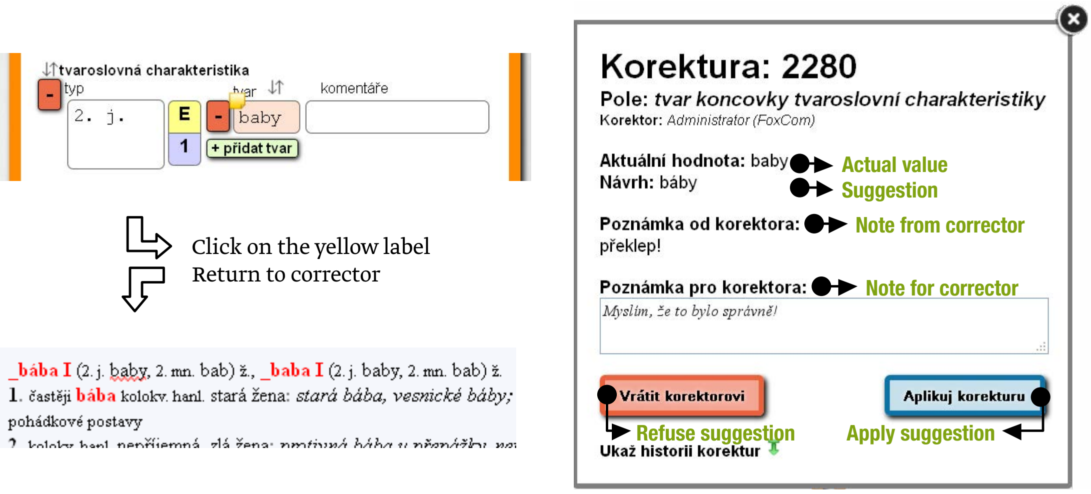
překlep!

Figure 6: The correction founding and sending it to editors.

By clicking on the information in the draft output, that the lexicographer wants to correct, he or she gets a popup window – see Fig. 6, where he or she can input his suggestion for correction, make a note about the correction for the author and send it to the author with a single click.

This is how the correction identification happens. There is a pending correction from this moment. This is indicated to the author of the entry (and not only him or her, but to other signed in users too) by highlighting the field that contains the corrected information – see Fig. 7.

By the click on the yellow "correction icon" nearby the highlighted field the popup will appear with the suggested correction from the lexicographer.



Click on the yellow label
Return to corrector

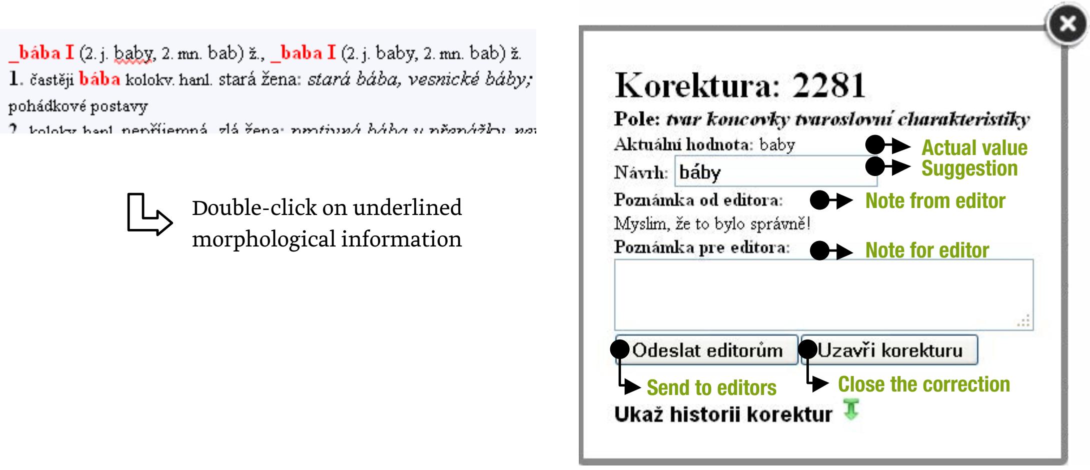
_bába I (2. j. baby, 2. mn. bab) ž., _baba I (2. j. baby, 2. mn. bab) ž.
1. častěji bába kolokv. hanl. stará žena: *stará bába, vesnické báby;* pohádkové postavy

Korektura: 2280
Pole: *tvar koncovky tvaroslovní charakteristiky*
Korektor: *Administrator (FoxCom)*
Aktuální hodnota: baby — Actual value
Návrh: báby — Suggestion
Poznámka od korektora: — Note from corrector
překlep!
Poznámka pro korektora: — Note for corrector
Myslím, že to bylo správně!
Refuse suggestion
Apply suggestion

Figure 7: The correction from the author's point of view.

The author may now decide whether he accepts the correction (take the green path number 1 in scheme on Fig. 5), or reject it (the blue path number 2 or the red path number 3 on Fig. 5). In the case he accepts the pending suggested correction, the system automatically updates the entry and the process successfully ends.

If author rejects the pending correction suggested by the (other) lexicographer – the corrector, it will be indicated in the draft (electronic) output – see Fig. 8. By clicking on it, the (other) lexicographer may view it together with the author's note on why it was rejected. At this stage the lexicographer who suggested the correction has two options: to close the correction (following the red path number 3 on Fig. 5) or to suggest a new one (following the blue path number 2 on Fig. 5). If the correction is closed, no changes in the entry are made and the process ends. If a new correction is designed, the process is started again.



_bába I (2. j. baby, 2. mn. bab) ž., _baba I (2. j. baby, 2. mn. bab) ž.
1. častěji bába kolokv. hanl. stará žena: *stará bába, vesnické báby;* pohádkové postavy

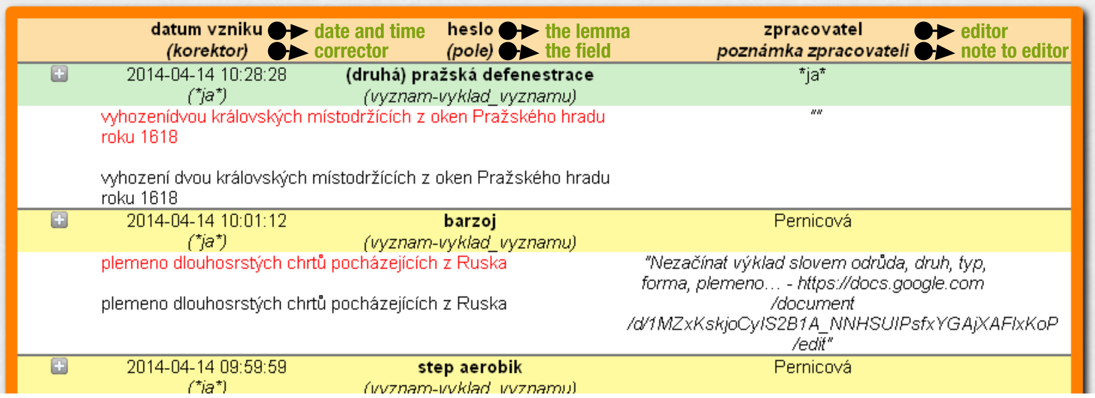
Double-click on underlined morphological information

Korektura: 2281
Pole: *tvar koncovky tvaroslovní charakteristiky*
Aktuální hodnota: baby — Actual value
Návrh: báby — Suggestion
Poznámka od editora: — Note from editor
Myslím, že to bylo správně!
Poznámka pre editora: — Note for editor
Send to editors
Close the correction

Figure 8: The correction refused by editor.

There is one more option for the lexicographer when he or she defines the correction that is not indicated in Fig. 5. He or she may immediately accept it – see Fig. 6 (the “Apply suggestion” button) - instead of creating a pending correction. Doing so, he or she directly updates the entry. This feature speeds up the process, when obvious typing errors are detected, because the pending correction does not have to wait for the author’s approval.

Every correction made using this editorial tool is recorded together with the old and new value of the time stamp, every action is signed by lexicographer, who changed the value or status, and their comments are also recorded. Thus, all the information that was ever corrected has an editorial history. It never gets lost and is always available in the editorial module pop-up. Thus, when the author is deciding whether to accept or reject some suggestion from another lexicographer, he can check the history of corrections made, find out who and when the suggestions were made. With the inclusion of a notes feature, he may even know why and under what circumstances they were made.



	datum vzniku (korektor)	heslo (pole)	zpracovatel poznámka zpracovateli
⊞	2014-04-14 10:28:28 (*ja*)	(druhá) pražská defenestrace (vyznam-vyklad_vyznamu)	*ja*
	vyhozenídvou královských místodržících z oken Pražského hradu roku 1618 vyhození dvou královských místodržících z oken Pražského hradu roku 1618		""
⊞	2014-04-14 10:01:12 (*ja*)	barzoj (vyznam-vyklad_vyznamu)	Pernicová
	plemeno dlouhosrstých chrtů pocházejících z Ruska plemeno dlouhosrstých chrtů pocházejících z Ruska		"Nezačínat výklad slovem odrůda, druh, typ, forma, plemeno… - https://docs.google.com/document/d/1MZxKskjoCyIS2B1A_NNHSUIPsfxYGAjXAFIxKoP/edit"
⊞	2014-04-14 09:59:59 (*ja*)	step aerobik (vyznam-vyklad_vyznamu)	Pernicová

Figure 9: The history of corrections.

5 Conclusion

Our DWS has been released and the lexicographic team uses it in their everyday work. Nevertheless, we are preparing additional modules for our DWS. This article was devoted to the editorial tool that has been deployed recently and we present it here for the first time.

We have strictly emphasized the quality of the user interface of our DWS. It must be designed according to the needs of the lexicographers that use it for the processing of large amount of lemmas.

Lexicographers are now processing lemmas using the described tools and preparing them to be published. Meanwhile we are preparing, except printed output, several applications, where published lemmas will be available for public. Using these applications like web pages, mobile applications for iOS or Android operating systems, users will be able to search and browse the dictionary on different devices.

Currently, we are preparing a very strong relation based search tool called xFilter which we will present sometime in the future.

6 References

Barbierik, K. et al. A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In Proceedings of the 7th international conference Slovko, 13-15 November 2013. Slovenská akadémia vied, Jazykovedný ústav Ľudovíta Štúra, pp. 9-26.

DEB II. Accessed at: http://deb.fi.muni.cz/index-cs.php [11/10/2013]

DEBDict. Accessed at: http://deb.fi.muni.cz/debdict/index-cs.php [11/10/2013]

IDM DPS. Accessed at: http://www.idm.fr/products/dictionary writing system dps/27/ [11/10/2013]

iLEX. Accessed at: http://www.emp.dk/ilexweb/index.jsp [11/10/2013]

Kochová, P., Opavská, Z., Holcová Habrová, M. (2014). At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project). *Poster presented on this conference.*

Matapuna. Accessed at: http://sourceforge.net/projects/matapuna/ [11/10/2013]

TshwaneLex. Accessed at: http://tshwanedje.com/tshwanelex/ [11/10/2013]

Abel, A. and A. Klosa 2012. 'The lexicographic working environment in theory and practice.' In R. V. Fjeld and J. M. Torjusen (eds.), *Proceedings of the 15th EURALEX International Congress.* Oslo: University of Oslo, 1–23.

Atkins, B. T. Sue and M. Rundell 2008. *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Acknowledgement.

This work has been supported by the grant project of the National and Cultural Identity (NAKI) applied research and development program A New Path to a Modern Monolingual Dictionary of Contemporary Czech (DF13P01OVV011).

Totalitarian Dictionary of Czech

František Čermák
Institute of the Czech National Corpus, Charles University
frantisek.cermak@ff.cuni.cz

Abstract

Due to a decades-long experience with the Communist totalitatian regime, an attempt was made to capture main features of its language, and thus the substance of the time, by a corpus-based dictionary. The corpus, based on large samples of Communist newspapers and booklets, which is now accessible on the web, served as basis for a frequency- based dictionary of the time; it has been provided with statistics according to time periods, with pertinent collocations, appearance in various genres and, most of all, with comparative figures with today´s standard as it is to be found in a large contemporary corpus of Czech. This, by far the first systematic description of the period in any language, is complemented by a detailed study of the lexis of the time, many typical excerps, etc.

Keywords: corpus; frequency dictionary; dictionary of a period; Communist totalitarian vocabulary

1 Goal and Historical Background

Twenty years after the decline of the Communist regime in Czechoslovakia, the first attempt (Čermák et al. 2010) has been made, on the basis of corpus-data, to map the language of the period of over four decades of the Communist rule (1948-1989) in the Czech-speaking part of the country, known now as Czechia. Whether too soon or too late, the decision was precipitated by realization how fast people tend to forget the totalitarian time, young generation not knowing it at all. However, it seems that it is through the vocabulary of a period that one gets to know it best. Needless to say that the objective of solving the situation is best served by corpus-based data offering an eminent source of information about ways of manipulation of the whole society through its rather effective propaganda. There seem to be very few similar dictionaries delving deep into the language of a period objectively and systematically, based on corpus data. Thus the book is both a lexicographic and linguistic record of that time as well as a valuable historical and sociological description of what had long been a taboo, namely propaganda, which, in many respects, drew amply on Hitler and Goebbels from an earlier period of fascism.

2 Data and Their Preparation

Since it was not possible to cover linguistically all texts from the period, the *Totalitarian corpus* (available on the web at korpus.cz) had to be limited to over 15 million positions made up of 422 762 word forms which equals to enormous 164 815 lemmas. The data have been drawn painstainkingly from the major and all-embracing Communist daily *Rudé právo* (Red Right) from three historically critical but different periods of the second half of the year 1952 (6. 6. - 31. 12. 1952), i.e. from the peak of the most primitive communist propaganda, the second quarter of 1969, i.e. from after the Soviet and Russian occupation of the country in 1968 (1. 4. - 31. 4. 1969) and from the first quarter of 1977 (3. 1. - 31. 3. 1977), i. e. time of deep social recession and depression. The newspaper data amounting to some 10 million words have been complemented by a careful and typical selection of 91 propaganda books and booklets (about 5 million words) from the same period. All of this data have been manually scanned and corrected having to come to grips with several different spelling reforms and various standardisation processes. No attempt has been made to cover a later period when it was felt that the political regime was characterized by stagnation and inertia as well as gradual crumbling of its strongholds which culminated in 1989, i.e. its definite downfall. The data have been tagged and lemmatized which made it necessary to adapt the existing taggers and lemmatizers. Of course, the fact that the Totalitarian corpus is freely accessible on the Web now means that any further analysis is possible, although the primary goal has been to cover the period´s typical vocabulary. It may be reasonably hoped that the core of the vocabulary is captured in dictionary form being accompanied by an extensive analysis, complements and comparison with today´s language.

3 The Dictionary of the Communist Totality

The *Totalitarian Corpus*, based on a prior decision to choose data from three historical periods of eminent linguistic interest (and the availabale propagandist books and booklets), had to be further limited before dictionaries based on it have been compiled, both capturing only the top selection of the most important lemmas.

One of the main goals of the project was to point out to users what used to be typical in the vocabulary of the period (**A**). The tricky business of corpus comparison of the period with today´s standard language, which is several decades remote from the time recorded in a large contemporary corpus of Czech (SYN 2005, see korpus.cz), has been projected into frequencies (**B**) in one of two dictionaries compiled on this basis, namely the **Specific Totalitarian Dictionary** (*Dictionary* in the following). Because this dictionary, based on such comparison, has been achieved due to use of statistical methods it is highly selective in its nature capturing only those lexemes whose frequency has had a high degree of deviation from today's use. The dictionary is alphabetical giving a summary of absolute (ABS) and relative (REL) frequency (in square brackets) and partial frequencies of each of the three periods examined (1952, 1969, 1977, given in ppm, i.e. parts per million). Numerical differencies between these three figures suggest a growing or falling use of the frequency of the words in the three periods. Obviously, many lexemes that have not seemed to be statistically different enough have been dropped and are not to be found here.

Where necessary, an additional brief explanatory note has been added identifying lexeme´s meaning. Typical collocations, that offer a more direct information about the use of the word are offered following a full dot • Should the lemma be out of use in today´s newspapers completely (i.e. not found in the contemporary 100-million corpus) it is preceded by a cross sign †, such as †**agrobiologický** (agrobiological). Positive, period-linked sign of a larger asterisk ★ standing before a collocation suggests a prominent use and specifically high frequency in the Communnist era, while the sign ✌ indicates, in contrast, a typical collocation of today, such as in **bezmezně** (boundlessly, infinitely, immensely) in „★ **b. oddáni** (*SSSR*) × ✌ ***b. věřit*** " (*immensely devoted* to USSR at the time, as against to *infinitely believe* in today´s usage). The overwhelming devotedness to the Soviet Union, often found elsewhere, is captured in this example prominently, showing also how the usage of the verb *believe* (věřit) has eventually returned to normalcy now.

Collocations have been obtained by getting a combination of predetermined values of three association measures of lexemes having a frequency higher than 5, namely of MI-score, log-likelihood and Dice. It is specifically collocations that reveal the usage of heavily loaded words of the time such as fight/struggle:

boj (struggle, fight) • *boj dělníků* (workers´ struggle); *boj lidu* (za mír) (people´s fight for peace); *boj národů* (za mír) (fight of nations for peace); boj proletariátu (proletariat´s struggle); *ideologický boj* (ideological struggle); *ideový boj* (struggle of ideas); *organisovat boj* (pracujících) (organise fight of the workers); *podporovat boj* (národů) (support fight of nations); *revoluční boj* (revolutionary fight); *rozvíjet boj* (proti imperialismu) (develop struggle against imperialism); *rozvinout boj* (za mír) (develop fight for peace); *stávkový boj* (fight of those on strike); *třídní boj* (class struggle); *řídit boj dělnické třídy* (direct the fight of the workers´s class), etc.

On a closer inspection, it is evident that the society lived then in a state of frenzy feeling both endangered from outside and feeling that it must fight for almost everything, including even the most common, everyday things.

4 A Sample of the Specific Dictionary

Let us have a look at the Dictionary and some of its lemmas:

Adenauer ABS 348, REL 27 ppm

[**83 ppm** | 1 ppm | 0 ppm] *Konrad Adenauer, kancléř NSR (1949-1963)*

Adolf ABS 92, REL 7 ppm

[**6 ppm** | 12 ppm | 5 ppm] • Adolf Hitler

agent ABS 1292, REL 100 ppm

[**263 ppm** | 22 ppm | 24 ppm] • agent buržoazie; agenti imperialismu; imperialistický agent; †titovský agent

agitace ABS 894, REL 69 ppm

[**104 ppm** | 12 ppm | **85 ppm**] • †agitace komunistů; názorná agitace; osobní agitace; politická agitace; účinnost ekonomické propagandy a agitace

agitátor ABS 811, REL 63 ppm

[**166 ppm** | 2 ppm | 23 ppm] • †kolektiv agitátorů; †příprava agitátorů; †seminář agitátorů; †schůze agitátorů; stranický agitátor

agrese ABS 1213, REL 94 ppm

[**188 ppm** | 50 ppm | 50 ppm] • hitlerovská agrese; imperialistická agrese; (definice) pojmu agrese; politika agrese

americký ABS 15089, REL 1170 ppm

[**2527 ppm** | 601 ppm | 476 ppm] • americký agresor; američtí barbaři; americký businessman; americký imperialismus; americký katan; americký monopolista; americké okupační úřady; americký okupant; †američtí podněcovatelé války; †americká soldateska; američtí supermani; provokace amerických imperialistů

angažovaný ABS 275, REL 21 ppm

[0 ppm | **38 ppm | 26 ppm**] • angažované země; †schůzka neangažovaných zemí

armáda ABS 6993, REL 542 ppm

[**1067 ppm** | 340 ppm | 260 ppm] • lidová armáda; (spartakiáda) spřátelených armád; západoevropská armáda; Svaz pro spolupráci s armádou (*viz Svazarm*)

balistický ABS 16, REL 1 ppm

[**0 ppm** | 1 ppm | 2 ppm] ★ (mezikontinentální) b. raketa × ✌ b. expertíza

bdělý ABS 86, REL 7 ppm

[**18 ppm** | 1 ppm | 2 ppm] ★ politicky b. × ✌ b. stav

bezmezně ABS 23, REL 2 ppm
[**4 ppm** | 0 ppm | 1 ppm] ★ b. oddáni (SSSR) × ✌ b. věřit

5 Basic Dictionary of the Communist Totality

To balance this key, though highly (i.e. statistically) prominent, dictionary, a second dictionary, that of Basic Dictionary (Základní slovník) has been compiled that is complementary to the Specific Totalitarian Dictionary. **The Basic Dictionary** is represented by almost ten thousand most frequent lemmas from the Totalitarian Corpus (9994, i.e. those having more than 63 occurrences). Thus a comparison of the specific period-related vocabulary obtained statistically and that of the systematic full dictionary is enabled, the latter enabling to find there also those words that were used but were not statistically interesting enough. This dictionary, based on descending absolute frequency ordering, shows all lemmas provided within the given frequency limits. The sum-total of frequency for each lemma is broken into three different frequency figures for each of the three periods covered. For cross-reference purposes, those lemmas given also in the Specific Totalitarian Dictionary are marked by asterisk.

Lemma	Overall	1952	1969	1977
a	498450	166271	130175	202004
v	422528	129014	126641	166873
být*	339917	110927	101912	127078
se	225925	67849	72378	85698
na	214183	70624	66385	77174
ten	120081	36929	36986	46166
s	106963	31667	33946	41350
který	104580	36545	30776	37259
že	99118	34209	31456	33453
z	95681	27971	31637	36073
strana*	52942	22081	12821	18040
všechen	51970	22246	11180	18544
rok*	49349	13494	15128	20727
práce*	44491	17044	8300	19147
jeho	39773	11046	11304	17423
aby	37883	17291	8835	11757
sovětský*	36644	22690	4495	9459
socialistický*	35990	7450	5373	23167
nový*	34985	13992	8061	12932

Looking at the first 30 most frequent lemmas of the Basic Dictionary it is just appaling to see what kind of priority some of its words used to have in the Communist-ruled society at the time (although some words have been left out). Thus, the very first noun is *strana* (party, frequency 22) followed by *práce* (work) while the very first adjective used then was *socialistický* (socialist, fr. 29), followed by *nový* (new). The prominent presence of nouns does not need much comment telling what had been dominant at the time: Communist party and work. Likewise, the adjective socialist is to be expected here while the particular adjective new would deserve a special study, not possible here, since every second thing, notion, aspect of life coming from the past had to, on ideological grounds, be rejected and replaced by something new, i.e. *Communist*, most often *Soviet-inspired*, since Soviet Union had been presented as a new Promised Land serving as a model and exemplary goal to be followed by everyone and serving as an official inspiration for everyone. Some extensive collocations of this type showing this in detail are added in the Supplement to the Dictionary.

6 Additional Features of the Dictionary

Next to the two dictionaries, there is a large analytical study attached (*Slovník komunistické totality: lexémy, nominace a jejich užití, Dictionary of the Communist Totality: lexemes, nominations and their Use*) preceding both dictionaries, based on the data added, scrutinizing and researching typical words and collocations more deeply paying special attention to principles of the Communist propaganda. Some attention is paid to comparison of joint features of the fascist propaganda and the Communist one and a semiotic analysis of how certain lexemes were used for manipulation is shown.

Since all data have been published only after they underwent a severe censorship, an attempt is also made to counterbalance this *officious (official)* language by a sample of the most important words of the *unofficial* language used by normal people and heard on the street that were not subject to the official parlance of the time and never printed. It is shown that beginnings of this second, unofficial language or rather vocabulary was born and used profusely wherever people were in private and not followed by the secret police; thus a parallel vocabulary of synonyms was born (e.g. *strana-partaj, komunista-komouš,* etc). The study is complemented by a number of typical samples of texts, vocabulary and collocations, showing, for example, the blind reliance on anything Soviet whose meaning has changed from a proper name so much as to just practically designate only what was „good, best, exemplary".

7 Notes and Conclusions

Slovník komunistické totality (Dictionary of the Commununist Totality) maps the gap between a general dictionary and one of an epoch or period enabling, in this case, comparison of two comparatively close time periods, an attempt rather rare in lexicography, not to speak of its wide sociological and political impact.

The fact that there is a substantial specific corpus behind it (which is not the case of German, Polish or Russian where similar books have been published) is both a possibility and offer to to anyone who wishes to study the period language more widely, and, should other dictionaries arise in other languages (from the period of Communism and Fascism primarily) it might become a contribution to a broader understanding of political and social movements spanning substantial parts of Europe not long ago - and not only that.

8 References

Slovník komunistické totality (Dictionary of the Commununist Totality)**,** eds. František Čermák, Václav Cvrček, Věra Schmiedtová, further coauthors Jan Kocek, Dominika Kovářiková, Karel Kučera, Renata Novotná, NLN Praha 2010

Arendtová, H. (2000). *Původ totalitarismu I a II.* Oikoymenh, Praha.

Bartošek, K., J. L. Margolin (1999). *Černá kniha komunismu: Zločiny, teror, represe;* Paseka, Praha, Litomyšl.

Bralczyk, J. (2001). *O języku polskiej propagandy politycznej lat siedemdziesiątych.* Wydawnictwo Trio, Warszawa.

Brenner, Ch. (2008). *Znormalizovaný totalitarismus? Paradigmata výzkumu socialismu I,* Ročenka textů zahraničních profesorů; FFUK, Praha.

Brousek, A. (1987). *Podivuhodní kouzelníci, čítanka českého stalinismu v řeči vázané z let 1945-1955.* Rozmluvy, Londýn.

Cvek, B. (2005). *Proč a v čem je komunismus vlastně totéž co nacismus*; Britské listy 26. 2. 2005. Dostupné z WWW: <http://www.blisty.cz/art/22182.html >.

Džilas, M. (1977). *Nová třída, kritika soudobého komunismu.* Demos, Curych.

Fidelius, P. (1983). *Jazyk a moc.* Edice Arkýř, Mnichov.

Fidelius, P. (2000). O totalitním myšlení, In *Kritické eseje.* Torst, Praha.

Głowiński, M. (1991). *Nowomowa po polsku.* Wydawnictwo PEN, Warszawa.

Hochel, B. (1991). Totalita v jazyku (a v nás). In *Kultúrny život*, 25, Bratislava.

Jánský, P. (2004). *Totalita slovem, písní a obrazem (Dějiny hrůzovlády KSČ).* Nakladatelství Music, Cheb.

Klemperer, V. (2002). *Deníky 1933-1941, Chci vydat svědectví.* Paseka, Praha, Litomyšl (*Victor Klemperer: Die Tagebücher 1933–1945. Kritische Gesamtausgabe. CD-ROM. Berlin 2007*).

Klemperer, V. (2003). *Jazyk Třetí Říše – LTI: poznámky filologovy.* Nakladatelství H&H, Jinočany.

Kartous, B. (2004). *Může totalita zahltit celou civilizaci?* Britské listy 10. 11. 2004. Dostupné z WWW: <http://www.blisty.cz/art/20532.html>

Macura, V. (1992). *Šťastný věk.* Pražská imaginace, Praha.

Nowak, P. (2002). *Swoi i obcy w językowym obrazie świata: język publicystyki polskiej z pierwszej połowy lat pięćdziesiątych.* Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.

Mokijenko, V. M., T. G. Nikitina (1998). *Tolkovyj slovar jazyka sovdepii.* Sankt-Peterburg.

Mokijeno, V. M. (2000). *Novaja russkaja frazeologija*. Opole.

Popper, K. (1994). *Otevřená společnost a její nepřátelé I. a II.* Oikoymenh, Praha.

Pisarek, W. (1972). *Frekwencja wyrazów w prasie*. Ośrodek Badań Prasoznawczych, Kraków.

Pisarek, W. (1976). *Język służy propagandzie*. Ośrodek Badań prasoznawczych, Kraków.

Pisarek ,W. (1983). *Analiza zawartości prasy*. Ośrodek Badań prasoznawczych,Kraków.

Röhrich, A. (2008). *Ideologie, jazyky, texty: Analýza a interpretace textů Rudého práva z roku 1953 a 1975 a Práva z roku 1997.* Bor, Liberec.

Šebesta, K. (2001). Studovat jazyk totality. In: *Institucionalizace (ne)odpovědnosti: Globální svět, evropská integrace a české zájmy.* Karolinum, Praha.

Šlosar, D. (1995). Jazyk totality a jazyk dneška, In *Spisovná čeština a jazyková kultura 1993.* FF UK, Praha.

Totalitarismus 3 (2007). Eds. I. Budil a T. Zíková. FF ZU, Plzeň.

Totalitarismus 4 (2008). Eds. I. Budil a T. Zíková. FF ZU, Plzeň.

Vítězové? Poražení? I. díl, Disent v období tzv. normalizace (2005). Eds. M. Vaněk a P. Urbášek, Prostor, Praha.

Vítězové? Poražení? II. díl, Politické elity v období tzv. normalizace (2005). Eds. M. Vaněk a P. Urbášek, Prostor, Praha.

Wierzbicki, P. (1986). *Struktura klamstva*, Glos, Warsawa.

Włodek, M. (2002). *Sondy do jazyka totality.* diplomová práce, FFUK, Praha.

Dictionary of Abbreviations in Linguistics: Towards Defining Cognitive Aspects as Structural Elements of the Entry

Ivo Fabijanić
English Department, University of Zadar, Obala k. Petra Krešimira IV., 2,
HR-23000 Zadar, Croatia
ifabijan@unizd.hr

Abstract

The first part of the article deals with general information about the project of compiling a dictionary of abbreviations in linguistics. It also contains a short overview of past research together with their main results. So far, some specific theoretical and practical solutions were proposed. Theoretical solutions refer to the Multi-level approach in collecting data for submorphemic word-formations, which consists of three aspects: 1) *Structure and Modes of Production*, 2) *Cognitive Aspects*, and 3) *Functional Aspects*. Practical solutions for the structure and modes of production have already been recommended with the results properly substantiated by the examples of abbreviations. The second part presents results of the analysis for the cognitive aspect of the multi-level approach. The semantic part reveals the relationship between an abbreviation type and the semantic (sub-)field it might be assigned to. Semiotic part is achieved by designating a specific interpretation of a sign's meaning, while the analysis of lexicalization and institutionalization confirms their contingent addition to this scientific lexicon. The motivational aspect takes two conceptual perspectives in consideration: the narrower and broader senses of the formation, and various semantic relationship patterns which led to the classification of fully and partially motivated abbreviations.

Keywords: dictionary; abbreviations; linguistics; micro-structure; analysis; multi-level approach; cognitive aspect; semantics; semiotics; motivation; lexicalization; institutionalization

1 Introduction

The overall aim of this paper is to provide specific theoretical and methodological models in preparing both a macro- and micro-structural framework for compilation of the future bilingual and bidirectional (English – English – Croatian), specialized and explanatory dictionary of abbreviations in linguistics. The dictionary would cover the core areas of linguistics and its interdisciplinary areas as well. The main triggers which motivated us to study the lexicon of abbreviations are, by all means, the lack of consistent categorisation and typology, as well as fixed boundaries between the respective types of abbreviations, which are, unfortunately, their most distinctive characteristics. The classification of abbreviations used here largely relies on López Rúa's (2006) work. We find this taxonomy appropriate because it clearly distinguishes certain abbreviation types. *Abbreviations* are divided into *simple* and *complex abbreviations*. According to López Rúa, an *initialism* is made of initial letters or occasionally the first two letters of the words in a phrase and combined to form a new sequence (2006: 677). The term initialism denotes an abbreviation created through the usage of initial letters, which applies to both alphabetisms and acronyms. The term alphabetism denotes an abbreviation pronounced as a series of letters of the alphabet, while the acronym denotes abbreviations pronounced as whole words. *Clippings* are either shortened words or syllables without a change in meaning or functions (López Rúa 2006: 676). *Blends* are created by " [...] joining two or more word-forms through simple concatenation or overlap and then by shortening at least one of them" (López Rúa 2006: 677).

2 Previous Research in the Field

Previously explained and described in the past research (cf. Fabijanić 2014, in print), the main objectives in compiling the dictionary and its main characteristics refer to the following concepts: bilingual and bidirectional type of dictionary according to the number of languages, alphabetical order of headwords according to the order of presentation, appropriate use of the data which should be provided within the microstructural confines, and scientific, identified by the domain-specific collection of abbreviations in linguistics from specialized publications. So far, I have proposed some specific theoretical and practical solutions to be utilized in compilation of the entries (Fabijanić 2014, in print). These solutions refer to the concept of multi-level approach in collecting data for submorphemic word-formations (cf. Fandrych 2008a).

The triaspectual multi-level approach is comprised of the following stages: 1) *Structure and Modes of Production*, i.e. the structural aspects and word-formation potential, word class, medium and origin; 2) *Cognitive Aspects*, i.e. semantic, semiotic and motivational aspects, lexicalization and institutionalization, and 3) *Functional Aspects*, i.e. stylistic and sociolinguistic aspects, pragmatic and text-linguistic aspects. The application of this interdisciplinary approach will give a fuller and a more transparent picture of various orthographic, morphological, semantic, stylistic and functional processes involved in the production and uses of abbreviations.

As for the first aspect of *Structure and modes of production*, abbreviations have been classified according to two criteria – *narrower* and *broader sense* (cf. Fabijanić 2014, in print). The narrower sense refers to those formed by using initial letters of each element in the expansion (pertaining mostly to alphabetisms), and pronounced either by individual names of letters or as a word. The broader sense implies the ways and processes of formation, more or less different from the orthographic norms (pertaining mostly to hybrid forms, acronyms, blends and clippings featuring some orthographic changes), in consequence of which, one or more initials are used for various smaller elements of the expansion (smaller than words, yet bigger than initials). Due to this, initials for graphemes, compounds, and affixes, grammatical and lexical words found in the final form of an abbreviation, as well as different orthographic changes, such as ellipsis, conversion, metathesis, addition, etc., were analysed and (sub-)classified.

For the purpose of their differentiation, a system of exclusive classification and subclassification of abbreviations was proposed (Fabijanić, Malenica 2013). Miscellaneous realisations of abbreviations are generally diversified into two main groups: those realised in the narrower sense and those in the broader sense. Abbreviations in the narrower sense are exclusively explained with an *LLL* descriptor for initials used in their formation. Abbreviations in the broader sense are represented with a whole set of additional different letters or initials (written either in capital or small letters) added to a three-letter descriptor: e.g. *l* for small letters, *P* for initial affixes, *N* for numerals, *S* for syllables, and *W* for a word. Further orthographic changes are explained by other descriptors, e.g. *E* for ellipsis, *C* – conversion, *M* – metathesis, and *A* for addition of a word or a diacritic sign not normally found in expansions. Comprehension and consequently classification of abbreviations depends on the degree of their (non-)coordination with the common orthographic norms.

The research in *Structure and modes of production* of abbreviations has proven that most of the alphabetisms are formed according to the criterion in the narrower sense, while the ratio of those formed in the narrower and the broader sense for acronyms (which were fewer in number than alphabetisms) was in favour of the broader-sense formations (cf. Fabijanić 2014 in print). As for the hybrid-form ratio, the broader-sense criterion is also more evident. The direct results of the analysis have attested the possibility of applying previously devised descriptors (Fabijanić, Malenica 2013), as well as some new descriptors, which have emerged in the analsyis of abbreviations in linguistics (e.g. *P-LL* for alphabetisms, *FLL* for acronyms and some clippings in broader sense, *SFL* and *FFL* for blends, *Lll* and *lll* for clippings in broader sense, and *LLW*, *PLW*, *L/LW*, *FLW*, *F-LW*, *S-LW*, *SFL*, *WLL*, *W-LL* for hybrid formations). The Structure and modes of production aspect will be described by labels referring to form(s) of abbreviations, medium, word class, origin (cf. § 7).

3 Aims and Objectives of the Current Research

The immediate aim of this research is to bring forth the results of the analysis for the second aspect of the Multi-level approach, i.e. the cognitive aspect inherent to non-morphematic word-formations. The *Cognitive aspect* deals with semantic, semiotic and motivational aspects, as well as with the lexicalization and institutionalization of abbreviations. The semantic part gives answers to the relationship between a given word-formation of a specific abbreviation type and the semantic (sub-)field it may be assigned to. Semiotic part is achieved by designating a specific interpretation of a sign's meaning, while the analysis of lexicalization and institutionalization of abbreviations confirms their contingent addition to the lexicon of a language or to the specific scientific lexicon. As far as motivation is concerned, the relationship between the structural pattern of abbreviations, their meaning(s) and phonemic, graphemic and sub-morphemic elements was analysed, which led to the classification of abbreviations into those fully or partially motivated.

4 The Corpus of Abbreviations

The abbreviations analysed in this article were taken from different dictionaries of general linguistics and dictionaries of various linguistic disciplines (e.g. phonetics and phonology, lexicography, etc.; cf. Sources). The corpus comprises 446 abbreviations, belonging either to the category of simple or complex abbreviations. There are 270 alphabetisms, 67 acronyms, 5 blends, 19 clippings, 22 simple abbreviations and 63 hybrid forms. Alphabetisms, simple abbreviations and clippings were mainly formed according to the criterion of narrower sense formations, while acronyms, blends and hybrid formations were mainly formed according to the criterion of broader sense (cf. Fabijanić 2014, in print).

The corpus provides additional information about each abbreviation in the following order: abbreviation, expansion, descriptor of abbreviation form, source, abbreviation type, and details of the analysis for lexicalization, institutionalization, semantics, semiotics, and motivation.

5 Research Methods

The explication of research methods in the article refers to the stages of analysis for the cognitive aspect of the Multi-level approach. They will be dealing with the description of methods applied for the analysis of semantics, semiotics, motivation, lexicalization, and institutionalization, i.e. the features of the mentioned subsidiary aspects which can be attributed to abbreviations.

The application of semantic aspect is understood through the possibility of assigning a specific abbreviation type to its semantic field, i.e. the (sub-)field of linguistics or the interdisciplinary disciplines. The practice of assigning a semantic field does not certainly mean that an abbreviation could have only been appointed to that specific field; on the contrary, each abbreviation can be assigned to other semantic fields as well, but what we wanted to point out by this practice is the immediate textual and contextual surrounding an abbreviation was found in. Semiotic aspect is realised by determining the relations between the elements of an abbreviation or a sign. Due to the fact that two elements of the sign – the sign itself and the object – are already present in the form of abbreviations and their expansions, I find the interpretant (the sense/meaning) to be the element which can and has to be analysed by the implementation of Peirce's three-graded diversification of sign's clarity or understanding, i.e. by implementing the grade of the immediate interpretant (sign's first meaning), the grade of the dynamic interpretant (the actual effect of the sign) or the grade of the final interpretant (the final interpretative result of the sign).

Motivation in the formation of abbreviations was analysed by taking two conceptual perspectives in consideration. The first concept is connected to the narrower and the broader senses of the formation, while the second one is connected to the classification of abbreviations according to various motivational patterns (e.g. homophony, homography, homonymy, metaphor).

Lexicalization in this work is understood within the confines of the synchronic sense in which the lexicalization of abbreviations corresponds to *the process of listing* and the *listedness* (cf. Hohenhaus 2005: 356). For the purpose of the analysis, abbreviations are classified into those that were or were not listed or lexicalized. Being specific in their formation, listing/listedness of abbreviations will inevitably be sub-classified according to their specific structure and modes of production. Therefore, abbreviations like simple abbreviations and clippings, due to their graphic and spoken arbitrariness, will not be listed (or can be considered to be in the process of listing), while alphabetisms, acronyms, blends, and hybrid formations will be listed. Finally, institutionalization "[...] refers to the stage in the life of a word at (or form) [brackets in original] the transitional point between the status of ex-nonce-formation-turned-neologism and that of generally available vocabulary item, i.e. a formation that is listed but not (necessarily) [brackets in original] lexicalized in the diachronic sense yet [...]" (cf. Hohenhaus 2005: 359). Institutionalization, in terms of this part of the research, refers to the fact whether an abbreviation as such can be considered as institutionalized within the lexicon of linguistics or not.

6 Cognitive Aspects: The Analysis

6.1 Semantic and Semiotic Aspect

As described in the previous section, the semantic aspect of abbreviations was realised in accordance with the nearest corresponding semantic field. The analysed abbreviations were allocated to various sciences and disciplines, both linguistic or non-linguistic ones, according to the principle of immediate context within the entries. Here are the semantic fields with the information on total number of allocated abbreviations and an example with its expansion: *Applied linguistics* (39; ASTP - 'Army Specialized Training Program'), *Cognitive linguistics* (1; ICM – 'Idealized Cognitive Model'), *Computational linguistics* (38; COBOL – 'COmmon Business-Oriented Language'), *Corpus linguistics* (9; BNC – 'British National Corpus'), *Neurolinguistics* (6; TDH – 'Trace-Deletion Hypothesis'), *Historical linguistics* (8; PIE – 'Proto Indo-European'), *Linguistic anthropology* (2; LISA – 'Language and Identity in Sociocultural Anthropology'), *Psycholinguistics* (12; PALPA – 'Psycholinguistic Assessment of Language Processing in Aphasia'), *Sociolinguistics* (19; LSPt – 'Language Status Politics'), *Pragmatics* (71; AP – 'Applying Pragmatics'), *Theoretical approaches* (4; TG – 'Transformational grammar'), *Phonetics and phonology* (16; VOT – 'Voice Onset Time'), *Morphology* (27; SG – 'SinGular'), *Syntax* (56; NP – 'Noun Phrase'), *Semantics* (13; SFH – 'Semantic Feature Hypothesis'), *Dialectology* (1; ADS – 'American Dialect Society'), *Stylistics* (1; DS – 'Direct Speech'), *Lexicology* (1; ALLEX – 'African Languages LEXical project'), *Lexicography* (25; OED – 'Oxford English Dictionary'), *Linguistic typology* (2; ASL – 'American Sign Language'), *Text analysis* (2; PISA – 'Procedures for Incremental Structure Analysis'), *Discourse analysis* (1; SA – 'Speech Acts'), *Literary* *linguistic analysis* (1; NI – 'Internal Narration'), *Translational studies* (1; TT-ST – 'Target Text–Source Text'), *Philosophy of language* (1; CF – 'Context-Free'), *Semiotics* (1; ISL – 'Iconicity in Sign Language'), *Cognitive pragmatics* (1; ICM – 'Idealized Cognitive Model'), *Cognitive technology* (1; CT/TC – 'Cognitive Technology/ Technological Cognition'), *Speech recognition* (1; HMM – 'Hidden Markov Model'), *Speech technology* (2; AVIOS – 'American Voice Input and/Output Society'), *Media* (2; BBC English – 'British Broadcasting Company English'), *Computational science* (9; WOZ – 'Wizard-of-OZ simulation'), *Medicine* (1; TBI – 'Traumatic Brain Injured patient'), *Literacy* (2; NLS – 'New Literacy Study'), *Communicology* (6; SAT – 'Speech Accommodation Theory'), *Education* (25; CABE – 'Central Advisory Board of Education'), *Associations* (20; EURALEX – 'European association for LEXicography'), *Organisations* (12; CALLSSA – 'Center for Applied Language and Literacy Studies and Services in Africa'), *Conferences* (1; LTRC – 'Language Testing Research Colloquium'), *Journals* (5; IJOAL – 'International Journal Of Applied Linguistics'), and *Databases* (1; LAPTOC – 'Latin American Periodical Table Of Contents').

Most of the abbreviations were allocated for the following fields: *Pragmatics* (71), *Syntax* (56), *Applied Linguistics* (39), *Computational linguistics* (38), *Morphology* (27), *Education* (25), and *Lexicography* (25), while the least allocated, i.e. a field with one example, are: *Dialectology*, *Stylistics*, *Lexicology*, *Discourse analysis*, *Literary analysis*, *Semiotics*, *Cognitive Pragmatics*, *Cognitive technology*, *Speech recognition*, *Medicine*, *Conferences*, and *Databases*. A cross-sectional view of abbreviation types in some semantic fields will disclose the following data about the most frequent abbreviation type and its formation structure: the most frequent type of abbreviations in the fields of *Pragmatics*, *Applied linguistics*, *Syntax* and *Computational linguistics* is the type of alphabetisms with the narrower sense formation structure of *LLL*.

Semiotic aspect refers to the analysis of three grades of interpretants: immediate, dynamic and final. Most of the abbreviations were classified within the class of sign having an immediate interpretant (approx. 230 abbreviations), followed by the class of final interpretants (approx. 120), and the ones with the dynamic interpretant (approx. 90). I believe that in case of abbreviations, the classification of interpretants is firmly connected to the cognition of relationship between the abbreviation type(s), its/ their expansion(s), variability of expansion, some inner features of different abbreviation types (e.g. those of acronyms' when compared to alphabetisms), and frequency of use. The immediate interpretant indicates "[...] the effect the sign first produces or may produce upon a mind without any reflection upon it" (cf. Semiotics and Significs 1909: 110-1). From the previous quotation, it might be possible to assume that primary effects on our understanding of some abbreviations can be considered *sui generis*. Such is the case for some alphabetisms, clippings and hybrid forms whose understanding is conditioned by their immediate (con-)textual surrounding or expansion, e.g. primary understanding of the alphabetism AAAL ('American Association for Applied Linguistics') is conditioned by its expansion provided in the text. Furthermore, clippings like ACC ('ACCusative'), ACT ('ACTive'), COP ('COPula'), or hybrid formations like, ALLEX ('American Association for Applied Linguistics'), BSAfE ('Black South African English'), CCSARP ('Cross-Cultural Speech Act Realization Project'), LCPt ('Language Corpus Politics'), LRs ('Language Rights'), ATN grammar ('Augmented Transition Network grammar') is by all means conditioned by that primary effect of understanding. Direct evidence of this claim is supported in the following example: the difference in understanding the ALLEX and EURALEX, although some extensional as well as structural and formational features are being shared, is evident in the finality of understanding of the latter hybrid (connotation for *EUR-* is more immediate than for *ALL-*).

With regard to the dynamic interpretant, understanding of abbreviations within this category is mostly conditioned by the variability of extensions or the form of an abbreviation. The dynamic interpretant "[...] is that which is experienced in each act of Interpretation and is different in each from that of the other [...]" (cf. Semiotics and Significs 1909: 110-1). I believe that this can be witnessed in the following examples of simple abbreviations, acronyms, alphabetisms and clippings: M for 'Movement' or 'Metapragmatic joker'; ACE for 'Automatic Content Extraction' or 'Australian Corpus of English'; CT for 'Cognitive Technology', 'Conversational Theory' or 'Centering Theory'; AUX or Aux for 'Auxiliary'.

The final interpretant is "[...] the one Interpretative result to which every Interpreter is destined to come if the Sign is sufficiently considered [...]" or "[...] the effect the Sign *would* [italics in original] produce upon any mind upon which the circumstances should permit it to work out its full effect [...]" (cf. Semiotics and Significs 1909: 110-1). Due to their completeness of graphic and phonetic forms, i.e. the possibility of being read as words, and their frequency of use, most of the analysed blends, acronyms, and some hybrids, which are very similar to acronyms, together with some infrequent alphabetisms, were classified into the class of abbreviations having the final interpretant. The sum of the meanings or the final interpretative result the signs would inevitably have, can be confirmed by the examples of: acronyms – COBUILD ('COllins Birmingham University Information'), ECHO – ('European Commission Host Organisation'); blends – ('AFRIcan association for LEXicography'), FORTRAN ('FORmula TRANslation'); hybrids – AUSTRALEX ('AUSTRalian Association for LEXicography'), ITSPOKE ('Intelligent Tutoring SPOKEn dialogue system'); alphabetisms – HTML ('HyperText Mark-up Language'), L1 ('First Language').

6.2 Motivational Aspect

As it has already been explained (cf. § 5), motivational aspect was analysed through two conceptual perspectives. The first concept takes into consideration the formational difference between various types of abbreviations, previously classified according to the aspects of narrower and broader sense:

> The narrower sense of their creation refers to those formed by using initial letters of each element in the expansion (mainly alphabetisms) [brackets in original], and pronounced either by individual names of letters or as a word. The broader sense implies the ways and processes of formation, more or less different from the orthographic norms (mainly hybrid forms, acronyms, blends and clippings featuring some orthographic changes) [brackets in original], in consequence of which, one or more initials are used for various smaller elements of the expansion (smaller than words, yet bigger than initials) [brackets in original]. (Fabijanić 2014; in print)

The motivational aspect for the abbreviations analysed in this work assumes that some of them are fully motivated, while the others are partially motivated. I find fully motivated abbreviations to be the ones which largely correspond to the norms of narrower sense creations, i.e. alphabetisms and acronyms, simple abbreviations, blends and clippings formed by the orthographic norm and in which every element of the expansion is traceable. Partially motivated are those that principally fall into the group of broader sense formations, i.e. alphabetisms, acronyms, simple abbreviations, clippings, blends, and hybrid formations which are not formed by the orthographic norms and in which more or less elements of expansions can be traced. The second conceptual perspective of the motivational aspect takes into consideration the specific patterns which motivated the emergence of abbreviations. These might be homophonous, homographic, homonymic, and metaphorical patterns. I shall provide some examples of different abbreviation types for each pattern. The homographic pattern, in which initals from extensions are repeated in abbreviations, is mostly evident in narrowly formed alphabetisms, e.g. IPA – 'International Phonetic Association', NLU – 'National Lexicographic Unit'. The homophonic pattern, in which part(s) of extensions or initials (in acronyms) are either echoed in a resultant abbreviation or make a word having different meaning in the general lexicon, can be detected in the example of clippings ACT ('ACTive'), INACT ('INACTive'), in the acronym ACE ('Automatic Content Extraction'), or the blend AFRILEX ('AFRIcan association for LEXicography). Sometimes the overlapping of patterns can be realized as in the example of BANA acronym ('Britain, Australasia, and North America') in which homographic and homophonic principles can be traced (the repetition of initials and homophony with other words and abbreviations like the surname *Bana*, a drink named *BANa* or *BANA* for '*British Acoustic Neuroma Association', 'Bulimia Anorexia Nervosa Association', 'Bath Area Network for Artists', etc.).* The homonymic principle is realised in examples of hybrids and acronyms like BIT – 'BInary digiT' (homonymic with *bit* 'amount of sth, part of sth') and CAM – 'Center of Auditory Memories' (homonymic with *CAM* 'Computer Aided Manufacturing', *cam* 'a wheel part which changes the movement of the wheel', *cam* 'a clipped form from *camera*'). *The homonymic principle can be disrupted by the addition or deletion of graphemes otherwise not found in original words, e.g. the acronym KWIC ('KeyWord In Context') in which <KWI> suggests the group of graphemes <qui>, while <C> suggests the group <ck> as in quick, or in case of the hybrid CHILDES database ('*CHIld Language Data Exchange System database') in which its form might suggest the ungrammatical plural form of the noun *child*. Metaphorical principle of motivation is evident in subsequent forms of acronyms or hybrid forms: NORM – 'Non-mobile, Older, Rural Male'; MACK – Multimodal Autonomous Conversational Kiosk'; CHAT – 'Cultural-Historical Activity Theory'; BASIC English – 'British, American, Scientific, International, Commercial English', while humorous touch is felt in metaphorically motivated FUG ('Functional Unification Grammar'), MUD ('Multi-User Domain'), DARE ('Dictionary of American Regional English'), LISA ('Language and Identity in Sociocultural Anthropology'), ELI ('English Language Institute'), PISA ('Procedures for Incremental Structure Analysis'), etc.

6.3 The Aspects of Lexicalization and Institutionalization

For the purpose of this research, Lipka's definitions on lexicalization and institutionalization of complex lexemes will be applied (2005: 4). According to Lipka, "[...] complex words [...] were coined according to productive morphological or semantic process [...], and they have [...] been affected – to a greater or lesser degree - by formal and/or semantic changes subsumed under the concepts of **lexicalization** and **institutionalization** [bold in original]." He defines lexicalization as: "[...] the process by which complex lexemes tend to become a single unit with a specific content, through frequent use. In this process, they lose their nature as a syntagma, or combination [of smaller units], to a greater or lesser extent" (1992: 107). In his later research (2005: 7), the definition of lexicalization was extended by the features of gradual, historical process which involve changes in phonology and semantics, as well as loss of motivation.

Since all the above definitions with some additional descriptions in Lipka's work (together with other work of specialists cited in his works), fit well to the topic of my research and since the lexicon to which the two processes can be applied is compatible with their patterns of verification, I shall propose a three-stage model of lexicalization for abbreviations, i.e. preliminary, primary and secondary stage. The *preliminary stage* is understood as a preparatory stage in the process of lexicalization. It refers to a small group of simple abbreviations whose final abbreviated form is too short or arbitrary to be considered either partially or fully lexicalized, e.g. A – 'Adjective', L – 'Location', M – 'Metapragmatic joker' or 'Movement', N – 'Noun'. The *primary stage*, with partially lexicalized abbreviations, is disclosed in formational incompleteness and variability of abbreviations. Thus we have alphabetisms with both small and capital letters (CmC – 'Computer mediated Communication'), clippings with small and capital letters, which, additionally, are not pronounced (Aux – 'Auxiliary', Utt – 'Utterance'), hybrid formations with small and capital letters (LHRs – 'Linguistic Human Rights', SaPs – 'Speech acts Projections'), and alphabetisms with various diacritics (R-A – 'Referentially Autonomous expression', CT/TC – 'Cognitive Technology/Technological Cognition'). The *secondary stage* with completely lexicalized abbreviations refers to those which are more easily recognized as lexical units, i.e. acronyms (LAD – 'Language Acquisition Device', LAPTOC – 'Latin American Periodical Table Of Contents'), hybrids in combinations with words (LISP language – 'LISt Processing language'), alphabetisms in narrower sense (MHG – 'Middle High German'), blends (AFRILEX – 'AFRIcan association for LEXicography), some clippings with consistent and utterable orthography (COP – 'COPula', FEM – 'FEMinine').

Institutionalization, in Lipka's view, refers to "[...] the sociolinguistic aspect of this process and can be defined as the integration of a lexical item, with a particular form and meaning, into the existing stock of words as a generally acceptable and current lexeme" (2005: 8). Lipka further defines institutionalization as the process in which a specific speech community (e.g. doctors, medical people, linguists, etc.) accepts the specific lexemes into the lexicon. (2005:11). He also states that "[both lexicalized and] institutionalized words, ie item-familiar ones, are registered and listed in good dictionaries [...]"

(2005: 12). If we take into consideration all the above viewpoints and aspects of institutionalization, as well as the facts about our corpus confirmed and compiled from specialized (mostly encyclopaedic) dictionaries, than we can conclude that all the analysed examples of abbreviations have been fully institutionalized. For the purposes of defining the entry structure, institutionalization will be explained by two attributes – affirmed or not affirmed institutionalization.

7 The Entry and its Elements

As it had previously been suggested, after having completed the analysis of *Structure and modes of production aspect* (cf. Fabijanić 2014, in print), an entry would consist of the following elements: a headword, (a) variant(s), pronunciation, the type of word-formation, the information on the word class (where applicable), a descriptor which will inform users about the mode of production, expansion elements in English and their translation in Croatian, the information about the medium and the origin of abbreviations, both in English and Croatian.

The following examples of different abbreviation forms (an alphabetism, an acronym, a clipping, a blend and a hybrid form) present the micro-structure of the future dictionary entry with the addition of elements for the *Cognitive aspect*, i.e. information on the semantic field, interpretation of the sign's interpretant (immediate, dynamic or final), information on the motivational aspect (fully or partially motivated), the information on the lexicalization (preliminary, partial and complete lexicalization) and institutionalization (institutionalized or not). The subsequent (simple) abbreviations have been suggested for the use within the entry: *T* (type), *E* (expansion), *M* (medium), *D* (descriptor), *O* (origin), *SF* (semantic field), *SI* (sign's interpretant), *MT* (motivation), *L* (lexicalization), and *I* (institutionalization).

- **ADS** [ˈeɪˈdiːˈes] **T**: *alph.*| **E**: *American Dialect Society*/Američko dijalektalno društvo | **M**: written/ pisani, spoken/govorni | **D**: *LLL* | **O**: *ADS* was founded in 1889 with the intention of creating a dictionary of American dialects. (ELL)/Američko dijalektalno društvo osnovano 1889. s namjerom stvaranja rječnika američkih narječja. | **SF**: dialectology/dijalektologija | **SI**: immediate/ neposredan | **MT**: full/potpuna | **L**: complete/dovršena | **I**: affirmed/potvrđena

- **ACE** [eɪs] **T**: *acr.* | **E**: 1) *Automatic Content Extraction*/Automatsko ekstrahiranje sadržaja, 2) *Australian Corpus of English*/ Korpus australskoga engleskog | **M**: written/pisani, spoken/govorni | **D**: 1) *LLL*, 2) *LLL E = prep.*; **O**: 1) The *ACE* program is a successor to ® *MUC* that has been running since a pilot study in 1999. (ELL)/*ACE* program provodi se još od pilot-projekta iz 1999., a naslijedio je MUC., 2) The corpus of Australian English compiled at Macquarie University using texts published in 1986. (HEL) / Korpus australskoga engleskog sačinjen na Macquarie sveučilištu iz tekstova objavljenih 1986. | **SF**: computational and corpus linguistics/ računalna i korpusna lingvistika | **SI**: dynamic/ dinamičan | **MT**: 1) full/ potpuna, 2) partial/djelomična | **L**: complete/dovršena | **I**: affirmed/potvrđena

- **AUX, Aux, aux** [-] **T**: *clip.* | **E**: *Auxiliary*/pomoćni | **M**: written/pisani | **D**: *Lll* | **O**: *Lat.* auxiliaris - 'giving aid' (RDLL)/*lat.* pomoćni - 'koji pomaže'. | **SF**: morphology/morfologija | **SI**: dynamic/dinamičan | **MT**: partial/djelomična | **L**: partial/ djelomična | **I**: affirmed/ potvrđena

- **AFRILEX** [ˈæfrɪˌleks] **T**: *blend* | **E**: *AFRIcan association for LEXicography*/Afričko leksikografsko udruženje | **M**: written/ pisani, spoken/govorni | **D**: *LSF E= noun, prep.* | **O**: *AFRILEX* was founded in 1995 and strives to promote all aspects of lexicography on the African continent. (ELL)/Organizacija osnovana 1995. u svrhu promocije svih aspekata leksikografije na afričkom kontinentu. | **SF**: associations, lexicography/udruženja, leksikografija | **SI**: dynamic/dinamičan | **MT**: full/ potpuna | **L**: complete/dovršena | **I**: affirmed/ potvrđena

- **ALGOL** [ˈælgɒl] **T**: *hybr. (syll+s.abb.)* | **E**: *ALGOrithmic Language*/algoritamski jezik | **M**: written/pisani, spoken/ govorni | **D**: *SSL* | **O**: Programming computer language appeared in 1958. (RDLL)/ Programski računalni jezik nastao 1958. godine. **I**: affirmed/ potvrđena | **SF**: computer science/računalne znanosti | **SI**: final/konačan | **MT**: partial/djelomična | **L**: complete/ dovršena | **I**: affirmed/ potvrđena

8 Conclusion

In closing, I would like to stress again the overall and immediate aims of this research. The overall aim is to provide the basis for the future dictionary of abbreviations in linguistics, which would be bilingual, bidirectional, specialized (domain-specific, technical), synchronic, explanatory, alphabetically arranged dictionary, informative and encyclopaedic in content, and serve both non-specialized and specialized audience. The solution for the lexicographic presentation of abbreviations is based on Ingrid Fandrych's *Multi-level approach* which is comprised of three aspects: *Structure and Modes of Production*, *Cognitive Aspect*, and *Functional Aspect*. The immediate aim of the research is to bring forth the results of analysis for the second aspect, i.e. the Cognitive aspect, which deals with semantics, semiotics, motivation, lexicalization and institutionalization.

Summing up the results of this research, I would like to state that the sources used in compiling the corpus of abbreviations in linguistics, have proved to be trustworthy, valuable and fundamental for the cognitive aspect, just as they were for the structures and modes of production. Furthermore, the research has proved that the criteria of narrower and broader sense classification can be reiterated for the elements of the cognitive aspect too.

I hope to have shown that the stages of the cognitive aspect can be defined by the recommended methods in analysing semantic, semiotic, motivational features of abbreviations, as well as the facts about their lexicalization and institutionalization. The semantic aspect was realised in accordance with the nearest corresponding semantic field for a specific abbreviation. The abbreviations were allocated to various sciences and disciplines, such as *Pragmatics, Syntax, Applied Linguistics, Computational linguistics*, etc. The analysis of the semiotic aspect was realised by the application of three grades of interpretants: immediate (sign's first meaning), dynamic (the actual effect of the sign) and final (the final interpretative result of the sign). Most of the abbreviations were classified within the class of sign having an immediate interpretant, followed by the class of final interpretants, and the class of the dynamic interpretant. The motivational aspect for the abbreviations analysed in this work assumes that some of them are fully motivated (those which largely correspond to the norms of narrower sense creations), while others are partially motivated (those that principally fall into the group of broader sense formations). The second conceptual perspective of the motivational aspect takes into consideration the homophonous, homographic, homonymic, and metaphorical patterns which motivated the emergence of abbreviations. As far as lexicalization is concerned, a three-stage model of lexicalization for abbreviations (preliminary, primary and secondary stage) is proposed. The preliminary stage is a preparatory stage in the process of lexicalization. The *primary stage*, with partially lexicalized abbreviations, is disclosed in formational incompleteness and variability of abbreviations. The *secondary stage* with completely lexicalized abbreviations refers to those which are more easily recognized as lexical units. In dealing with the aspect of institutionalization, the crucial moment for the application of dichotomous differentiation between affirmed and not affirmed institutionalization, is recognized through the fact that abbreviations have been lexicographically attested.

Finally, with regard to the micro-structure of the future entry, its second level will consist of five structural elements which will be represented by the following abbreviations/symbols: *SF* (semantic field), *SI* (sign's interpretant), *MT* (motivation), *L* (lexicalization), and *I* (institutionalization).

9 References

Fabijanić, I., F. Malenica (2013). "Abbreviations in English medical terminology and their adaptation to Croatian". In *JAHR*, Vol. 4, No. 7., Faculty of Medicine, Rijeka.

Fabijanić, I. (2014). "Dictionary of abbreviations in linguistics: towards a bilingual, specialized, single-field, explanatory dictionary". In Proceedings of the conference *Planning Non-existent dictionaries: Lexicographic Typologies*, Lisbon. In print.

Fandrych, I. (2004). Non-Morphematic Word-Formation Processes: A Multi-Level Approach to Acronyms, Blends, Clippings and Onomatopoeia. Unpublished PhD Thesis. University of the Free State: Bloemfontein.

Fandrych, I. (2008a). "Pagad, Chillax and Jozi: A Multi-Level Approach to Acronyms, Blends, and Clippings". In Nawa Journal of Language and Communication. Vol.2, No. 2.

Fandrych, I. (2008b). "Submorphemic elements in the formation of acronyms, blends and clippings", In Lexis 2: "Lexical
Submorphemics / La submorphémique lexicale".

Hohenhaus, P. (2005). "Lexicalization and Institutionalization". In Handbook of Word-Formation (Eds. Štekauer, P. and
R. Lieber), Springer.

Lipka, L. (1992). "Lexicalization and Institutionalization in English and German". In *Linguistica Pragensia*, Issue 1/ 1992, Faculty of Arts, Charles University of Parague, Prague.

Lipka, L. (2002). English lexicology, Lexical Structure, word semantics and word-formation. Tuebingen: Gunter Narr.

Lipka, L., S. Handl, W. Falkner (2004). "Lexicalization & Institutionalization: The State of the Art in 2004". In *SKASE Journal of Theoretical Linguistics*, Vol. 1/No. 1., University in Košice.

Lipka, L. (2005). "Lexicalizatuion and Institutionalization: revisited and extended". In *SKASE Journal of Theoretical Linguistics*, Vol. 2/No. 2, University in Košice.

López Rúa, P. (2004). "Acronyms & Co.: A typology of typologies", In *Estudios Ingleses de la Universidad Complutense*. Vol. 12, pp. 109-129, Madrid.

López Rúa, P. (2006). "Non-Morphological Word Formation". In *Encyclopedia of Language and Linguistics(2nd Edition).* Vol. 2., Elsevier: Oxford.

Malenica, F., I. Fabijanić (2013). "Abbreviations in English Military Terminology". In *Brno Studies in English*. Vol. 39, No. 1., Faculty of Arts, Masaryk University, Brno.

The Commens Dictionary of Peirce's Terms. Accessed at http://www.helsinki.fi/science/commens/terms [30/03/2014].

Semiotic and Significs: The Correspondence Between Charles S. Peirce and Victoria Lady Welby. Ed. by Charles S.
Hardwick & J. Cook (1977). Bloomington: Indiana University Press. Accessed at http:// www.helsinki.fi/science/commens [30/03/2014].

10 Sources

Filipović, R. (1990). Anglicizmi u hrvatskom ili srpskom jeziku: porijeklo – značenje – razvoj. Školska knjiga: Zagreb.

Anglicisms in European Languages. Manfred Goerlach (ed.). Oxford University Press: Oxford. 2005.

Concise Encyclopedia of Pragmatics. Jacob L. Mey (ed.). Elsevier: Oxford. 2009.

Hartmann, R. R. K., Gregory James (1998). *Dictionary of Lexicography*. Routledge: London

Encyclopedia of Language and Linguistics, 2nd edition. Keith Brown (ed.). Elsevier: Oxford. 2004.

Roach, Peter (2009) *English Phonetics and Phonology, Glossary: A Little Encyclopaedia of Phonetics*. (http://www.cambridge.org/elt/peterroach)

Kristal, Dejvid (1985). *Enciklopedijski rečnik moderne lingvistike*. Nolit: Beograd.

The Handbook of English Linguistics. Aarts, B., April McMahon (eds.). Blackwell Publishing: London. 2006.

Bussman, Hadumod (1998). Routledge Dictionary of Language and Linguistics. Routledge: London.

Trask, Robert Lawrence (2005). *Temeljni lingvistički pojmovi*. Školska knjiga: Zagreb.

La definizione delle relazioni intra- e interlinguistiche nella costruzione dell'ontologia IMAGACT

Gloria Gagliardi
Università degli Studi di Firenze
gloria.gagliardi@unifi.it

Abstract

IMAGACT è un'ontologia interlinguistica che rende esplicito il *range* di variazione pragmatica associata ai predicati azionali a media ed alta frequenza in italiano ed inglese. Le classi di azione che rappresentano le entità di riferimento dei concetti linguistici, indotte da *corpora* di parlato da linguisti madrelingua, sono rappresentate in tale risorsa lessicale nella forma di scene prototipiche (Rosch 1978). Tale metodologia sfrutta la capacità dell'utente di trovare somiglianze tra immagini diverse indipendentemente dal linguaggio, sostituendo alla tradizionale definizione semantica, spesso sottodeterminata e linguo-specifica, il riconoscimento e l'identificazione dei tipi azionali. L'articolo illustra i criteri generali che hanno ispirato il *mapping* inter-/intra- linguistico dei dati derivati da *corpora* per la formazione dell'ontologia, le questioni di natura teorica e tecnica poste dalla costruzione della risorsa e le soluzioni adottate. Vengono descritte le tipologie e la natura delle relazioni tra le entità del database nella sua versione 1.0, e le modalità generali con cui i materiali linguistici annotati sono stati organizzati in una struttura dati coerente.

Keywords: ontologia; verbi di azione; relazioni interlinguistiche

1 Introduzione

I verbi d'azione veicolano informazioni essenziali per la corretta interpretazione delle frasi e quindi per la comprensione del linguaggio. Le relazioni che strutturano questa parte di lessico sono tuttavia molto complesse, in quanto non è possibile stabilire una corrispondenza biunivoca tra predicati ed eventi (Moneglia & Panunzi 2010). I verbi d'azione più frequenti nella comunicazione quotidiana sono infatti "generali", ovvero possono essere applicati in modo produttivo a classi di azioni pragmaticamente e cognitivamente diverse, come mostra la variazione pragmatica del verbo italiano.

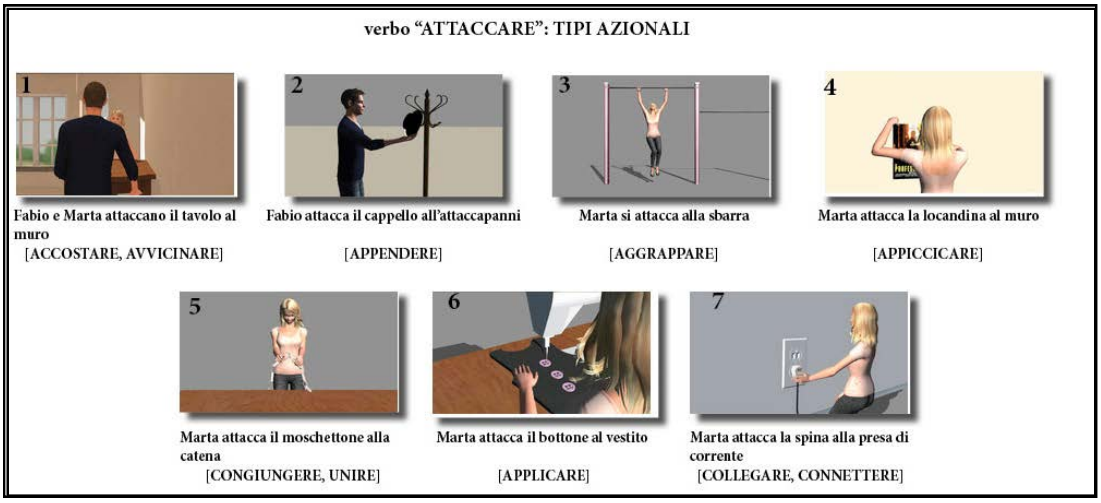

verbo "ATTACCARE": TIPI AZIONALI
1
Fabio e Marta attaccano il tavolo al muro
[ACCOSTARE, AVVICINARE]
2
Fabio attacca il cappello all'attaccapanni
[APPENDERE]
3
Marta si attacca alla sbarra
[AGGRAPPARE]
4
Marta attacca la locandina al muro
[APPICCICARE]
5
Marta attacca il moschettone alla catena
[CONGIUNGERE, UNIRE]
6
Marta attacca il bottone al vestito
[APPLICARE]
7
Marta attacca la spina alla presa di corrente
[COLLEGARE, CONNETTERE]

Fig. 1: Variazione pragmatica del lemma italiano *attaccare*.

Tale variazione corrisponde alla competenza semantica referenziale dei parlanti, ed è quindi un dato essenziale per la modellizzazione dell'informazione lessicale. Tuttavia, essa è solo saltuariamente censita dai dizionari tradizionali e dalle più note ontologie e risorse computazionali, come ad esempio Wordnet (Fellbaum 1998) e Verbnet (Kipper-Schuler 2005).

Il progetto IMAGACT contribuisce a superare questa lacuna attraverso la realizzazione di un'ontologia interlinguistica dell'azione che esplicita lo spettro di variazione pragmatica associata ai predicati in italiano e in inglese (Moneglia *et al.* 2012). Le classi di azioni che identificano il riferimento di ogni verbo sono state individuate a partire dall'annotazione di grandi *corpora* rappresentativi dell'uso linguistico parlato spontaneo, e quindi associate attraverso una procedura di *mapping* inter-/intra- linguistico ad una serie di scene prototipiche, in grado di elicitare nell'utente la comprensione della classe di eventi rappresentata (Rosch 1978).

La metodologia di induzione delle classi di azioni e l'utilizzo di prototipi in sostituzione delle definizioni per la rappresentazione del riferimento, due tra gli aspetti più innovativi di IMAGACT, hanno però sollevato alcune questioni di strutturazione dell'informazione nella fase di formazione dell'ontologia (par. 3.1).

Verranno descritti, a partire da alcuni *case study*, i problemi e le soluzioni adottate per la costruzione della risorsa, ovvero le modalità secondo le quali il materiale annotato è stato organizzato all'interno di una struttura dati coerente, che, pur mantenendosi aderente all'intuizione dei parlanti madrelingua, risulti di facile consultazione per l'utente finale.

2 Induzione delle Classi Azionali da Corpus

Una strategia efficace per apprezzare l'uso dei verbi *action-oriented* è la diretta osservazione delle loro occorrenze nel parlato spontaneo, in cui il riferimento all'azione è decisivo. In IMAGACT è stata dunque adottata una procedura di tipo *bottom-up*; le classi di azioni sono state indotte da risorse linguistiche di parlato disponibili, su licenza, per scopi scientifici:

- *corpus* Inglese: una selezione del British National Corpus (BNC) di circa 2 milioni di parole;
- *corpus* Italiano: una collezione di risorse di parlato in lingua italiana (LABLITA corpus, LIP, CLIPS) per un totale di 1,6 milioni di parole (Moneglia *in press*; Gagliardi 2014).
- I materiali linguistici sono stati sottoposti ad una articolata procedura di annotazione (Moneglia *et al.* 2012; Frontini *et al.* 2012); i dati risultanti consistono di:
- due elenchi di verbi, uno per l'italiano e uno per l'inglese;
- per ogni verbo, una serie di "tipi" azionali, ovvero le classi di azioni fisiche tra loro tipologicamente e cognitivamente diverse che rientrano nell'estensione del predicato (fig. 1);
- per ogni tipo azionale, uno o più *Best Example*, cioè le istanze più rappresentative di tutte le strutture tematiche e delle proprietà aspettuali individuate;
- per ogni *Best Example*, l'insieme delle occorrenze che costituiscono la variazione del verbo nel *corpus*, standardizzate da linguisti madrelingua in frasi semplici ed annotate a vari livelli.

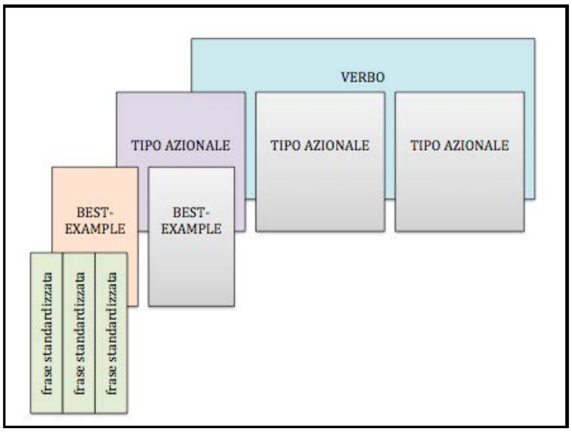


Fig. 2: Risultati della procedura di annotazione IMAGACT.

L'ontologia interlinguistica è stata costruita mediante il ricongiungimento dei tipi azionali in un'unica galleria di scene prototipali. Il requisito generale che ha ispirato la formazione della risorsa è che l'immagine standard associata a ciascun tipo sia facilmente riconoscibile e garantisca la corretta individuazione del concetto azionale, indipendentemente dalla lingua e dalla cultura di origine dell'utente.

3 Mapping

3.1 Criteri generali

La costruzione di un'ontologia coerente dal punto di vista linguistico e formale a partire dai materiali estratti da *corpora* ha rappresentato una sfida molto impegnativa, sia a livello pratico, considerata l'enorme mole di dati da riconciliare in una struttura unitaria, che a livello teorico, data la novità della metodologia di induzione delle classi azionali.
I dati in *input* hanno influenzato fortemente la forma e la struttura del database e la concezione stessa della procedura di *mapping* inter-/intra- linguistico. La tipizzazione della variazione è infatti condizionata dalla semantica del verbo in oggetto: il senso del lemma, operando come "punto di vista" sulle categorie azionali, ha determinato la granularità dell'annotazione, anche a parità di eventi predicabili. Si considerino ad esempio le variazioni dei lemmi *attaccare* ed *appendere* riportate in fig. 3.

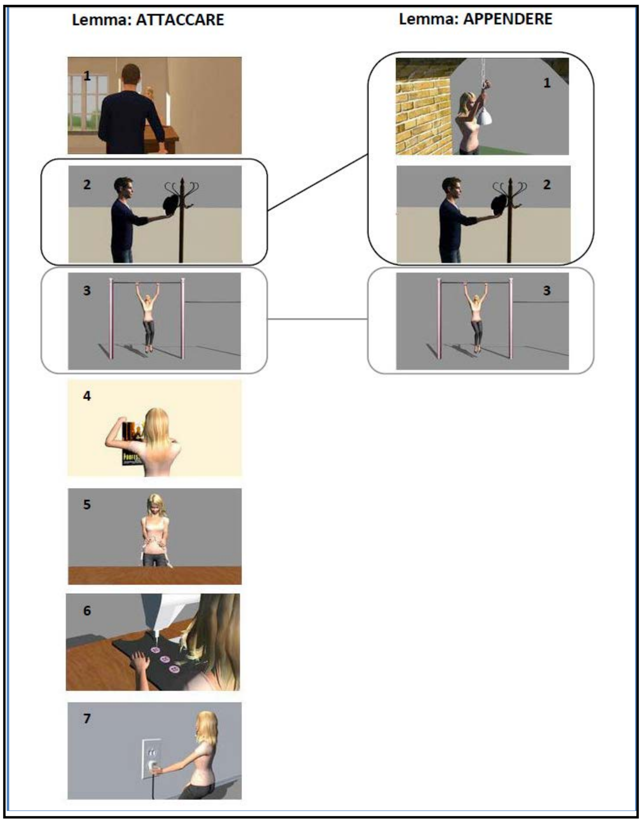


Fig. 3: Classi azionali dei lemmi attaccare e appendere a confronto.

I tipi azionali 1 e 2 del verbo *appendere* sono stati categorizzati come unico tipo (il 2) nel lemma *attaccare*: ciò significa che l'annotatore, tipizzando la variazione primaria del lemma *attaccare*, non ha ritenuto di dover distinguere gli eventi sulla base dello stato risultante del tema (il tema "pende dal riferimento" in 1, e non "pende" in 2); al contrario, il medesimo tratto è stato considerato rilevante per il lemma *appendere*. Eventi che costituiscono più classi azionali all'interno della variazione di un predicato più specifico sono stati insomma categorizzati come unica classe per predicati di maggiore generalità.

3.2 Ipotesi di lavoro

Per gestire casi come quello appena illustrato, nel corso della pianificazione della struttura del database sono state prese in considerazione due ipotesi di lavoro. Una prima soluzione prevede che la granularità dell'annotazione del lemma meno generale venga riprodotta nel lemma più generale. Riprendendo l'esempio in fig. 3, il tratto "sospensione" verrebbe considerato pertinente sia per il lemma *appendere* che per il lemma *attaccare*. Ciò avrebbe come conseguenza il fatto che il database ammetta un'unica tipologia di relazione, l'equivalenza. Ne risulterebbe un DB di notevole semplicità strutturale, in cui è sempre possibile stabilire relazioni 1:1 tra tipi azionali. Un *mapping* così concepito porterebbe però ad una anti-economica sovragenerazione di tipi azionali per i verbi generali.

La seconda soluzione prevede che nella struttura dei dati vengano introdotte relazioni implicite di tipo IS_A. La scelta avrebbe essenzialmente due conseguenze: il database creato conterrebbe gerarchie *implicite* di tipi, e uno stesso tipo potrebbe essere rappresentato nell'ontologia da più scene. A ciò corrisponderebbe un aumento della complessità delle relazioni nel DB; la soluzione, tuttavia, consentirebbe di mantenere contenuto il numero di tipi azionali e soprattutto di garantire l'aderenza della tipizzazione all'intuizione dei parlanti madrelingua.

In ragione della sua maggior coerenza rispetto ai requisiti progettuali, è stata scelta la seconda soluzione.

3.3 Relazioni del DB IMAGACT

Le entità del DB IMAGACT 1.0 sono organizzate mediante due tipologie di relazione:

- Relazione tipo-tipo;
- Relazione tipo-scena.

Nella prima categoria rientra la relazione L_EQ, "*local equivalence*". Nel quadro teorico adottato in IMAGACT (Moneglia 1997) viene definita "equiestensionalità" o "equivalenza locale" la possibilità per due (o più) predicati di applicarsi allo stesso evento o insieme di eventi, sulla base di proprietà di senso. Tale proprietà è rappresentata *indirettamente* nel database dall'appartenenza di una stessa scena alla variazione primaria di due o più lemmi (fig. 4).

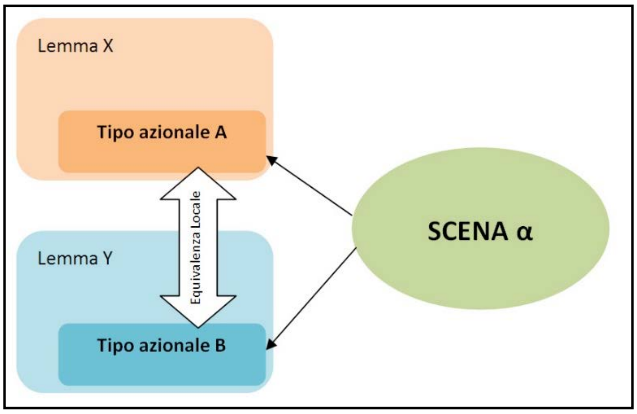


Fig. 4: Relazione di Equivalenza Locale (L_EQ) nel DB IMAGACT.

Dati i tipi azionali a, b, c ed i lemmi X e Y, la relazione ha le seguenti caratteristiche:

- $a \in X,\ b \in X \Rightarrow \neg L\_EQ\ (a, b)$
- $a\mathfrak{R}b \Rightarrow b\mathfrak{R}a$ (simmetria)
- $a\mathfrak{R}b,\ b\mathfrak{R}c \Rightarrow a\mathfrak{R}c$ (transitività)

Tipi azionali e scene vengono invece collegati in IMAGACT secondo due modalità (fig. 5):

- PRO, "*prototipo*": la scena è un prototipo per il tipo;
- INST, "*istanza*": la scena rappresenta una possibile realizzazione (non prototipica) della classe di eventi rappresentati in un tipo.

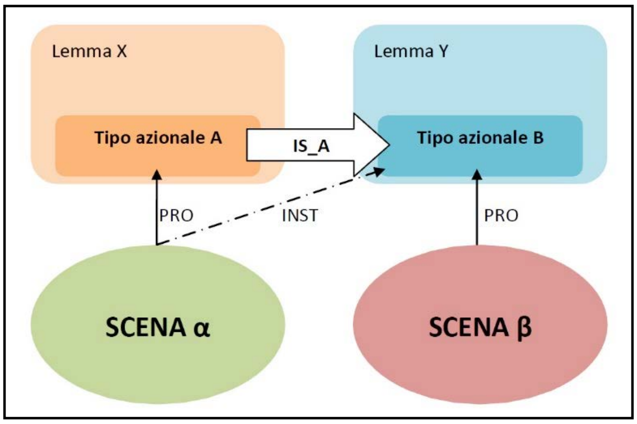


Fig. 5: Relazioni PRO e INST nel DB IMAGACT.

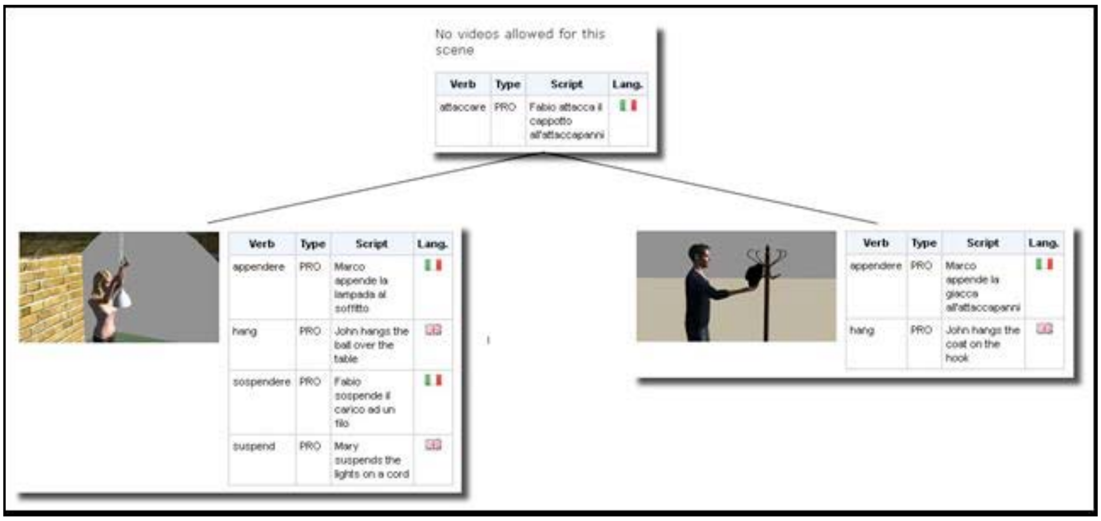
Ogni tipo azionale del database IMAGACT ha associata una, ed una sola, scena con relazione PRO; può invece avere, opzionalmente, più scene connesse con relazione INST.

Ciò corrisponde al fatto che un tipo di azione altamente prototipico per un verbo (es. "*attaccare/appendere* la lampada al soffitto" in relazione alla variazione primaria del lemma *appendere*), possa corrispondere ad una istanza periferica per un altro verbo (l'evento di "*attaccare/appendere* la lampada al soffitto" è una fra le possibili istanze per il lemma *attaccare*, al pari di "*attaccare/appendere* il cappello all'attaccapanni"). Il fenomeno, che non ha natura logica, è connesso alla maggiore o minore marcatezza pragmatica dell'evento e a fattori semantici ancora da investigare.

3.4 Il concetto di "Famiglia di Prototipi"

La scelta di una strategia gerarchizzante ha avuto come effetto l'introduzione in IMAGACT del concetto di "famiglia di prototipi": laddove siano presenti differenze di granularità di annotazione dovute al senso del lemma annotato e tali differenze appaiano consistenti e/o interessanti, classi di azioni distinte all'interno della variazione di un predicato specifico possono essere associate a costituire un unico *cluster* di prototipi in predicati più generali. Con la dicitura "famiglia di prototipi" si intende dunque in IMAGACT l'insieme delle scene connesse ad un predicato allo scopo di esplicitare differenziali linguistici.

In fig. 6 è mostrata la soluzione strutturale adottata per l'esempio discusso nei paragrafi precedenti. Dal punto di vista dell'architettura dell'ontologia, ciascuna scena corrisponde a un nodo della gerarchia.



No videos allowed for this scene

Verb	Type	Script	Lang.
attaccare	PRO	Fabio attacca il cappotto all'attaccapanni	🇮🇹

Verb	Type	Script	Lang.
appendere	PRO	Marco appende la lampada al soffitto	🇮🇹
hang	PRO	John hangs the ball over the table	🇬🇧
sospendere	PRO	Fabio sospende il carico ad un filo	🇮🇹
suspend	PRO	Mary suspends the lights on a cord	🇬🇧

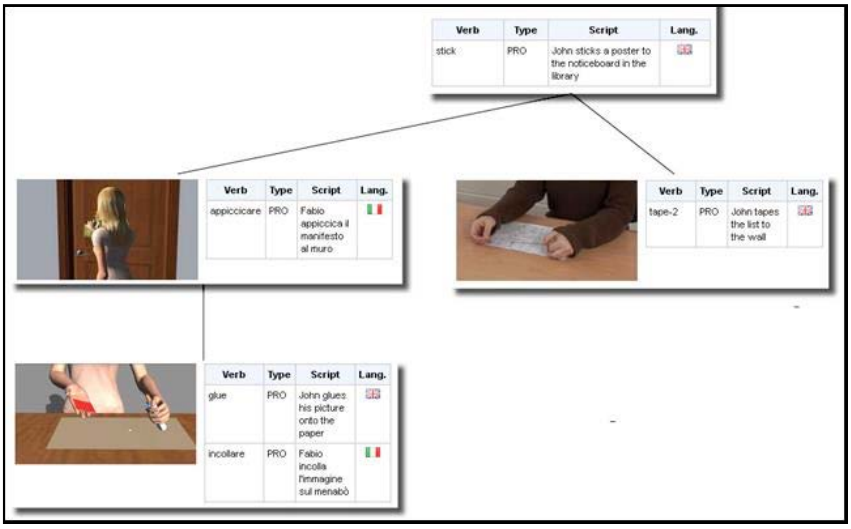
Verb	Type	Script	Lang.
appendere	PRO	Marco appende la giacca all'attaccapanni	🇮🇹
hang	PRO	John hangs the coat on the hook	🇬🇧

Fig. 6: Esempio di "Famiglia di Prototipi": *attaccare.*

3.5 Troponimi e denominali

Una struttura dati così concepita consente anche di gestire agevolmente varie tipologie di iponimi, e la loro relazione con i tipi azionali dei verbi generali (fig. 7). Tra questi:

- troponimi, ovvero iponimi che esplicitano la modalità con cui l'azione viene compiuta dall'agente (es. *appiccicare* vs. *attaccare*);
- denominali, ovvero iponimi che esplicitano uno specifico materiale o oggetto di cui l'agente si serve per realizzare l'azione (es. *incollare*, *to glue*, *to tape* vs. *attaccare*).



Verb	Type	Script	Lang.
stick	PRO	John sticks a poster to the noticeboard in the library	🇬🇧

Verb	Type	Script	Lang.
appiccicare	PRO	Fabio appiccica il manifesto al muro	🇮🇹

Verb	Type	Script	Lang.
tape-2	PRO	John tapes the list to the wall	🇬🇧

Verb	Type	Script	Lang.
glue	PRO	John glues his picture onto the paper	🇬🇧
incollare	PRO	Fabio incolla l'immagine sul menabò	🇮🇹

Fig. 7: Esempio di mapping di verbi denominali e troponimi: to stick, appiccicare, to glue, incollare, to tape.

4 Conclusioni

La versione 1.0 del database IMAGACT, rilasciata in data 1/09/2013, contiene 521 verbi ad alta e media frequenza per l'italiano e 550 per l'inglese, connessi ad una galleria di 1010 scene. La risorsa è interrogabile all'URL http://www.imagact.it/.

La metodologia illustrata ha permesso la generazione di tale ontologia al suo stadio attuale: le classi azionali individuate a partire dai lemmi delle due lingue sono state organizzate in una struttura dati coerente, conciliando la necessità di correttezza formale con la volontà di mantenere aderente la tipizzazione della variazione all'intuizione dei linguisti madrelingua che hanno prodotto l'annotazione. Per facilitare l'estensione della struttura dati ad altre lingue, il database è attualmente in fase di revisione e di semplificazione nell'ambito del progetto MODELACT ("From individuation to modelling in natural language ontology of action. Grounding the definition of action concepts on language infrastructures").

5 Bibliografia

Fellbaum, Ch. (1998). *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage comprehensive verb lexicon. PhD thesis, University of Pennsylvania, Philadelphia, US.

Gagliardi, G. (2014). Validazione dell'Ontologia dell'Azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI). PhD thesis, Università degli Studi di Firenze, Italia.

Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G. & Panunzi, A. (2012). Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In: M. Zock & R. Rapp, *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, CogALex III*, The COLING 2012 Organizing Committee, pp. 69-80.

Moneglia, M. (1997). Prototypical vs. not-prototypical verbal predicates: ways of understanding and the semantic types of lexical meanings. In: *Vestnik Moskovkogo Universitatea (Moscow State University Bullettin)*, 2, pp.157–173.

Moneglia, M. (*in press*). The Semantic variation of action verbs in multilingual Spontaneous speech Corpora. In: T. Raso, H. Mello (eds.), Spoken Corpora and Linguistics Studies, Amsterdam: Benjamin.

Moneglia, M. & Panunzi, A. (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In: I. Korzen & E. Cresti (eds.) *Language, Cognition and Identity. Extension of the Endocentric/Esocentric Typology*. Firenze: FUP, pp. 27-46.

Moneglia, M., Monachini, M., Calbrese, O., Panunzi, A., Frontini, F., Gagliardi, G. & Russo, I. (2012). The IMAGACT cross-linguistic ontology of action. A new infrastructure for natural language disambiguation. In: N. Calzolari et al. (eds.), *Proceedings of the Eigth International Conference on Language Resources and Evaluation – LREC'12*, pp. 948-955.

Rosch E. (1978). Principles of Categorization. In: Rosch E. and Lloyd B.B. (eds.), *Cognition and Categorization*. Hillsdale, NW: Erlbaum. 27-48.

Acknowledgements

Il progetto IMAGACT è stato finanziato dalla regione Toscana nell'ambito del programma PAR.FAS. (linea di azione 1.1.a.3) . Ulteriori ricerche sul database IMAGACT, incluso questo articolo, sono state realizzate grazie al contributo del progetto MODELACT (2013-2016), finanziato nell'ambito del programma nazionale Futuro in Ricerca.

A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition

Ellert Thor Johannsson, Simonetta Battista
A Dictionary of Old Norse Prose, University of Copenhagen
nkv950@hum.ku.dk, sb@hum.ku.dk

Abstract

Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose (ONP) is a historical dictionary project at the Department of Scandinavian Research at the University of Copenhagen and covers medieval Old Norse material found in prose texts from the oldest written documents to early modern times (Old Icelandic 1150-1540 and Old Norwegian 1150-1370). After the publication of the first four volumes of an intended thirteen volume printed edition, a decision was made to change the format of the dictionary to a digital edition available online. This new medium has had to accommodate the already published printed material as well as unprinted and not fully edited material. In this article, we discuss how the change in ONP's format has provided the user with some benefits in working with the dictionary material, but also some new challenges. We compare the features of the printed volumes to the features of the online version and in doing so address some key questions that relate to the perspective of the user and how the digital version can be further improved.

Keywords: historical dictionary; electronic publication; digitalization

1 Introduction

Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose (ONP) was established in 1939 with the original intention of publishing a supplement to the already renowned Old Norse dictionaries of the 19th century, the works of Cleasby & Vigfusson (1874) and Johan Fritzner (1886, 1891, 1896). Furthermore ONP was limited to prose language, as a leading work on poetic language had then recently been published (Jónsson 1931) and therefore the poetic vocabulary was deemed sufficiently accounted for.

As the dictionary work progressed, the editorial staff soon realized that the project would fare better as an independent new dictionary based on its own principles and procedures. Plans were laid out for a new historical scholarly dictionary of Old Norse that would strive to represent the actual original medieval material by adhering to rigorous philological standard. This included retaining the spelling of scholarly text editions and using a system of references or sigla not only referring to a particular edition but also to each specific manuscript that edition is based upon.

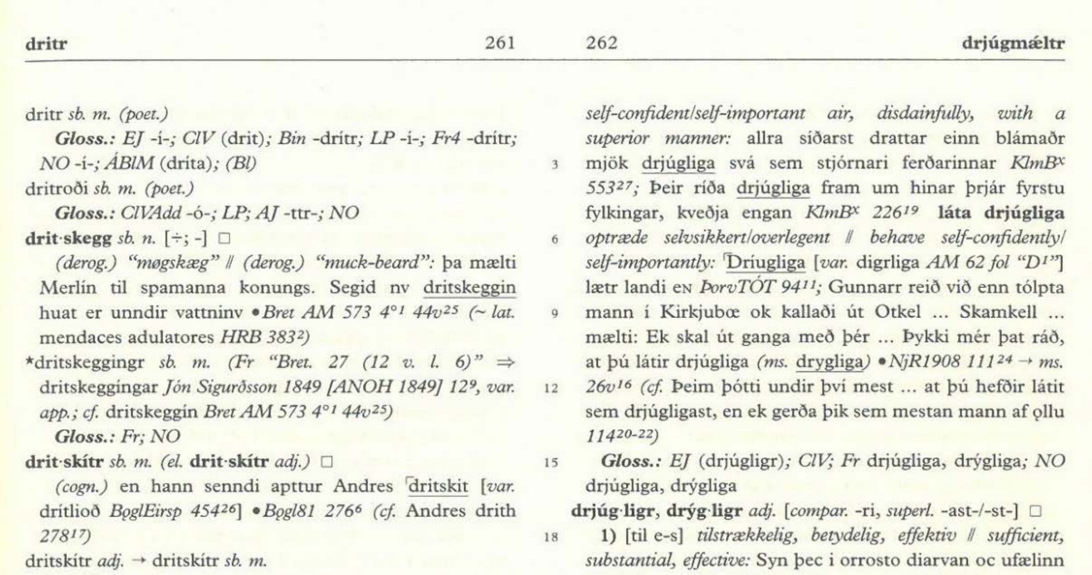
For the first decades, the work consisted mostly of excerpting Old Norse texts and building an impressive collection of dictionary citations consisting of around 750.000 handwritten slips. When the collection of citations was considered extensive enough to cover thoroughly the vocabulary from all known Old Norse Prose texts, plans were conceived to begin publication of series of printed volumes. In a publication celebrating the 25-year anniversary of the dictionary, the current chief editor stated the intentions of the dictionary staff for the publication to be complete within 25 years (Widding 1964: 21). In spite of this ambitious plan, the printed edition of the actual dictionary did not commence until 1989 with a volume of indices. This volume was the first in a series of estimated thirteen printed volumes scheduled for publication over the next decades. The printed publication continued with three more printed volumes (ONP1-3, covering the alphabet from *a-* to *em-*) at roughly five year intervals, the latest of which appeared in 2004. The remaining nine volumes would have thus likely taken around 45 years in production. Therefore, in 2005 a radical decision was made about the future of the dictionary. Instead of printed volumes, it should become a digital publication freely available on the web.

After the printed publication of the dictionary had been suspended, the work began on preparing the material for digitalization and electronic publication. This preparation work entailed various types of tasks including converting old reference sigla to the latest scholarly editions, scanning all citation slips under relevant headwords, as well as scanning scholarly editions and obtaining the necessary copyrights. This work finished in 2009 and in 2010 the first version of the digital ONP appeared online. Unlike the printed volumes, the digital edition is not a refined finished product. The decision to change the format of the dictionary also entailed making the dictionary material available online in a relatively raw form rather quickly. The first version of the digital edition consisted of two somewhat different components: on the one hand, the already published volumes in digital form and on the other hand the basic dictionary material, i.e. the collection of citations as handwritten slips, scanned under the appropriate lemma. The basic idea was that after making all the basic material accessible online the work of the editors would continue, gradually organizing the collection of citations, writing definitions and structuring the raw material into a more dictionary like format. In the long term, the online version aims to resemble the printed edition in terms of scope and attention to detail, although the nature of this new medium as well as the nature of the material has required a new approach and different editorial procedures from the ones applied to the printed publication.

This paper illustrates the difference between the printed edition and the digital edition of ONP with special focus on the user's perspective. Some of the questions that we address in our discussion are: What is required or expected of the user to be able to take advantage of the features of the dictionary? What are the pros and cons from the user's perspective of both paper and digital publication? How does the digital format offer the user different ways of accessing and working with the lexicographic material? What are the benefits and disadvantages of the current form of the dictionary? What improvements would be of further benefit to the user?

2 The basic characteristics of ONP

ONP has many features in common with similar dictionaries, but adds many layers to the information given. Figure 1 demonstrates what kind of information is displayed in the printed form of the dictionary and how it is structured.



dritr 261

dritr *sb. m. (poet.)*
Gloss.: *EJ* -í-; *ClV* (drit); *Bin* -drítr; *LP* -í-; *Fr4* -drítr; *NO* -í-; *ÁBlM* (dríta); *(Bl)*

dritroði *sb. m. (poet.)*
Gloss.: *ClVAdd* -ó-; *LP; AJ* -ttr-; *NO*

drit·skegg *sb. n.* [÷; -] □
(derog.) "møgskæg" // (derog.) "muck-beard": þa mælti Merlín til spamanna konungs. Segid nv dritskeggin huat er unndir vattninv •*Bret AM 573 4°^{1} 44v^{25}* (~ *lat.* mendaces adulatores *HRB 383^{2})*

*dritskeggingr *sb. m. (Fr "Bret. 27 (12 v. l. 6)"* ⇒ dritskeggingar *Jón Sigurðsson 1849 [ANOH 1849] 12^{9}, var. app.; cf.* dritskeggin *Bret AM 573 4°^{1} 44v^{25})*
Gloss.: *Fr; NO*

drit·skítr *sb. m. (el.* **drit·skítr** *adj.)* □
(cogn.) en hann senndi apttur Andres ⸢dritskit [*var.* drítlioð *BǫglEirsp 454^{26}*] •*Bǫgl81 276^{6} (cf.* Andres drith *278^{17})*

dritskítr *adj.* → dritskítr *sb. m.*

262 **drjúgmæltr**

self-confident/self-important air, disdainfully, with a superior manner: allra síðarst drattar einn blámaðr mjök drjúgliga svá sem stjórnari ferðarinnar *KlmB^{x} 553^{27};* Þeir ríða drjúgliga fram um hinar þrjár fyrstu fylkingar, kveðja engan *KlmB^{x} 226^{19}* **láta drjúgliga** *optræde selvsikkert/overlegent // behave self-confidently/self-importantly:* ⸢Driugliga [*var.* digrliga *AM 62 fol "D^{1}"*] lætr landi eɴ *ÞorvTÓT 94^{11};* Gunnarr reið við enn tólpta mann í Kirkjubœ ok kallaði út Otkel ... Skamkell ... mælti: Ek skal út ganga með þér ... Þykki mér þat ráð, at þú látir drjúgliga *(ms.* drygliga*)* •*NjR1908 111^{24} → ms. 26v^{16} (cf.* Þeim þótti undir því mest ... at þú hefðir látit sem drjúgligast, en ek gerða þik sem mestan mann af ǫllu *114$^{20-22}$)*
Gloss.: *EJ* (drjúgligr)*; ClV; Fr* drjúgliga, drýgliga*; NO* drjúgliga, drýgliga

drjúg·ligr, drýg·ligr *adj.* [*compar.* -ri, *superl.* -ast-/-st-] □
1) [til e-s] *tilstrækkelig, betydelig, effektiv // sufficient, substantial, effective:* Syn þec i orrosto diarvan oc ufælinn

Figure 1: From ONP3.

The main features are:

- two categories of lemma, normal vs. bold type; the actual entries are in bold, while words which are outside the corpus are shown in normal type. These can be poetical words (e.g. *dritr* and *dritroði* in Figure 1), foreign words which are not morphologically integrated into Old Norse, or the so-called "ghosts" (e.g. *dritskeggingr* in Figure 1). A "ghost" is a word found in other dictionaries, which is based on obsolete editions or misreading of a manuscript, and therefore has now "disappeared";
- non-normalized orthography, i.e. the orthography of the relevant manuscript or scholarly edition is kept in the dictionary citations;
- two target languages: Danish and English;
- references to foreign parallel texts (esp. Latin);
- detailed system of sigla indicating not only reference to an edition but also the actual manuscript for each section of the text (in some cases different manuscripts are used within the same edition);
- morphological information (inflectional pattern and verb conjugation) based on texts (mainly the actual examples found in the dictionary database);
- syntactic information (especially verb compliments and prepositional use);
- phrases and collocations;
- the citations are taken from actual Old Norse texts (editions or manuscripts) and are not modified in any way;

- the definitions are based on detailed analysis of the excerpted dictionary citations as well as secondary literature;
- references to glossaries and occasional references to secondary literature at the bottom of each dictionary entry.

3 The typical user and the prerequisites for using ONP

The user of ONP is typically a scholar of Old Norse language or literature, or medieval history and culture. This can be deduced from two "active" sources: the feedback forms on ONP's website and specific requests addressed to the dictionary staff. Students of these fields also use ONP, especially those who are advanced enough in their studies to use original text material.

Using ONP has always required some degree of expertise. The dictionary is intended to be primarily used as an academic or scholarly research tool. The fundamental assumption is therefore that the user can read and understand the language and is familiar with non-normalized text editions, i.e. scholarly editions of Old Norse texts that seek to render the original spelling of the medieval source. Orthography in medieval texts can be highly irregular and often inconsistent, even within the same manuscript or document. This is a fact the user needs to be aware of.

Related to this is the assumption that the user possesses knowledge to be able to "translate" or "convert" non-normalized forms into standard dictionary entry forms. This is a twofold task. First the user has to identify the normalized spelling of the word in question. Secondly the user has to know the basic form of the entry in order to look it up. For example when the user comes across forms like *diarvan* or *djorf* in a text edition he has to know that they are graphical and inflectional variations of the adjective *djarfr* 'daring', which is the normalized standard form of the headword (in the nominative singular masculine) to be able to look it up in the dictionary.

The user is also required to have some knowledge of grammar, or at least familiarity with the grammatical terminology and the presentation of grammatical information. The presentation of grammatical information in the dictionary entries is explained in some detail in the volume of indices (ONP:Registre 1989), which also includes a detailed list of all abbreviations ONP uses.

4 Comparing and contrasting the printed vs. the digital edition

The basic division of ONP into printed volumes and digital publication is fairly clear cut. As the publication of printed volumes has been suspended the digital version online consists of all the available material, including the previously printed material. This has already been mentioned above in section 1. However the online version is further divided into material that has been edited in the framework of the online version and material that has not been edited yet. Yet another level exists as the online edited material can be further divided into semantically edited material (primarily nouns) and structurally edited material (primarily verbs). Figure 2 gives an overview of this multi-level structure:

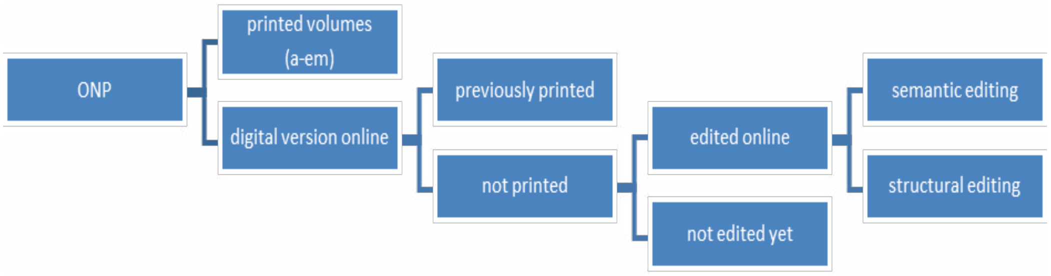


Figure 2: The structure of ONP's published material.

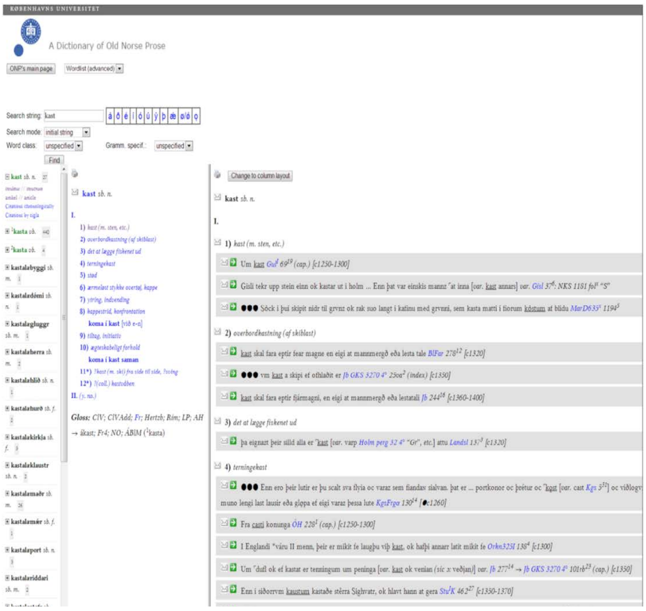
It is therefore not a straight forward task to compare the printed edition of the dictionary and the digital edition from the user's perspective. The multi-level structure of the material means that contrasting the printed edition vs. the digital edition gives a different picture depending on what exactly is being compared.

- If we compare a dictionary entry from one of the printed volumes with the same entry in the online version, we would find them almost identical, although the online version shows all available examples and not only the representative ones selected to appear in print. These "previously unpublished citations" appear outside of the structure of the dictionary article so the user needs to figure out where they belong in the article structure.
- If we compare a noun from the printed edition to a semantically similar noun found in the online version we would see some clear differences. There would be a difference between nouns that have been edited online and those, which have been not: the edited nouns have the structure of a dictionary entry with numbered definitions similar to the printed volumes. The citations are more accessible as they are typed up and correctly placed under the appropriate definition; the nouns that have not yet been edited are only available as a collection of all the citation slips under a normalized headword.
- When comparing verbs, the difference between the printed volumes and the online version is even more apparent since verbs have only been edited online for their structure, but not yet for semantic content. The user would find useful information about the verb constructions and verbal modifiers, but no definitions.

Looking at what these different types of comparisons have in common we can see that the main feature of the online version is the amount of actual examples made available. This is especially helpful when the user has good knowledge of the language and is able to himself find his way to a useful example by going through the list of citations. However there are still some structural features missing in many cases. Even though the large number of examples is beneficial to some it might be overwhelming to others, especially in those headwords where there are hundreds of examples. This can make searching for something specific very time consuming and in the absence of dictionary article structure, quite difficult.

5 Interacting with the dictionary material using the online version of ONP

There are several features of the online version that make accessing and working with the material easier than in the printed edition. Search in the printed edition is limited to alphabetical ordering of the lemmas. The online version offers a variety of tools to facilitate the search procedure. The main advantage is the possibility to tailor the search to the user's needs, e.g. search for all words that contain a particular derivational suffix, or compounds which have a common element, e.g. words where –*bátr* 'boat' is the second part. There are other ways to make the search more precise, e.g. by limiting the results to certain parts of speech, or in case of nouns, grammatical gender.



Search string: kast
Search mode: initial string
Word class: unspecified Gramm. specif.: unspecified

kast sb. n. 27
struktur // structure
artikel // article
Citations chronologically
Citations by sigla
¹kasta vb. 440
²kasta vb. 4
kastalabyggi sb. m. 1
kastaladómi sb. n. 1
kastalagluggr sb. m. 1
kastalaherra sb. m. 2
kastalahlið sb. n. 1
kastalahurð sb. f. 2
kastalakirkja sb. f. 3
kastalaklaustr sb. n. 2
kastalamaðr sb. m. 28
kastalamær sb. f. 1
kastalaport sb. n. 3
kastalariddari sb. m. 2

kast sb. n.

I.
1) kast (m. sten, etc.)
2) overbordkastning (af skiblast)
3) det at lægge fiskenet ud
4) terningekast
5) stød
6) ærmeløst stykke overtøj, kappe
7) ytring, indvending
8) kappestrid, konfrontation
koma í kast [við e-n]
9) tiltag, initiativ
10) ægteskabeligt forhold
koma í kast saman
11*) ?kast (m. ski) fra side til side, ?sving
12*) ?(coll.) kastodben
II. (y. no.)

Gloss: ClV; ClVAdd; Fr; Hertzb; Rím; LP; AH → ákast; Fr4; NO; ÁBlM (¹kasta)

kast sb. n.

I.

1) kast (m. sten, etc.)

Um kast Gul^I 69^19 (cap.) [c1250-1300]

Gisli tekr upp stein einn ok kastar ut i holm ... Enn þat var einskis mannz ˹at inna [var. kast annars] var. Gísl 37^6: NKS 1181 fol^x "S"

●●● Sóck i þui skipit nidr til grvnz ok rak suo langt i kafinu med grvnni, sem kasta matti i fiorum köstum af blidu MarD635^x 1194^5

2) overbordkastning (af skiblast)

kast skal fara eptir fear magne en eigi at mannmergð eða lesta tale BlFar 278^12 [c1320]

●●● vm kast a skipi ef ofhlaðit er Jb GKS 3270 4° 25va^2 (index) [c1350]

kast skal fara eptir fjármagni, en eigi at mannmergð eða lestatali Jb 244^16 [c1360-1400]

3) det at lægge fiskenet ud

þa eignazt þeir silld alla er ˹kast [var. varp Holm perg 32 4° "Gr", etc.] attu Landsl 137^3 [c1320]

4) terningekast

●●● Enn ero þeir lutir er þu scalt sva flyia oc varaz sem fiandan sialvan. þat er ... portkonor oc þrétur oc ˹kǫst [var. cast Kgs 5^31] oc viðlogv muno lengi last lausir eða glǫpa ef eigi varaz þessa lute KgsFrga 130^14 [●c1260]

Fra casti konunga ÓH 228^1 (cap.) [c1250-1300]

I Englandi *váru II menn, þeir er mikit fe laugþu viþ kast, ok hafþi annarr latit mikit fe Orkn325I 138^4 [c1300]

Um ˹dufl ok ef kastat er tenningum um peninga [var. kast ok venian (sic x veðjan)] var. Jb 277^14 → Jb GKS 3270 4° 101rb^23 (cap.) [c1350]

Enn i siðorrvm kaustum kastaðe stérra Sighvatr, ok hlavt hann at gera Stu^IK 462^27 [c1350-1370]

Figure 3: Example of an edited dictionary article in ONP Online.

A helpful search feature is the alphabetically ordered list of lemmas that appears on the left side of the screen when a search string is entered. This gives a quick overview of the words that have the same initial letters and might inspire the user to look up related words and compounds.

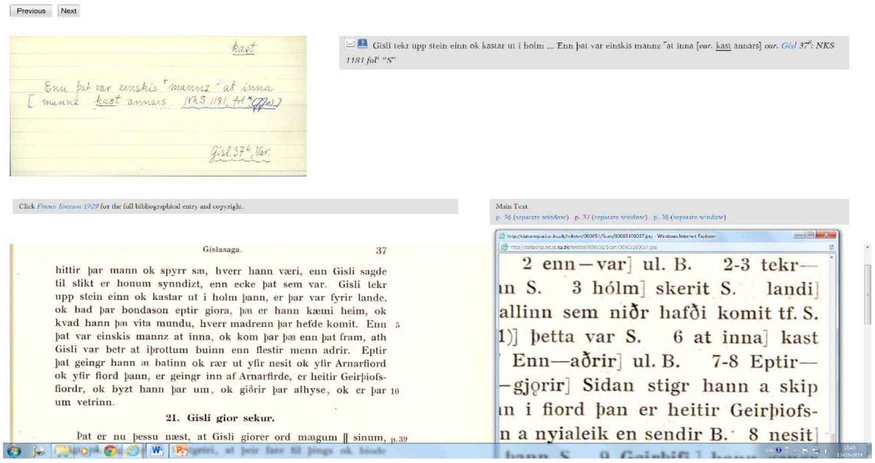
When it comes to material that has already been edited there are several additional features that facilitate working with the material. The structural overview (the middle column in Figure 2) is one such feature providing the user with a quick impression of different definitions from which to select the examples to view. This facilitates working through the list of citations, especially when headwords have a large number of citations associated with them.

The glossary section is located at the bottom of the structural overview. This list of abbreviations illustrates in which earlier dictionaries and glossaries the word is found. If the word is found under a different entry in these older works this is also indicated. Such glossary section is also a feature of the printed edition, but the online edition has the advantage of providing a hyperlink to the glossary or dictionary in question. Currently this feature is limited to the dictionary of Fritzner (1886, 1891, 1896): If the user clicks on the abbreviation *Fr.* a new tab opens displaying the relevant article.

On each level of the article an envelope icon is displayed. By clicking on this icon a new window will pop up giving the user the possibility to bring suggestions and comments to the attention of the editors. This can be done anonymously.

Once the user has chosen which citations to investigate first, she can click on a single definition and get a list of all relevant examples.

Clicking on the green arrow next to a specific citation will bring up a separate window:



kast

Enn þat var einskis ⸢mannz ⸣at inna
[mannz kast annars NkS 1181, fol* (77a)]

Gísl 37⁶, Var.

Gísli tekr upp stein einn ok kastar ut i holm ... Enn þat var einskis mannz ⸢at inna [*var.* kast annars] *var.* Gísl 37⁶: *NKS 1181 fol˟ "S"*

Click Finnur Jónsson 1929 for the full bibliographical entry and copyright.

Gislasaga. 37

hittir þar mann ok spyrr sæ, hverr hann væri, enn Gisli sagde til slikt er honum synndizt, enn ecke þat sem var. Gisli tekr upp stein einn ok kastar ut i holm þann, er þar var fyrir lande, ok bad þar bondason eptir giora, þæ er hann kæmi heim, ok kvad hann þæ vita mundu, hverr madrenn þar hefde komit. Enn 5 þat var einskis mannz at inna, ok kom þar þæ enn þat fram, ath Gisli var betr at iþrottum buinn enn flestir menn adrir. Eptir þat geingr hann æ batinn ok rær ut yfir nesit ok yfir Arnarfiord ok yfir fiord þann, er geingr inn af Arnarfirde, er heitir Geirþiofs-fiordr, ok byzt hann þar um, ok giðrir þar alhyse, ok er þar 10 um vetrinn.

21. Gisli gior sekur.

Þat er nu þessu næst, at Gisli giorer ord mægum ‖ sinum, p. 39

Main Text
p. 36 (separate window) p. 37 (separate window) p. 38 (separate window)

2 enn—var] ul. B. 2-3 tekr—
ın S. 3 hólm] skerit S. landi]
allinn sem niðr hafði komit tf. S.
1)] þetta var S. 6 at inna] kast
Enn—aðrir] ul. B. 7-8 Eptir—
—gjǫrir] Sidan stigr hann a skip
ın i fiord þan er heitir Geirþiofs-
n a nyialeik en sendir B. 8 nesit]

Figure 4: Closer look at a specific citation: citation slip, typed citation and scanned page from the edition (ONP Online).

Here the user will find access to a variety of additional information by further clicking on the relevant links. By clicking on the reference siglum the user has immediate access to the index entry of that particular text. This is displayed in yet a separate window. Here the user can quickly find out

more information about the text edition the dictionary refers to, i.e. what manuscript are used and where the text is found in those manuscripts (folio range). There might also be some other sources ONP has used in referring to this particular texts, such as other editions or even manuscripts that are not used by the editor in the scholarly edition the citation is taken from. This is the information that was also published in the printed volume of indices (ONP:Registre 1989). The user can seek further information about the scholarly edition by clicking on the title and on a separate page get a full entry from the ONP's bibliography. Once this information has been displayed the user has the option of browsing the bibliography, i.e. see what other scholarly editions or secondary literature this same editor might have helped create. The relative ease of access to all this secondary information can save scholarly users a lot of time as they often need the original citations to refer to in their own work.
Another feature that is of great benefit for the user is the access to the relevant scholarly edition itself, since ONP provides the possibility to view the actual printed page the citation is taken from. This is a huge improvement compared with the printed dictionary where the user had to get a hold of the actual book to get a closer look at the context of a specific citation. Almost all the works cited in the online edition are displayed in such a way. Due to copyright reasons the viewing of the editions is limited to three pages, i.e. the relevant page cited by the dictionary, the page immediately preceding it and the one immediately following it. This allows the user to gain a better grasp of the meaning of the word and further orient herself in the text the word is taken from.

5.1 Benefits of the digital edition

We have already in our discussion touched upon some of the benefits and disadvantages of the digital edition of ONP from the perspective of the dictionary user. We will sum up the most important features that are of benefit to the user:

- Multiple search capabilities: Chance to tailor the search according to needs and interest of the user.
- Access to all the material: The user has access to all the raw dictionary material. All the citation slips are available as well as all edited dictionary entries. The user is able to make some use of ONP's collection of citations, even though he is not able to derive the full benefits of an edited dictionary entry.
- Possibility of greater context: The user has some access to the actual edition the dictionary citations are taken from and is able to browse through relevant parts of editions.
- Interactive communication: The online edition offers an easy way of bringing comments and suggestions to the attention of the dictionary staff.
- In the edited part, citations are assigned to a definition: The editor of the dictionary entry strives to assign each citation to a definition, but noting if the use of the word is ambiguous or possibly can be interpreted as belonging to a different definition from the one it is assigned to.

5.2 Additional benefits for the editors

In the previous discussion it has become evident that the digital edition of ONP has many benefits to the users in terms of quick and easy access to the relevant information. Some of those benefits also extend to the editors. There are mainly two that we wish to highlight. The first is that the editor receives feedback, suggestions or comments from users regarding the dictionary entries that have already been edited and made available on the web. The editor can then take the necessary measures to improve the article in question and make corrections or additions which will be made available next time the online version is updated. This is a new type of interaction between users and editors with potentially mutual benefits.

This is closely related to another benefit the digital edition gives to the editor, i.e. the possibility to correct and improve on his/her contribution to the dictionary, even after it has been published on the web. This is of course a huge change from printed publication, where a lot of the editor's work involved reading and re-reading his own work and works of others in order to make sure that no errors or misprints would find their way into the final published product. This process can be quite tedious and is not necessary to the same extent when the publishing platform is a digital version on the internet. In practical terms this means that now the editor can spend less time proof-reading and more time editing actual dictionary entries, thus increasing the productivity.

Both of these benefits, the possibility to interact with the user and the possibility to improve published dictionary entries, change the nature of the dictionary making process slightly. The process of editing and preparing the printed volumes of the dictionary was a collective effort where many editors where often involved in the production of an individual dictionary entry. The editing procedure for the digital publication is in its nature such that it relies less on the collective effort and more on the individual editor. As a result the dictionary entries are now signed with the initials of the editor in charge of each particular entry. Although the production of the dictionary is still very much a result of collaboration and teamwork, it is the individual responsibility of each editor to make sure his/her dictionary entries adhere to the rigorous standards ONP holds itself to and to respond to the user when the need to do so arises. This makes the editing work more transparent and gives the user a chance to bring suggestions directly to the person responsible for a particular headword if needed.

6 Room for improvement and future plans

Even though many aspects of the digital edition of ONP provide the user with increased possibilities to interact with the dictionary material and editors there is still very much room for improvement. In some respects the printed volumes of the dictionary provide the user with more consistent distribution of information, because of the discrepancies between levels of the online edition discussed in section 4 above. The main difference is between edited material and the unedited part where the information is only accessible as a more loose structure collection of citations. However it is our hope that these discrepancies will gradually diminish as the editing work progresses and the dictionary material keeps getting updated. Currently around 140.000 citations have been semantically edited online and around 170.000 have been structurally edited. This means that around 300.000 citations are not yet edited.

6.1 Focus areas for improving ONP online

One area where there is room for improvement involves the target language in the edited part of the digital edition. The current editorial procedure calls for the target language to be either English or Danish. In practice this means that in the latest dictionary entries published online, the definitions are monolingual and mostly in Danish. This limits the user base of the dictionary somewhat, although we hope to bring back the consistent bilingualism of the printed volumes in a not so distant future.

Another area of focus involves reference to new editions. ONP has always tried to keep up with the latest scholarly advances in the field of Old Norse studies. If a new edition of a text appeared then all of the dictionary's citations involving that text would have to be updated to the new edition to reflect those advances. This means there is some internal inconsistencies in the printed volumes, i.e. some of the references in ONP3 did not exist when ONP1 came out. When the change to digital publication was put in place it was decided that point of reference should be the material as it was at that point in time. This means that there have been no updates to new editions after 2005. An historical scholarly dictionary like ONP needs to keep up with the latest research. Otherwise it slowly but surely will become obsolete. It is our hope that part of the continuous improvement of the digital edition will include updating the reference sigla to the latest available editions.

6.2 Potential future improvements of benefit to the user

The current plan for ONP calls for the continuation of the editorial work and gradually assigning all unedited citation to structured dictionary entries available in the online version. In addition to completing the editing work, there are several other features, which will increase usability of the digital edition. Besides the ones already discussed, such as increasing the consistency between different levels of the online version as well as consistently providing the user with definition in two languages, we foresee other beneficial improvement that will become part of ONP online in the future.

The search options now are limited to the normalized lemma list. Part of the dictionary database includes a notation of by-forms and alternative forms, some of which are quite common although not searchable through the regular search options. A first step in increasing the search possibilities would be to link such non-normalized by-forms and alternative forms to their respective headwords and allow them to be part of the searchable material. The search options can be improved further. It is possible with relatively little effort to add a feature that searches not only in the list of lemmas, but also all citations that have been typed in. Of course, this feature would be rather primitive to begin with as the citation texts are not lemmatized and not even normalized, but it definitely would be helpful to the user in certain cases.

Another feature that might in the long term become part of the online edition is a search feature, which would limit the search to certain periods, places (scriptoria), geographical areas (e.g. West Iceland) or even certain manuscripts (e.g. all texts found in a particular manuscript). The structure of the dictionary database allows for many kinds of searches, which are already available for the staff. These features are though still in a relatively early stage and have not been laid out in practical terms just yet but they would give the user new alternative ways to interact with the dictionary material.

7 References

Cleasby, Richard & Gudbrand Vigfusson (1874). *An Icelandic-English Dictionary*, Oxford: Clarendon Press.

Fritzner, Johan (1886, 1891, 1896). *Ordbog over Det gamle norske Sprog* 1-3, rev. ed., Kristiania: Den norske forlagsforening.

Jónsson, Finnur (1931). *Lexicon poeticum antiquæ linguæ Septentrionalis / Ordbog over det norsk-islandske skjaldesprog*. Original edition by Sveinbjörn Egilsson. 2nd ed. Copenhagen: Kongelige nordiske oldskriftselskab.

ONP Online. Accessed at: www.onp.ku.dk [07/04/2014].

ONP:Registre = Ordbog over det norrøne prosasprog : Registre / A Dictionary of Old Norse Prose: Indices. (1989). Copenhagen: The Arnamagnæan Commission.

ONP1 = Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (1995). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 1: a-bam. Copenhagen: The Arnamagnæan Commission.

ONP2 = James E. Knirk, Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (2000). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 2: ban-da. Copenhagen: The Arnamagnæan Commission.

ONP3 = Helle Degnbol, Bent Chr. Jacobsen, James E. Knirk, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (2004). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose* 3: de-em. Copenhagen: The Arnamagnæan Commission.

Widding, Ole (1964). Den Arnamagnæanske Kommissions Ordbog, 1939-1964: Rapport og plan, Copenhagen: G.E.C.GADS Forlag.

Making a Learner's Dictionary of Academic English

Diana Lea
Oxford University Press
diana.lea@oup.com

Abstract

This paper gives an account of the development of the *Oxford Learner's Dictionary of Academic English*, a dictionary for non-native-English-speaking students who are studying academic subjects at tertiary level through the medium of English. First, a corpus of academic English was created, using high-quality texts from a broad range of disciplines, maintaining a balance between textbooks, containing typical student reading material, and journal articles, modelling expert academic writing. Drawing on the corpus, and on previous research in the field, a core headword list of "general academic vocabulary" was drawn up. This list was continually supplemented with necessary defining words, complements, collocates, synonyms and antonyms of these words as the work of compiling the dictionary entries progressed. In compiling the entries, particular challenges were encountered in reconciling the academic and pedagogic requirements of the dictionary. This can be seen, for example, in decisions about sense division and the wording of definitions and also in the selection of example sentences from the corpus. Editors had to find ways to represent academic language faithfully whilst making it accessible to learners. The result is a genuinely academic learner's dictionary that should offer real help to learners with their academic writing.

Keywords: learner's dictionary; academic English; EAP; corpus

1 Introduction

Academic vocabulary has received considerable research attention, in particular with the effort to identify a core academic vocabulary, as distinct from general English vocabulary on the one hand and discipline-specific technical vocabulary on the other. Coxhead (2000) proposed the Academic Word List (AWL), a list of 570 word families, divided into ten sublists, found to account for around 10% of the words in a corpus of academic English, as opposed to 1.4% of the words in a fiction corpus. The AWL was generally well received by teachers and has been quite widely exploited in published materials (Coxhead 2011). More recently, however, Paquot (2010) and Gardner and Davies (2013) have proposed alternative lists, addressing some of the perceived shortcomings of the AWL, notably its exclusion of the 2,000 word families of the General Service List (West 1953) as already "known" to students at this level, and its construction around whole word families, regardless of discrepancies in frequency (Paquot 2010: 17) and even core meaning (Gardner and Davies 2013: 3) between different word family members. Hyland and Tse (2007), however, have questioned the whole validity of a single, cross-disciplinary, core academic vocabulary, partly on the basis that the same words may be used in widely different ways in different disciplines.

In contrast, there has been much less attention paid to the idea of a dictionary of academic English, as opposed to a word list. Kosem (2008) surveyed a number of dictionaries marketed for university students, mostly aimed at native speakers, and concluded that, apart from supplementary material on academic writing, these dictionaries differ little in content from the general-purpose dictionaries on which they are based. Learners' dictionaries, such as the *Longman Dictionary of Contemporary English* and the *Oxford Advanced Learner's Dictionary* have started to acknowledge the interest in academic vocabulary and writing, by marking words in the AWL and including their own academic writing supplements, but they remain essentially dictionaries of general English. In this paper, I shall give an account of the *Oxford Learner's Dictionary of Academic English (OLDAE)* (2014), which I believe to be the first widely available, genuinely academic, learner's dictionary.[1]

I shall begin by outlining the principles and parameters that were established at the start of the project. I shall then describe some of the challenges that were encountered in the course of the project, notably building the academic corpus on which the dictionary is based; determining the headword list; and above all, reconciling the academic and pedagogic requirements, especially with regard to writing the definitions and selecting the example sentences. I shall then conclude by evaluating the achievements and limitations of the dictionary and suggesting some possible future developments.

2 Principles and Parameters

The principles and parameters of *OLDAE*, as a learner's dictionary of academic English, will be found to differ quite widely from those proposed by Kosem (2010). Although I agree with the general thesis that no one is a native speaker of academic English and that both native and non-native-English-speaking students should be viewed as "apprentice writers" when it comes to academic writing (Kosem 2010: 49), I believe the particular needs of these two groups are sufficiently different that a single dictionary cannot serve both equally well. *OLDAE* is therefore designed to serve the needs of non-native-English-speaking students of English for Academic Purposes (EAP), at a range of levels from the B1 student on a foundation course, to students at C2 level writing their Masters' dissertations. The dictionary is also, however, partly for practical reasons, much smaller in scale than that proposed by Kosem. It largely excludes general and technical English, focusing essentially on a core academic vocabulary across the disciplines, but taking a broad view of what this might encompass,

[1] I am aware of the Louvain EAP Dictionary (Granger and Paquot 2010), an innovative online dictionary-cum-writing-aid, which can claim to be the very first learner's dictionary to be based on analysis of academic corpora. However, it is currently only available to staff and students at the Université catholique de Louvain; moreover, with only around 900 headwords, it may be arguably more a writing-aid-with-dictionary-entries than a complete dictionary in itself.

not attempting to identify a single definitive list for all students, but exploring the nuances of usage of so-called "general academic" words across different disciplines. It is specifically intended to help EAP students with their academic writing across a range of genres. The most fundamental principle underpinning the dictionary was that it should be based on a thorough analysis of genuine academic writing; this meant constructing a new corpus, the 85-million-word Oxford Corpus of Academic English (OCAE).

3 The Corpus

Output from a corpus can only ever be as representative and appropriate as the corpus itself; the content of OCAE therefore needed to match, as nearly as possible, the materials that target users of the dictionary would be reading and writing themselves. In terms of "reading" content, this was provided by higher education textbooks, mostly aimed at undergraduate level, which at 42 million words constituted just under half the corpus. "Writing" content was more challenging: ideally, what was needed was some 40-50 million words of very high-quality student essays and dissertations, but creating such a corpus was beyond our resources. Instead, we substituted expert academic writing from journals, monographs and handbooks, adding up to a further 43 million words; this assured the quality, and meant the findings on the usage of academic vocabulary would be sound. However, it did mean text of a much higher level than our students would be attempting to write, which made the selection of authentic but user-friendly example sentences rather more challenging – but I shall return to this point later. In terms of the balance of disciplines in the corpus, we tried to match this approximately to the profile of disciplines being studied by international students at English-medium universities. Figure 1 shows the breakdown of the corpus into different subject areas.

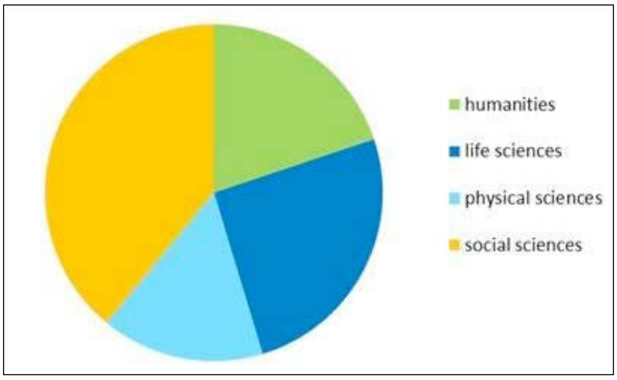
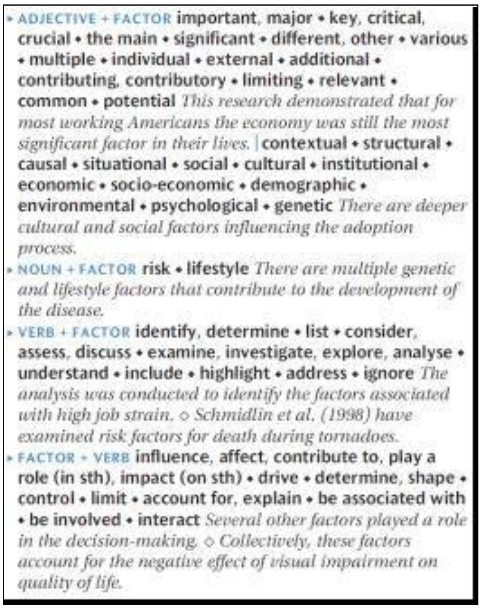


Figure 1: Breakdown of texts by subject area in the Oxford Corpus of Academic English.

Natural sciences (divided into life sciences and physical sciences) and social sciences each account for around 40% of the corpus, with the remaining 20% made up of humanities texts. The largest single disciplines were business and medicine, at around 8% each.

The Headword List

The headword list was built up organically as the work of compiling the dictionary progressed. We began with a core list, comprised of words from the AWL, after checking them against the corpus. We added to this four word lists of our own, extracted from the four subcorpora of OCAE, as compared with a fiction corpus. As work began on compiling entries for these words, the headword list was rapidly augmented with necessary defining words, complements, collocations, synonyms and opposites of the words in the initial list. Collocations were an especially rich source of additional headwords. If dictionary users were to be enabled to use the core words productively, the linguistic contexts in which they could be used would be all-important. This meant generous treatment of collocations in the dictionary entries: for nearly 700 of the most important, collocationally prolific words, a separate section of the entry lists collocations in the style of a collocations dictionary (see Figure 2). These collocation entries then fed back into the main headword list of the dictionary as it was obviously important that no word should be listed as a collocation without being defined and exemplified in its own separate entry. In this way, the headword list expanded beyond a core of 3-4,000 academic words, to encompass higher-level words that are more context-specific, though still mostly at the subtechnical level, as well as presenting in appropriate academic contexts the functional words that are actually basic to all forms of discourse. The process of enhancing the headword list continued throughout the compiling and editing process and was completed by a trawl through the corpus for items of a certain level of frequency that had not been picked up and that might also warrant inclusion. It would have gone way beyond the scope of the project to include every word that a particular student might wish to look up; the aim was to give thorough coverage to the core words that all students would need, plus a generous helping of supplementary words that would be useful to many.



- ADJECTIVE + FACTOR **important, major • key, critical, crucial • the main • significant • different, other • various • multiple • individual • external • additional • contributing, contributory • limiting • relevant • common • potential** *This research demonstrated that for most working Americans the economy was still the most significant factor in their lives.* | **contextual • structural • causal • situational • social • cultural • institutional • economic • socio-economic • demographic • environmental • psychological • genetic** *There are deeper cultural and social factors influencing the adoption process.*
- NOUN + FACTOR **risk • lifestyle** *There are multiple genetic and lifestyle factors that contribute to the development of the disease.*
- VERB + FACTOR **identify, determine • list • consider, assess, discuss • examine, investigate, explore, analyse • understand • include • highlight • address • ignore** *The analysis was conducted to identify the factors associated with high job strain.* ◇ *Schmidlin et al. (1998) have examined risk factors for death during tornadoes.*
- FACTOR + VERB **influence, affect, contribute to, play a role (in sth), impact (on sth) • drive • determine, shape • control • limit • account for, explain • be associated with • be involved • interact** *Several other factors played a role in the decision-making.* ◇ *Collectively, these factors account for the negative effect of visual impairment on quality of life.*

Figure 2: Collocations of *factor* from the *Oxford Learner's Dictionary of Academic English* (2014).

5 Editorial Policy

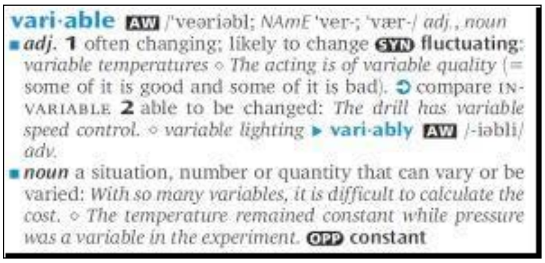
There is insufficient space here to give a full account of the editorial policy of the dictionary. However, many of the challenges can be considered as different manifestations of the one central challenge: how to reconcile the academic and pedagogic requirements of the dictionary. We wished to make this both an academic dictionary and a learner's dictionary, but there were cases where a compromise was called for. I shall focus on two key aspects here: the definitions and the example sentences.

5.1 Definitions

The starting point for most of the definitions in the dictionary was the definitions in the 8th edition of the *Oxford Advanced Learner's Dictionary (OALD)* (2010). These definitions have the advantage of accessibility for learners, especially as they are written within a carefully controlled defining vocabulary of 3,000 words (in fact reduced to 2,300 for *OLDAE*). In many cases, *OALD* definitions were retained unchanged. However, there were important decisions to be made, often not only over the wording of definitions but over the content. One example is that of *variable* as both noun and adjective. The *OALD* entry (Figure 3) distinguishes two separate senses for the adjective and just one, coverall sense for the noun. The derivative *variably* is nested in the adjective, undefined.

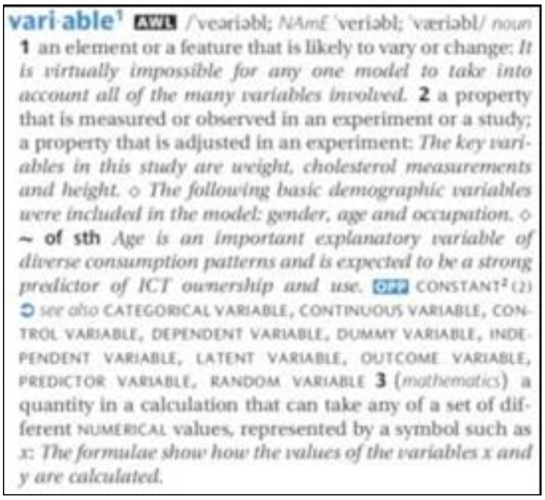
OLDAE, however, recognizes that, for academic purposes, *variable* is a very important word, and EAP students need to know a lot more about it. The noun entry (Figure 4) separates out two further, much more specific

meanings that are important in an academic context: one that is relevant to all experimental sciences, and a basic meaning in mathematics that students from a wide range of disciplines will need to know. As well as conveying much more information than the *OALD* entry, this splitting of senses enables the definitions to be much more precisely worded: students have more to read in this entry, but the burden of interpreting and applying what they have read is actually lighter, because these definitions spell things out much more clearly. The adjective entry (Figure 5) distinguishes four separate meanings, each with different synonyms and antonyms, whilst *variably* (Figure 6) is recognized as an important academic word in its own right, not just a derivative of *variable*, and isgiven its own entry with two distinct meanings.



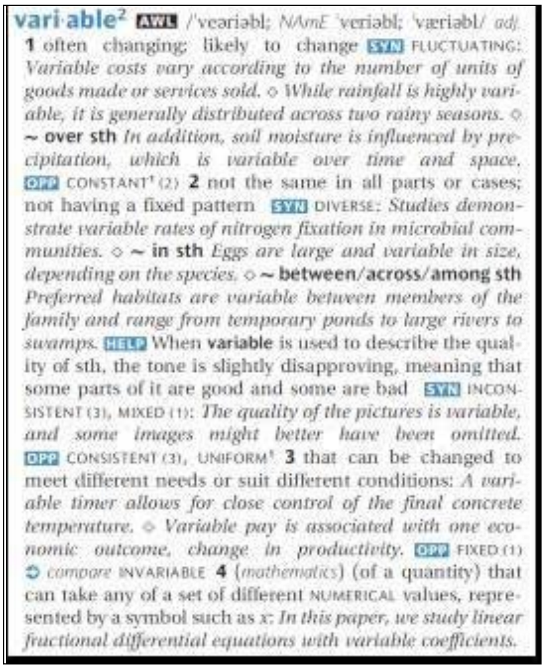
vari·able AW /ˈveəriəbl; *NAmE* ˈver-; ˈvær-/ *adj., noun*
■ ***adj.*** **1** often changing; likely to change SYN **fluctuating**: *variable temperatures* ◇ *The acting is of variable quality* (= some of it is good and some of it is bad). ↻ compare INVARIABLE **2** able to be changed: *The drill has variable speed control.* ◇ *variable lighting* ▸ **vari·ably** AW /-iəbli/ *adv.*
■ ***noun*** a situation, number or quantity that can vary or be varied: *With so many variables, it is difficult to calculate the cost.* ◇ *The temperature remained constant while pressure was a variable in the experiment.* OPP **constant**

Figure 3: Entry for *variable* from the *Oxford Advanced Learner's Dictionary,* 8th edition (2010).



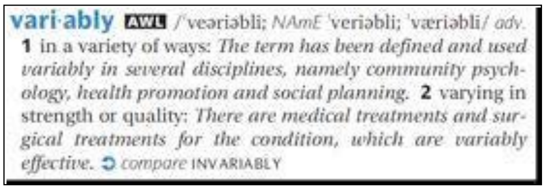
vari·able^1 AWL /ˈveəriəbl; *NAmE* ˈveriəbl; ˈværiəbl/ *noun*
1 an element or a feature that is likely to vary or change: *It is virtually impossible for any one model to take into account all of the many variables involved.* **2** a property that is measured or observed in an experiment or a study; a property that is adjusted in an experiment: *The key variables in this study are weight, cholesterol measurements and height.* ◇ *The following basic demographic variables were included in the model: gender, age and occupation.* ◇ **~ of sth** *Age is an important explanatory variable of diverse consumption patterns and is expected to be a strong predictor of ICT ownership and use.* OPP CONSTANT^2 (2) ↻ *see also* CATEGORICAL VARIABLE, CONTINUOUS VARIABLE, CONTROL VARIABLE, DEPENDENT VARIABLE, DUMMY VARIABLE, INDEPENDENT VARIABLE, LATENT VARIABLE, OUTCOME VARIABLE, PREDICTOR VARIABLE, RANDOM VARIABLE **3** (*mathematics*) a quantity in a calculation that can take any of a set of different NUMERICAL values, represented by a symbol such as *x*: *The formulae show how the values of the variables x and y are calculated.*

Figure 4: Entry for *variable, noun* from the *Oxford Learner's Dictionary of Academic English* (2014).



vari·able^2 AWL /ˈveəriəbl; *NAmE* ˈveriəbl; ˈværiəbl/ *adj.*
1 often changing; likely to change SYN FLUCTUATING: *Variable costs vary according to the number of units of goods made or services sold.* ◇ *While rainfall is highly variable, it is generally distributed across two rainy seasons.* ◇ **~ over sth** *In addition, soil moisture is influenced by precipitation, which is variable over time and space.* OPP CONSTANT^1 (2) **2** not the same in all parts or cases; not having a fixed pattern SYN DIVERSE: *Studies demonstrate variable rates of nitrogen fixation in microbial communities.* ◇ **~ in sth** *Eggs are large and variable in size, depending on the species.* ◇ **~ between/across/among sth** *Preferred habitats are variable between members of the family and range from temporary ponds to large rivers to swamps.* HELP When **variable** is used to describe the quality of sth, the tone is slightly disapproving, meaning that some parts of it are good and some are bad SYN INCONSISTENT (3), MIXED (1): *The quality of the pictures is variable, and some images might better have been omitted.* OPP CONSISTENT (3), UNIFORM^1 **3** that can be changed to meet different needs or suit different conditions: *A variable timer allows for close control of the final concrete temperature.* ◇ *Variable pay is associated with one economic outcome, change in productivity.* OPP FIXED (1) ↻ *compare* INVARIABLE **4** (*mathematics*) (of a quantity) that can take any of a set of different NUMERICAL values, represented by a symbol such as *x*: *In this paper, we study linear fractional differential equations with variable coefficients.*

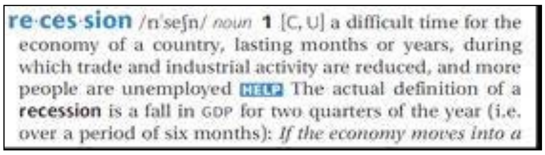
Figure 5: From the entry for *variable, adjective* from the *Oxford Learner's Dictionary of Academic English* (2014).



vari·ably AWL /ˈveəriəbli; *NAmE* ˈveriəbli; ˈværiəbli/ *adv.* **1** in a variety of ways: *The term has been defined and used variably in several disciplines, namely community psychology, health promotion and social planning.* **2** varying in strength or quality: *There are medical treatments and surgical treatments for the condition, which are variably effective.* ⊃ *compare* INVARIABLY

Figure 6: Entry for *variably* from the *Oxford Learner's Dictionary of Academic English* (2014).

A slightly different challenge is posed by a word like *recession* (Figure 7):



re·ces·sion /rɪˈseʃn/ *noun* **1** [C, U] a difficult time for the economy of a country, lasting months or years, during which trade and industrial activity are reduced, and more people are unemployed HELP The actual definition of a **recession** is a fall in GDP for two quarters of the year (i.e. over a period of six months): *If the economy moves into a*

Figure 7: From the entry for *recession* from the *Oxford Learner's Dictionary of Academic English* (2014).

The main definition that comes first is closely based on the definition of this word offered in the *OALD*. However, as our economics adviser pointed out, it is not, strictly speaking, a definition at all, but a description. On closer inspection, this will be found to be true of many "definitions" offered in general learners' dictionaries, and the dictionaries in general are all the better for it. They offer learners the degree of understanding they need in a form that is accessible to them. For the EAP student, however, the case is different. The student of economics (or history or geography or a number of related subjects) is not well served by a mere description of a recession, when it is in fact a very precisely defined economic term. Our solution was to offer the description first, followed by the "actual definition", clearly signalled as such. (The economics adviser, it must be confessed, was not happy with this solution, and felt that only the exact definition should be offered; however, this was a view offered from an entirely academic perspective, with no concession to the particular needs of foreign learners, and so the editor's view – that both general description and precise definition should be offered – prevailed.)

5.2 Example sentences

Selecting the example sentences was probably the most challenging aspect of compiling most entries and policy on this evolved over the course of the project, in some cases necessitating late revision of earlier compiled entries. Consultation with academics and EAP tutors at the planning stage impressed on the editors the need for extreme caution when lifting and editing examples from the corpus. Some were uncomfortable with the idea of editing corpus text at all. However, when faced with the reality of raw corpus text, set against the practical needs of the intended users of the dictionary, it became clear that many of the selected corpus examples would need some degree of editing to render them useful and appropriate for learners. Potential difficulties with unedited corpus text were numerous: very high-level vocabulary; difficult constructions; extremely long sentences; obscure and distracting detail; general oddness. Editors also had to take into account the fact that the academic genres in the corpus – textbooks and journal articles – were not the genres that students themselves would be writing. Textbook examples were often tempting, as they were clear and accessible, but many textbooks employ a tone of "expert speaking to student" that would not be appropriate in a student essay.[2]

We initially approached the task of selecting example sentences from an academic corpus with the feeling that it was in some way a different task from selecting examples for a general learner's dictionary from a general English corpus. Experience persuaded me, however, that this task was not in fact different in kind, though perhaps it was different in level of difficulty. The most useful examples are the most typical, which often means the most general:

> Taylor makes the following argument:...
>
> This approach yields dramatically lower estimates.
>
> Several other factors played a role in the decision-making.
>
> The most persuasive argument against this idea comes from Foster (2009).

Examples like this may not be taken directly from any one text; often they are a distillation of a number of different concordance lines, all of them very similar. Other examples – the majority – do contain context derived from a particular source text; and, where appropriate, may be taken from that text unedited. This helps them to feel more authentic; nonetheless, it is important that the context does not get in the way of understanding the linguistic point being presented in the example. The examples are intended to "feel authentic" but they cannot actually be authentic – even if completely unedited, they are inauthenticated the moment they are taken from their context and set in italic type in a learner's dictionary. Ultimately, though, the needs of the learner trump other considerations. Learners using this dictionary are not expected to immediately start writing fluent expert academic texts; what they need to acquire is a style that approaches more closely an appropriate academic style, whilst still being accessible from the level they are currently at.

6 Conclusion

The learner's dictionary is not an academic genre but a pedagogic one. A learner's dictionary of academic English needs to pay close attention to the rules and conventions of academic writing, and represent them as faithfully as it can, but the learner's needs still take precedence. *OLDAE* is designed

[2] The use of "we" is a case in point. Textbook writers use it frequently with the meaning of "you and I", to include the (student) reader in a comment such as, "In this chapter, we shall see ..." This style is not employed in research articles, where experts are writing for other experts, and it is not to be recommended for students writing for a tutor or examiner. A usage note at the entry for we in the dictionary explains this point.

to meet the specific needs of tertiary level students writing assignments in English in a wide range of disciplines. It covers a generous "core" academic vocabulary, showing not only the meanings of words, but how to use them in context. We hope it will be a valuable new resource for students. Its limitations are largely those imposed by the relatively limited size and scope of the project. Coverage does not go much beyond the "subtechnical" level of vocabulary, but it is assumed that the technical vocabulary of the student's own discipline will be explained as part of their subject course. The traditional format of print dictionary plus CD-ROM may also limit its appeal for some of today's students. However, there is a lot of scope for presenting, combining and expanding the content in different ways to make it even more useful and accessible to a wider range of users. For example, a customizable online subscription model could allow users to combine *OLDAE* content with both more general content and more technical content from other dictionaries, according to the subject they are studying. To make this a reality would require rather more work, both editorial and technical, but it seems worth aspiring to.

7 References

Coxhead, A. (2000). A New Academic Word List. In *TESOL Quarterly.* 34(2), pp. 213-238.

Coxhead, A. (2011). The Academic Word List 10 years on: research and teaching implications. In *TESOL Quarterly*, 45(2), pp. 355-362.

Gardner, D., Davies, M. (2013). A New Academic Vocabulary List. In *Applied Linguistics.* First published online August 2, 2013. Accessed at: applij.oxfordjournals.org [25/03/14].

Granger, S., Paquot, M. (2010). The Louvain EAP Dictionary (LEAD). In *Proceedings of the XIV EURALEX International Congress, 6-10 July 2010.* Leeuwarden, The Netherlands, pp. 321-326.

Hyland, K., Tse, P. (2007). Is There an "Academic Vocabulary"?. In *TESOL Quarterly.* 41(2), pp. 235-253.

Kosem, I. (2008). Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy? In *Proceedings of the XIII EURALEX International Congress,15-19 July 2008.* Barcelona, Spain, pp. 1575-1584.

Kosem, I. (2010). Designing a model for a corpus-driven dictionary of Academic English. PhD thesis. Aston University, Birmingham, UK.

Longman Dictionary of Contemporary English. 5th edition (2009). Harlow: Pearson Education Limited.

Oxford Advanced Learner's Dictionary. 8th edition (2010). Oxford: Oxford University Press

Oxford Learner's Dictionary of Academic English (2014). Oxford: Oxford University Press

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis.* London, New York: Continuum International Publishing Group, pp. 11-17, 55-61.

West, M. (1953) *A General Service List of English Words.* London: Longman, Green and Co.

The Danish Thesaurus: Problems and Perspectives

Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen
Society for Danish Language and Literature
sn@dsl.dk, ltj@dsl.dk, hl@dsl.dk

Abstract

In this paper, we present a new thesaurus for Danish and discuss some of the problems and decisions that the compilers have been faced with. The thesaurus is compared to other thesauri: Roget, Dornseiff and particularly to its smaller Danish predecessor Andersen. The different steps in the compilation process are outlined, with special attention being devoted to word ordering at the lowest level (alphabetical vs. semantic) and to the use of stylistic labels. In a comprehensive thesaurus, the conclusion is that semantic ordering is more useful for the user and that stylistic labels are necessary.
Keywords: thesaurus; onomasiological dictionary; semantic ordering; stylistic labels

During the last four years, a new Danish Thesaurus (*Den Danske Begrebsordbog*, DDB) has been compiled at the Society for Danish Language and Literature (DSL). Funded by the Carlsberg Foundation, the thesaurus is the first of its kind in 70 years and will at first be published as a printed book. The future plans are to publish it in an electronic version online at DSL's dictionary site *ordnet.dk* where it will be integrated with other dictionary resources at DSL. As it is completely based on, and systematically linked to, the approx. 100,000 lemmas and 135,000 word senses of the corpus-based The Danish Dictionary (DDO), we will have a thesaurus of modern Danish which contains all types of words, not least a wealth of compounds which is a common word type in a Germanic language like Danish. All types of fixed phrases and a number of frequent collocations from the DDO are given in the thesaurus. Well represented in the DDO, interjections and interjectional phrases are listed in the thesaurus together with semantically related verbs. Taxonomic vocabulary such as words for plants, animals, food, diseases, technical devices etc., is classified from a layman's point of view in common language, and all domains are thoroughly represented, including taboo areas. As an important side effect, the approach of common sense id numbers in the DDO and the DDB allows us to enrich a large number of the DDO lemmas and senses with supplementary information about synonyms and semantically related words in a future updated version of the DDO simply by making use of the thesaurus data. Experiments have likewise been carried out on the reuse of thesaurus data in the Danish WordNet, also based on the DDO and using the same sense id numbers (Nimb & Pedersen 2012; Nimb et al. 2013).

Unlike the comprehensive approach that we have adopted, some thesauri, for example the Norwegian Rosbach (2001) and our predecessor for Danish, *Dansk Begrebsordbog* (Andersen 1945, DB), focus instead on the core concepts of the language. We will discuss and compare these two approaches, focusing specifically on the difference between the DDB and the DB. Afterwards we will argue that a comprehensive approach has as consequence 1) that a semantic word ordering is necessary, and 2) that stylistic labels are necessary. But first a brief account of the different stages involved in the compilation of the DDB.

1 The Thesaurus-making Process

The basis of the compiling process of the DDB is an XML document representing all senses and their corresponding lemma forms in the DDO, supplied with all relevant semantic information from the dictionary: definition, domain, synonyms etc. On the basis of the sense unit document, the stock of words for the thesaurus is retrieved. As a case in point, consider the compilation of the section 'Bicycling'. First, the lexicographer retrieves all sense units where the string *cykel* is part of the definition and all units of which the corresponding lemma string begins or ends in *cykel*, picks out the relevant sense units and inserts them into the thesaurus document. Synonyms and near synonyms from the DDO as well as words discovered by introspection are added and linked to the DDO if not already part of its stock of lemmata. The selection of words are grouped into semantic categories, such as *persons* riding a bike (cyclist, rider etc.), concrete *objects* (bikes, saddles etc.), *events* (to cycle, to ride a bike etc.) or *properties* of the persons, objects etc., headed by the best representative of the group (annotated in the structure). They may also be put into groups where only a thematic relation holds between the words; this option is mainly used for cases that are difficult to categorize: single words with few near synonyms but nevertheless belonging to the thematic section. In this way, we make sure that not only the prototypical words and senses are covered but also the grey or peripheral areas of the vocabulary or the radial members of the categories established (cf. Dirven & Verspoor 2004: 33).

All the groups are tagged with formalized information about their type of category, and within each semantic group, the words are presented in a logical order according to semantic criteria (see below). In the digital document, semantics is the sole basis for classification. In a later phase, data is converted in order to present the material in four main divisions according to word class for the printed thesaurus (see below). But initially, verbs and their verbal nouns belong to one group tagged with the type 'act' and sometimes also followed by interjections related to the act. Words designating properties are grouped together, no matter if the property is expressed in the form of an adjective (*happy*), a noun (*happiness*), a verbal expression (*to tread/walk on air*) or an adverbial (*on cloud nine*), here exemplified by English words and expressions.

If we again turn to the theme of bicycles, the section 'Bicycling' is one out of 29 sections in the chapter 'Sports and leisure', which in turn is one of 22 chapters that make up the full thesaurus. In total there are 888 sections in the thesaurus – some chapters have only 20 sections while others have more than 60. The chapters and sections were based on the division found in Dornseiff's *Der deutsche Wortschatz nach Sachgruppen* (2004), but have been thoroughly adjusted in all the cases necessary. New sections were added for mental diseases, for contraception and abortion, for medicine, labour market and unemployment, to mention but a few. Other sections were removed, such as the German sections *Bote* (‘messenger’, instead included in the section on communication), and *Autoindustrie* (‘motor industry’), since they were either judged less important categories in modern Danish conceptualization or could easily be covered by other sections.

For various – not necessarily technical or logical – reasons, the grant was allocated for a printed dictionary. That is the reason why this paper is primarily concerned with the printed dictionary, but in all aspects of the compilation and editing process, data has been organized in a way that allows digital exploitation in a later phase, whether as an independent online dictionary or as an integrated semantic component of the DDO.

The manuscript for the printed thesaurus is produced on the basis of the thesaurus XML document. For each section, all words are extracted automatically and presented in four main groups according to word class (nouns, verbs, adjectives, remaining words) with clear marks (in the form of bullet points) of the shifts between semantic categories within each word class group, e.g. between the nouns for persons and the verbal nouns. The logical order within each semantic category, e.g. within the list of persons, respectively verbal nouns, is maintained in the text. The annotated initial words of each semantic group are automatically presented as keywords in the printed text. Some manual adjustment is still needed after the conversion, e.g. in the cases where a word end up being the only member of a group, or in case highlighted words conflict or have too large scope.

2 Comprehensive versus Restricted Thesauri

Where the aim of the DDB is to present all types of words in the DDO including many compounds, the editors of the predecessor DB chose to present only a restricted selection of Danish concepts. Andersen himself states that the vocabulary of the thesaurus was reduced to only a part of the vocabulary found in the contemporary monolingual Danish dictionary *Dictionary of the Danish Language* (ODS), and that he intentionally did not cover the vocabulary in detail, in contrast to the German thesauri which he also used as a model to improve the overall macro structure (Dornseiff’s 2nd edition from 1940 and 3rd edition from 1943), and also in contrast to the approach of the contemporary Swedish thesaurus (Bring 1930, cf. Andersen 1945: X-XI). This should be viewed in light of the fact that a large amount of already categorized words were at his disposal in a manuscript of c. 1,000 pages established during a period of 25 years by a schoolteacher in Copenhagen. The manuscript was considered suitable for publishing provided that a thorough revision and modernisation was carried out, for which Andersen was responsible. He left out a substantial amount of compounds and dialect words, as well as many words for natural objects and other concrete objects, the advantage of this being, he writes, a dictionary which is “klarere og mere hændig” (‘clearer and more handy’; Andersen 1945: X).

Another recent Scandinavian thesaurus, Rosbach (2001) has a limited stock of words for Norwegian. In our opinion, the restricted collection of words might be a reason why neither Andersen (1945) nor Rosbach (2001) have become widely used dictionaries in their respective countries, compared to the position of *Roget's Thesaurus* (2002) within the English-speaking community (Hüllen 2009: 40 and 44). Roget and also Dornseiff (2004) for German constitute valuable linguistic resources, offering a large variety of words and expressions as it is the main purpose of thesauri, namely to support language variation and provide linguistic inspiration to users in text producing situations (cf. Hüllen 2009: 29 and 46).

3 The Comprehensive Approach: Lexicographic Challenges

The comprehensive approach involves some lexicographic challenges since some of the methods of the restricted thesauri are not applicable when the contents increase, both in terms of the number of lexical units and the types of words and expressions included. The first challenge concerns the order in which the words are presented; the second concerns the treatment of words which are not part of the unmarked standard vocabulary.

3.1 Semantic versus Alphabetical Word Ordering

The DB organizes words in word classes which are divided into different semantic groups (objects, events, persons etc.) separated by dashes. Persons are always preceded by double dashes as the final element of a noun group. Within each semantic group (between dashes) words are listed in alphabetical order, meaning that the first word is not necessarily a keyword for the whole group. For instance, words within the category 'milk' are presented in the order "Fløde, Kærnemælk, Mælk, skummet Mælk, Piskefløde, Sødmælk, Yoghurt" (cream, buttermilk, milk, skimmed milk, double cream, whole milk, yoghurt). In the case of words for 'water', we find "Brøndvand, Drikkevand, Gaasevin, Isvand, Kildevand, Kommunevand, Vand" (well water, drinking water, plain water, ice water, spring water, tap water, water). In both cases, the hypernyms (*milk* and *water*) are placed between different types of hyponyms and do not function as keywords for the established semantic category.
Dornseiff (1940 and 1943) used the same alphabetical presentation. In the 8th edition (2004), though, we find an important difference to this: a keyword, typically the hypernym, has been moved from its place in the alphabetical order to the initial position of the group. The only cue to the keyword function is the break in the alphabetical order, which may be challenging for the user. Consider the case of words for different types of dairy products in Dornseiff (2004): "Milch · Buttermilch · Joghurt · Kefir · Magermilch · Rahm · Sauermilch" (milk, buttermilk, yoghurt, kefir, skimmed milk, cream, curdled milk). One notes also that the choice of alphabetical order inevitably breaks down logical ordering, placing *skimmed milk* between different types of sour milk rather than next to *milk*, of which it is a direct hyponym. Another example from Dornseiff (2004) concerns words for coffee: "Kaffee · Blümchen · Cappuccino · Espresso · Lorke · Milchkaffee · Mokka · MuckeFuck" (coffee, weak coffee (informal), cappuccino, espresso, bad weak coffee (dialect), café au lait, mocha, coffee substitute/ersatz coffee (deroga-

tory)), where one could argue that the informal word *Blümchen* and the dialect word *Lorke* have too prominent positions and would according to logic be better placed after the different types of coffee. Where a restricted thesaurus, such as the DB, can get away with an alphabetically ordered list of a small amount of words within a semantic category, this is in our opinion not suitable for the comprehensive thesaurus. The higher the amount of words, the more disturbing the alphabetical order becomes. This is the case in Dornseiff (2004) where a given word is most likely to be semantically more closely related to the initial keyword of the whole semantic category than it is to its immediate neighbours. In that way, the semantics of one word cannot be used to understand the next word and thereby activate a forgotten word in the user's mind. Furthermore, the distance from a word to the keyword can easily get very long. By contrast, Roget (2002) presents words only in semantic order and has in fact done so from the very first edition (Hüllen 2009: 40), a principle also adopted by Bring (1930). Initial keywords at the highest level of the taxonomy are graphically highlighted in Roget (2002) by italic types.

> *soft drink*, teetotal d., non-alcoholic beverage; water, drinking w., filtered w., eau potable, spring water, fountain; soda water, soda, ..., coffee, café au lait, café noir, black coffee, white coffee, decaffeinated coffee, decaf, Irish coffee, Turkish c., espresso, cappuccino, latte ...

Example 1: Roget's Thesaurus (2002), soft drinks (excerpt).

In the DDB, we implement the same type of semantic ordering as Roget. It relies entirely on the lexicographer's judgment (cf. Hüllen 2009: 29), following two principles which go hand in hand, as reflected in linguistic theories on prototypes. The first implies that the prototypical, or central, members of a category are presented before less prototypical, radial members (see for example Dirven & Verspoor 2004: 17 for a description). An example is the section in the DDB on furniture, where chairs and beds are placed before lamps and carpets. The second principle is based on the idea of basic level categories in a language, from which a division of the vocabulary into three conceptual levels may be derived: a generic, a basic and a specific level (Dirven & Verspoor 2004: 37). Following this principle, basic and general level terms will be placed before specific level terms in the thesaurus. In many cases, the general level term is used as the title of the section in question and the basic level terms are highlighted as keywords. Illustrated by English words, general level terms such as *animal*, *plant* and *furniture* constitute section titles and are in their corresponding sections presented before basic level terms such as *dog*, *tree* and *bed*. Both the general and the basic level terms (i.e. *animal* as well as *dog*) are marked as keywords and listed in the thesaurus before the specific terms, in this case kinds of dogs, trees or beds, for example *poodle*, *oak tree* and *double bed*. In the case of coffee, the DDB presents the words as seen in example 2.

kaffe, mokka (uformelt); espresso, café au lait, caffé latte, cappuccino, macchiato; filterkaffe, stempelkaffe, pulverkaffe, kolbekaffe, tyrkisk kaffe; sort kaffe; jordemoderkaffe (uformelt), mokka; en lille sort;

termokaffe; varmekaffe; mosevand (slang); en kop kaffe, kaffetår, refill; morgenkaffe, formiddagskaffe, eftermiddagskaffe, aftenkaffe

(**coffee**, mocha (informal); espresso, café au lait, caffé latte, cappuccino, macchiato; drip coffee, press pot coffee, instant coffee, coffee made in a coffee maker, Turkish coffee; black coffee; very strong coffee (informal), (a cup of) strong coffee; a cup of black coffee laced with spirits; thermo jug coffee; warmed-up coffee; bog water (slang); a cup of coffee, cup of coffee, refill; morning coffee, mid-morning coffee, afternoon coffee, evening coffee)

Example 2: Coffee words in the DDB.

An argument against the semantic ordering principle is that it becomes more difficult for users to find a specific, already known word as they cannot rely on the alphabet when browsing through a group. Instead, they must look via words that come closest in meaning in the text, and this is not always an easy task. To support overview and browsing as much as possible, we have chosen to highlight more keywords in the text than Roget does. In the case of soft drinks in example 1, also hypernyms at a lower level in the taxonymy will be highlighted, i.e. also *water*, *soda water* and *coffee*. Furthermore, keywords at the highest taxonomic level are presented in boldface in the printed DDB in order to obtain a clearer visual signal than the italics used in Roget. As a consequence, finally, the index is organized such that for each entry word the following information is given: entry, keyword, section indicated by its number, part-of-speech. This is similar to the solution in Roget and different from Dornseiff that refers only to section number and title.

3.2 Stylistic Labels

Comprehensive thesauri cover a much broader range of stylistic varieties, accentuating the need for stylistic labels. In this section we will discuss the types of information used in the DDB and compare it with DB, Roget and Dornseiff.

The DB does not have any information about register, for example we find several informal words for nose simply listed without any comment (*Mule* ‘muzzle’, *Næse* ‘nose’, *Næsebor* ‘nostril’, *Snabel* ‘trunk’, *Snude* ‘snout’, *Snydeskaft* ‘hooter, conk’, *Tryne* ‘snout’, *Tud* ‘snout, schnozzle’). Nor does Dornseiff use stylistic labels, which in our opinion sometimes leads to confusing lists of words, for example those concerning coffee mentioned above where dialect words (*Lorke* (bad weak coffee), informal words (*Blümchen* (weak coffee), and *MuckeFuck* (substitute/ersatz coffee) are presented alongside standard German words for coffee (*Kaffee, Cappuccino, Espresso, Milchkaffee, Mokka*). Roget, on the other hand, uses labels to some extent but gives no precise description as to the use of these. Since the DDB contains a substantial number of slang and informal words from the DDO, we have in line with Roget chosen to assign labels of the stylistic and temporal status to the words which are not part of the standard vocabulary. The labels we use are extracted from the information in the DDO but presented in simplified form. The detailed set of values in the DDO are here converted into five types: four stylistic (derogatory, informal, slang, jocular) and one temporal (archaic). Regional language is rare in the DDO and therefore not included in the set of labels. The method of direct transfer, however, is problematic, due to the fact that the labels were originally applied in relation to other senses of the same word and maybe a few synonyms given in the DDO entry, not to a large group of synonymous expressions as it is in the DDB. Manual adjustment is therefore needed in many cases. Only in the event where a few labelled words occur between numerous unmarked words in a semantic group, can they be kept as they are, without further editing. In other cases, we need to adjust the text in order to achieve homogeneous information about linguistic style. This is particularly relevant where the labels of adjacent words clash when seen in context. In these cases, we harmonize the information by choosing one label to cover both, if possible. For example, *lampe* ('lamp') and *pære* ('light bulb') are two expressions for 'brain' in Danish, the former labelled 'slang', the latter 'informal' in the DDO. In the DDB, both are labelled 'informal' as there is no clear-cut boundary between slang and informal language. In the case of large groups of synonymous expressions of the same stylistic value, the label of the initial keyword indicates that the following words have the same value. Information about temporal status is treated independently of other labels. Example 3 shows the case of derogatory words for women in Danish, before (1) and after (2) editing.

(1) **kælling** (neds.), kone, gås (neds.), tante (neds.), tøs (neds.), hundyr, furie, rappenskralde (neds.), strigle (neds.), mokke, hystade, skude (neds.), sæk, (neds.), tudse (neds.), so (neds.), smatso (neds.), klidmoster (neds.), kran (slang), madamme (neds.), sladretaske (neds.), rendemaske (slang, gammeldags), sladderkælling, havgasse (neds.), harpe (neds.), ribs (neds.)

(**bitch** (derogatory), woman, silly goose (derogatory), aunt (derogatory), hussy (derogatory), female animal, fury, shrew (derogatory), termagant (derogatory), fishwife (derogatory), battleaxe (derogatory), virago (derogatory), cow, hag (derogatory), gadabout, .. etc.

(2) **kælling** (neds.), gås, tante, tøs, hundyr, madamme, klidmoster, furie, mokke, harpe, strigle, ribs, hystade, rappenskralde, havgasse, sladderkælling, sladretaske, rendemaske (gammeldags), skude, kran, sæk, tudse, so, smatso

bitch (derogatory), silly goose, aunt, hussy, female animal, fury, shrew (old-fashioned), termagant, fishwife, battleaxe, virago, cow, hag, gadabout, .. etc.

Example 3: Derogatory words for 'woman' in the DDB. The group is initiated by one label instead of keeping the automatically inserted labels from the DDO on each word.

A manual adjustment is required in approx. 1/8 of the 888 semantic groups in the DDB, concentrated in certain semantic areas such as men, women, drinking alcohol, body parts, sexuality and bodily functions, but also physical punishment, conflict, scolding and others. For the larger part of the DDB, however, stylistic labels occur only sparsely and are typically kept the way they appear in the DDO.

4 Conclusion

It is a challenge to compile a comprehensive thesaurus which truly reflects the vocabulary of a semasiological monolingual dictionary. Bringing such a project to a completion within a period of just a few years' time can only be successful by using computational methods and by means of a well-structured model which guides the lexicographer's categorization of the words and, maybe most importantly, offers ways of placing the many radial concepts of the language. The alphabetical ordering of the words adopted by previous works is impracticable and must be replaced by semantic guidelines to ensure a consistent logical order within the large vocabulary of each category. The close connection between the two dictionaries makes it possible to reuse data in various ways. In this paper, we have shown how stylistic labels from the dictionary can be transferred to the thesaurus, and in the future our plan is to extract information about the semantic relations between words in the opposite direction, from the thesaurus into the dictionary, in this way adding an onomasiological component to the way users may access dictionary data.

5 References

Andersen, Harry (1945). *Dansk Begrebsordbog*. København: Munksgaard.

Bring, S. C. (1930). *Svenskt ordförråd ordnat i begreppsklasser*. Stockholm: Hugo Gebers Förlag.

DDO = *Den Danske Ordbog*. Accessed at: http://ordnet.dk/ddo [11/04/2014].

Dirven, Rene & Marjolijn Verspoor (2004). *Cognitive Exploration of Language and Linguistics*. Philadelphia, PA, John Benjamins Publishing Company, USA.

Dornseiff, Franz (2004). *Der deutsche Wortschatz nach Sachgruppen*, 8. Auflage, Berlin/New York: Walter de Gruyter.

Dornseiff, Franz (1943). Der deutsche Wortschatz nach Sachgruppen, 3. Auflage.

Dornseiff, Franz (1940). Der deutsche Wortschatz nach Sachgruppen, 2. Auflage.

Hüllen, Werner (2009). Dictionaries of synonyms and thesauri. In A. P. Cowie: *The Oxford History of English Lexicography, vol. II, Specialized Dictionaries*. Oxford: Oxford University Press, pp. 25-46.

Nimb, S. & B. S. Pedersen (2012). Towards a richer wordnet representation of properties – exploiting semantic and thematic information from thesauri. In *LREC 2012 Proceedings*. Istanbul, Turkey, pp. 3452-3456.

Nimb, S., B. S. Pedersen, A. Braasch, N. H. Sørensen & T. Troelsgård (2013). Enriching a wordnet from a thesaurus. In Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA). Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740).

ODS = *Ordbog over det danske Sprog*. Vol. 1-28 (1918-1956). Det Danske Sprog- og Litteraturselskab og Gyldendal. Online version at: http://ordnet.dk/ods [11/04/2014].

Roget, Peter Mark (2002). *Roget's Thesaurus*, 150th anniversary edition edited by George Davidson. London: Penguin.

Rosbach, Johan Hammond (2001). *Ord og begreper. Norsk tesaurus*. Oslo: Pax Forlag A/S.

From a Dialect Dictionary to an Etymological One

Vilja Oja, Iris Metsmägi
Institute of the Estonian Language
vilja.oja@eki.ee, iris.metsmagi@eki.ee

Abstract

Typically, loanwords from different sources are presented in different entries of a dialect dictionary, but the etymologies of the dialect words are often obscure. An etymologist, on the other hand, has to consider the phonetic shape and developments of a stem, its areal distribution and meanings in dialects to find out the etymology. What are the cooperative prospects of etymologists and dialectologists? The paper compares the presentation of dialect words in two dictionaries currently compiled on the web, namely, in the Estonian dialect dictionary and in the Estonian etymological dictionary. Despite the different specifics of the two dictionaries a number of similar problems have cropped up. Also, comparisons of the material have yielded essential information enabling solutions for both sides. Across dialects, a loanword often displays numerous phonetic variants, while a variant may easily sound untypical of the concrete dialect. The possible donor can be traced considering the occurrence of the dialect word in the traditional area of loanwords of a certain origin and the semantic relationship of the word with the presumed donor language. Cooperation between etymologists and dialectologists has contributed a lot to making a distinction between homonymy and polysemy, to identification of folk etymologies (words as well as semantic nuances), to distinguishing between separate loanwords and derivatives of the same stem, etc. A shared electronic environment enables bilateral specification of the linguistic material, if necessary.

Keywords: dialect vocabulary; etymology; homonymy; polysemy; semantic change; Estonian

1 Introduction

The study is focused on the presentation of dialect words in an etymological dictionary. In 2003 the Institute of the Estonian Language (IEL) launched the project of an Estonian Etymological Dictionary. The first edition of the Estonian etymological dictionary (EES – *Eesti etümoloogiasõnaraamat*) compiled at the Institute of the Estonian Language was published in spring 2012. This is an approximately small dictionary for a wide readership, with about 6600 entries. The entry list of the EES was based on the word stems contained in the Estonian normative dictionary (ÕS 2006). Thus it includes the stems of standard Estonian and a small selection of dialect vocabulary (see Metsmägi 2010). The work has revealed that quite a significant part of Estonian word stems still lack a satisfactory scholarly explanation of their origin. Now, an extended and much more thorough edition is being prepared on the basis of the EES. The entry list of the new Estonian etymological dictionary (henceforth: EED) will be augmented by about 1500 additional stems: dialect words and other older stems (e.g. words denoting obsolete tools, archaic occupations etc.) and the latest loanwords. The large Estonian etymological dictionary previously available, the *Estnisches etymologisches Wörterbuch* by Julius Mägiste (1982–1983, 2nd edition 2000; 12 volumes, 4106 pp.; EEW) contains dialect words as well. Unfortunately, the author was unable to finish it or edit due to his death. For that reason, there are numerous misprints and defective entries. Another reason for the dictionary being incomplete is that in exile the author ceased to have access to the materials kept in Estonia.

Standard Estonian was developed on the basis of local dialects. Estonian dialects, however, are based on the vernaculars of several Finnic tribes, not just one. The evolution of the dialects was affected by both the local conditions and various socio-historical factors. Conscious and deliberate development of Estonian as a standard language for the whole Estonian territory was only started in the 19th century (see, e.g. Kask 1984). Estonian vocabulary cannot be unambiguously divided into dialect and standard words. Often a word has different meanings in different dialects, which – like unusual phonetics – may help in tracing the origin of the word. In addition, dialects are a treasury of archaic terms and obsolete expressions that do not belong to the modern standard. Some hints on the origin of Estonian dialect words can be found in the Estonian-German dictionary compiled by F. J. Wiedemann in the 19th century (Wiedemann 1973). The etymology of Estonian dialect vocabulary has also been discussed in monographic studies of loanwords (e.g. Ariste 1933, Koponen 1998, Must 2000, Vaba 1997).

The Archive of the Estonian dialects and Finno-Ugric languages (EMSUKA) at the Institute of the Estonian Language includes a collection of Estonian dialect vocabulary, which contains about three million paper slips carrying phonetic, morphological and semantic data of dialect words as well as usage examples. This lexical collection serves as basis for the Estonian dialect dictionary (EMS – *Eesti murrete sõnaraamat*) currently compiled. So far, 25 fascicles (5085 pages) of the EMS (*a – matkama*) have been published. As preliminary work for the EMS a concise dialect dictionary (VMS – *Väike murdesõnastik*) was compiled in the 1980s. The VMS represents a list of possible entry words for the EMS, together with their areal distribution and the approximate meaning of the non-standard words. The dictionary has been published both in print and electronically. At present both the Estonian dialect dictionary and the Estonian etymological dictionary are being compiled online using the EELex program created at the Institute of the Estonian Language. The program enables representation of database material in a book layout form. Both dictionaries will be published electronically. The electronic version enables linking the dictionaries with the rest of the online dictionaries of the Institute of the Estonian Language.

2 Word Selection

The dialect words to be added to the new Estonian etymological dictionary (EED) are drawn from the material stored in the Estonian dialect dictionaries EMS and VMS. The data will be elaborated using the collections of the Archive of the Estonian dialects (EMSUKA), if necessary. One major dilemma that relates to the preparation of any dictionary is what words to choose, what ones to leave out. The EMS is exhaustive, including the entire vocabulary available from all Estonian dialects. Some words have been recorded from a few sub-dialects only while some others are known practically throughout the Estonian territory. Limits have been set on the presentation of words not specific to dialect usage, e.g. foreign words, new standard words, special terms, slang and nursery language (see Must 1968, Oja 1996).

In the first place the entry list of the new Estonian etymological dictionary is to be supplemented by dialect words representing lexical peculiarities of larger regions, such as South Estonian dialects[1], e.g. *kurst* 'twisted handful of flax, hair, straw etc.' < Latvian *gùrste* (Vaba 1997: 107); *kuuas* 'axe handle' < Latvian dial. *kuôts* 'handle of a tool etc.' (Vaba 1997: 108-109), or the Insular dialect of Estonian, e.g. *kurt* '(fir or pine) cone' < Estonian-Swedish[2] *kott*, *kotte*, Old Swedish *grankuttar* '(fir or pine) cone', cf. Finnish-Swedish *kort* id. (Ariste 1933: 69).

Old genuine stems, missing in the standard vocabulary but found in the dialects will be added to the new etymological dictionary even if sparsely recorded. For example, a rather widespread dialect word *kosk* 'thick bark' is possibly an ancient (Uralic) genuine stem to which Ugric and Samoyed equivalents have been suggested (SKES 222; SSA 1: 409–410; UEW 179–180). The word *keri* ~ *kere* 'linden bark' used in some Estonian dialects (EMS II: 1026) belongs to the Finno-Ugric layer of the genuine stems (SKES 183; SSA 1: 345; UEW 148–149). The dialect word *kärg* 'woodpecker' has etymological equivalents in Finnic and Volgaic languages (SKES: 261; EEW: 1138; UEW: 652; SSA 1: 476). The genuine stems with obvious etymological equivalents only in Finnic as the closest cognate languages will be added to the entry list as well.

On the other hand, the entry list will be augmented by words that clearly testify to being borrowed. Among the dialect words there are some old loanwords that were borrowed in the period before Estonian had become a separate language, i.e. the Indo-European, Indo-Aryan, Baltic, Germanic, Scandinavian and Slavic (Old Russian) loanwords. In general, they have etymological equivalents in other Uralic resp. Finnic languages. For example South Estonian *mehiläne, mihiläne* or *mehine* 'bee' (see VMS 2: 22) and the other Finno-Ugric words with the same stem (e.g. Finnish *mehiläinen*, Hungarian *méh*, *mihe* etc.) have been borrowed from Proto-Indo-European or Proto-Indo-Iranian (SKES: 339; SSA 2: 156; UEW: 271). Another word for bee, *mesilane* and variants (VMS 2: 25), used in other dialects as well as in standard Estonian is a derivative from the noun *mesi* 'honey', which is another Indo-European or In-

[1] About grouping the Estonian dialects see, e.g. Pajusalu 2003: 231.

[2] In Middle Ages the coasts and islets of Estonia were populated by Swedes. Their dialect is called Estonian-Swedish (see e.g. Blumfeldt 1961).

do-Iranian loanword (EES 280–281; UEW: 273; SSA 2: 161). The noun *rend* used for a long dining table in the Insular dialect has been borrowed from Baltic via the now extinct Curonian language *rend* 'table' (Vaba 2009: 779–780). The Insular dialect word *vada* 'seine net' (in close cognate languages: Karelian and Veps *vada*, Livonian *vadā*, Finnish *vata*) has been borrowed from Proto-Germanic, cf. Swedish, Norwegian *vad*, Old Norse *vaðr*, Middle Low German (MLGm) *wade*, German *Wate* (LÄGLOS III: 381; SKES: 1671–1672; SSA 3: 417–418).

In South-East Estonian and Eastern dialects the word *mugel* (*mugla, mukl*) stands for 'spent lye' (Wiedemann 1973: 609; VMS 2: 35). This is a Russian loanword borrowed from Old Pskov dialect (EEW: 1558; Koponen 1998: 127; Ojansuu 1922: 139). In North and Central Veps the same word *mugl* stands for 'lye'. The term for 'soap' used in eastern Finnic languages is a more recent loanword derived from the same Russian stem: Finnish, Karelian, Veps, Ingrian, Votic *muila, muil* etc. < Russian *mýlo* (ALFE 1: 205–206; Kalima 1952: 123–124).

As a rule, the transparent loanwords belong to some group of younger loanwords, borrowed only after Estonian had become a separate language, i.e. Low German, (High) German, Swedish, Russian, Latvian and Finnish loanwords. For example *kink* 'haunch; ham' (EMS III: 165) < MLGm *schink(e)* 'ham' (SKES I: 195); *tohv* 'kind of cloth' < German *Stoff* 'material; textile' (EEW: 3202); *turslag* 'strainer' < German *Durchschlag* 'strainer; sieve' (EEW: 3373); *tutspard* 'moustache' < German *Stutzbart* ' tile beard' (EEW: 3383); *kisla* 'very sour, over sour' < Russian *kíslyj* 'sour' (Must 2000: 97–98); *robutama* 'work quickly and intensively, toil, drudge; work quickly but carelessly' < Russian *rabótat'*, dial. *rabotát'* 'to work' (EEW: 2510; Must 2000: 333–334).

It is not uncommon that a stem of Indo-European origin has been borrowed into different Finnic languages and even to the dialects of the same language in different times and via different routes. Sometimes the words of the same origin have been borrowed repeatedly. For example North-Estonian *tulk* 'interpreter, translator' was borrowed from Old Russian, but the Finnish *tulkki* is a Scandinavian loan and the standard Estonian word *tõlk* id., is a newer loan from Russian (SKES: 1391; EEW: 3350; SSA 3: 324).

3 Entry Structure

In both dictionaries, an entry contains the following components: headword, grammatical information of the stem, dialectal variants with information on the regional distribution and word meaning or meanings. In addition the EMS gives usage examples of the word from different dialects. The etymological dictionary presents information on the origin of the word stem and the equivalents of the stem in cognate languages, with comments and bibliography. Both dictionaries make ample use of cross references and reference entries or subentries. In the electronic version, the references function as links to the referred entries.

In both dictionaries the headword is either a standard word or a dialect word in a standardized shape. The main form of an Estonian noun is nominative singular (or nominative plural in the case of *plurale tantum*) and the lemmatic form of a verb is the infinitive ending in *-ma*. In both dictionaries the headwords are ordered alphabetically, but in the EMS the place of *h*-initial words depends on the vowel following *h*-: *a*- (*ha*-), *e*- (*he*-), etc. as word-initial *h*- is lacking in most Estonian dialects.

The entry lists of both dictionaries have been built up on stem basis, in the way that the stems having different etymologies go to separate entries. In the dialect dictionary (EMS) the predictable phonetic variants of a word are all in the same entry, but irregular variants are presented in different ones. By way of exception, some irregular variants of loanwords are found together in the same entry (in more detail see Neetar 1992). In the etymological dictionary (EED) an entry will cover all words or stems originating in a common etymological source, i.e. derivatives, lexicalized inflected forms and stem variants (synchronic as well as diachronic ones). Only the separately borrowed derivatives of a stem, i.e. derivatives containing affixes of the donor language get separate EED entries. In the EMS, separate entries are provided for each suffixed derivative as well as for the compounds included. In the EED, only the compounds borrowed as a single item are given in separate entries, e.g. *leierkast* 'barrel organ' < German *Leierkasten*.

According to the specifics of either dictionary their main emphasis lies on different aspects. The major part of an EMS entry deals with the details of the dialectal variants of the word and with its use in dialects: meanings, sub-senses and examples. If necessary, a semantic group is provided with subheads for figurative senses, idioms and phrases. The etymological dictionary is focused on the origin of word stems. The etymology is described in the EED by cognate language equivalents and/or the loan source. The entry of EED will also be supplied with a bibliographical component, containing a survey of the etymological treatments of the entry stem (i.e. the references).

4 Some Crucial Problems

One of the trickiest issues facing the authors of either dictionary is classification of the linguistic material, in particular, choice of the entry word. Many similar stems make one wonder which of the words should be presented in the same entry and which should be given separate entries. In dialect dictionary, a polysemantic word gets several entries, the entry word *nukk,* for example, starts seven articles (VMS 2: 105). Collocations are usually presented under several entries (see Oja 1996). In an etymological dictionary, it is essential to discriminate between words originating from the same loan source and those originating from different ones. Thus a detailed analysis of dialectal variants and meanings is required. The different meanings of a word as well as different phonetic variants may originate from different sources. The sources in their turn may be mutually connected, e.g. there are parallel loans from different Germanic languages – Middle Low German, (High) German, and Swedish. Actually the situation is even more complicated, because the words borrowed from the local German dialect, the so-called Baltic German, and from the local Swedish dialect called Estonian Swedish, should be treated as separate loanword groups as well. In addition, sometimes different meanings have been borrowed into Estonian from the same source but in different times. In these cases the word variants or meanings will be discussed in separate EED entries.

4.1 Polysemy and Homonymy

In a general case, the senses of a polysemantic word are presented in one and the same entry in the Estonian etymological dictionary, but in the dialect dictionary they are sometimes found under different entry words. The latter case applies, e.g. to words *koot*^{1} 'flail, agricultural tool for manual threshing; mobile part of various objects' (EMS III: 635–636) and *koot*^{2} 'part of animal foot; part of human foot (lower part of the leg; thigh); usually pl., facet. human feet'(EMS III: 636–637). Notably, the two senses differ quite radically in their geographic distribution: *koot*^{1} mainly occurs in South Estonian dialects, and also in the Central and Eastern dialects and in the Pärnumaa region, but it is practically absent in the Insular dialect, Läänemaa region and in the North-Eastern Coastal dialects, whereas *koot*^{2} occurs in the North Estonian areas, apart from single reports from South Estonian dialects. Etymologically speaking, the root is the same: < MLGm *kote*, *kute* 'ankle; hoof; pastern' (Saareste 1924: 204; EEW III: 946). Thus, its primary sense is 'part of leg', semantically transferred to 'manual threshing tool' in Estonian (Saareste 1924: 204). The geography of the word in the sense of a tool (*koot*^{1}) is eloquent of the spread of an innovative two-piece threshing tool: it was first introduced in southern Estonia (Mark 1932: 369–370, ERL: 96), while the older one-piece tool called *vart* 'threshing stick' was favoured the longest in the West-Estonian islands and in some places in western Estonia (Manninen 1929: 45–46; ERL: 342). The South-East Estonian *prunt's* (and variants) 'skirt' is a Latvian loanword < East-Latvian regional *bruņči* id. that probably originates from the colour word *brūns* 'brown' < MLGm *brūn* id. (Saareste 1924: 166; Vaba 1997: 168–169). As long as the Latvian dialect word was unknown the South Estonian term for skirt used to be associated with the Estonian dialect word *pruńt, prońt* 'pleat' (EEW: 2183).

4.2 Folk Etymology and Semantic Changes

There are several ways of loanword adaptation. Being uncertain of the real semantic background of foreign terms people often associate them with a similarly sounding familiar word. This way folk etymology may change the phonetic shape as well as the meaning of the loanword beyond recognition. Such loanwords tend to have exceptionally many dialect variants that are poorly motivated phonetically. Semantic change in loanwords, being caused by cultural differences, local specifics, taboo, etc., will complicate semantic analysis as well as detection of the origin of the word. Note that in dialects a word may be used in its original meaning lacking in the standard language or in a sense that is closer to the original meaning than standard usage (Oja & Metsmägi 2013).

For example, a wallet or purse may be humorously called *tengelpung* in Estonian. The components of the compound word have been associated with the well-known loanwords *teng* 'money' (< Russian *denga* id.) and *pung* arch. 'wallet, pouch' (< MLGm *punge* 'pocket' or < Swedish *pung* 'wallet, pouch'). Actually, the Estonian *tengelpung* (in dialects also *tenkelpuuh, tenkelpus* etc.) has been borrowed from the Baltic German dialect (< German *Denkelbuch* 'old style paper notebook, pocketbook' < *denken* 'think, believe, plan, imagine' + *Buch* 'book') (Ariste 1942: 20; Viires 1960: 158.)

The name of juniper (the plant in the genus *Juniperus*) is *kadakas* in Estonian. In the second half of the 19th and beginning of the 20th century a similar noun *kadakas* and the compound *kadaka/saks* (*saks* 'squire, vulg. German') or *kadaka/sakslane* (*sakslane* 'German') were used for Estonians who tried to look like Germans and spoke (usually incorrect) German (EMS II: 453). Although folk etymology would associate the Estonian words *kadakasaks, kadakasakslane, kadakas* '(half-) Germanized Estonian' with 'juniper', the disdainful words have nothing to do with the tree. Instead, it is a loanword borrowed from the German compound word *Katensaße* 'slum dweller, craftsman' (< *Kate* 'hut, shanty' + *Saße* 'place of residence'), which has been folk-etymologically modified to sound like certain familiar words (Saari 2004: 119–120). The word *katekismus* 'catechism' has dialect variants *katekeskmus*(*s*) and *katekeskmine* showing that the word has probably been folk-etymologically connected to the South-Estonian dialect word *kat's* (vocalic stem *kate*) 'two' and *kesk*(*mine*) 'middle; between'.

4.3 Derivatives or Separate Borrowings

Sometimes the question is if we have to do with a borrowed derivative, i.e., whether the derivative and the stem have been borrowed separately or not. The donor language may have been the source of the stem as well as of one or more of its derivatives. Some old Indo-European loanwords in the Finnic language group have undergone morphological and semantic adaptation to such an extent that they have come to be regarded as genuine native derivatives. For example, the structure of the Estonian *raudjas* 'russet' appears to be *raud* 'iron' + diminutive suffix -*jas*, but it is most likely a loan from a Baltic colour word, cf. Latvian *raũds, raũdis* 'reddish brown' from Proto-Indo-European **reud*^{h}-), whereas the term *raud* 'iron' is a Germanic loan from Proto-Germanic **rauðan-* (Oja 2004: 37–38).

Newer (esp. MLGm) loanwords display a lot of cases where a noun has been borrowed in parallel with a zero-affixed verb of the same stem, e.g. Estonian noun *kink* 'gift, present' < MLGm *schenke* 'act of giving; (welcome) present, etc.' (EEW: 834) and the verb *kinkima* 'make a present, donate, give' < MLGm *schenken* 'make a present, to give' (Ariste 1940a: 12; EEW: 834) or another example: *rööv* 'robbery' < MLGm *rōf* 'robbery; booty' (Ariste 1940b: 110) and *röövima* 'to rob' < MLGm *roven* 'to rob' (Ariste 1980: 34; SKES: 908; SSA 3: 122).

5 Conclusion

Word origin can be specified by following its dialectal variants as well as similar words in cognate languages and contact languages, considering both their areal distribution and meaning. The variation of the phonetic shape and the meaning of words in Estonian dialects may suggest different borrowing times or travelling routes. The areal distribution of loanwords helps to pinpoint the centres of cultural innovations. An available dialect dictionary is of great help to word etymologization and etymological lexicography, offering concentrated and systematized information on the areal distribution, phonetic variation, and meanings of words. However, its entry list is most extensive, containing the whole dialect vocabulary, including compound words and derivatives, and thus an etymological lexicographer has to make a selection including the following: (1) dialect words representing lexical peculiarities of larger regions; (2) old genuine words; (3) loanwords with an obvious source; (4) derivatives borrowed separately from the stem.

A common issue in compiling both dictionaries is the arrangement of the highly variable linguistic data, in the face of polysemy and homonymy, folk etymology and semantic changes. Depending on the specifics of the dictionary, the final solutions may differ for the dialect dictionary and for the etymological dictionary. In a dialect dictionary the material cannot be presented systematically enough without considering word etymology. Hence an etymological dictionary makes an effective supplement to a dialect dictionary, helping to understand the background of the diversity of meanings and phonetic variation. Thus the best policy would probably be parallel compilation of the two dictionaries in a close cooperation of both teams, which is, however, extremely problematic to organize.

6 References

ALFE = *Atlas Linguarum Fennicarum. Itämerensuomalainen kielikartasto. Läänemeresoome keeleatlas. Ostseefinnischer Sprachatlas. Лингвистический атлас прибалтийско-финских языков.* ALFE 1–3. Chief ed. T. Tuomi. Eds. S. Suhonen (1), T.-R. Viitso (2), V. Rjagoev (3). Suomalaisen Kirjallisuuden Seuran Toimituksia 800/1295. Kotimaisten kielten tutkimuskeskuksen julkaisuja 118/159. Helsinki: Suomalaisen Kirjallisuuden Seura, Kotimaisten kielten tutkimuskeskus 2004–2010.

Ariste, P. (1933). *Eestirootsi laensõnad eesti murretes. Die Estlandschwedische Lehnwörter in der estnischen Sprache.* Acta et Commentationes Universitatis Dorpatensis B XXIX. Tartu.

Ariste, P. (1940a). *Georg Mülleri saksa laensõnad* [Georg Müller's German loanwords]. Acta et Commentationes Universitatis Dorpatensis B XLVI.1. Tartu.

Ariste, P. (1940b). Saksa laensõnadest 16. sajandi eesti kirjakeeles [About German loanwords in the 16th century literary Estonian]. In *Eesti Keel* 3–4. Tartu, pp. 108–112.

Ariste, P. (1942). Etümoloogilisi märkmeid [Etymological notes] I. In *Acta et Commentationes Universitatis Dorpatensis* B XLIX. 1. Tartu, pp. 1–26.

Ariste, P. (1980). Deutsche Lehnwörter im Wotischen. In *Specifičeskie osobennosti leksiki i grammatiki ural'skih jazykov.* Acta et Commentationes Universitatis Tartuensis 517. Fenno-ugristica 6. Tartu, pp. 27–38.

Blumfeldt, E. (1961). Estlandssvenskarnas historia [History of Estonian Swedes]. In: *En bok um Estlands svenskar* I. Stockholm: Kulturföreningen Svenska Odlingens Vänner, 63–178.

EES = Metsmägi, I., Sedrik, M. & Soosaar, S.-E. (2012). *Eesti etümoloogiasõnaraamat* [Estonian Etymological Dictionary]. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

EEW = Mägiste, J. (1982–1983). *Estnisches etymologisches Wörterbuch.* Helsinki: Finnisch-Ugrische Gesellschaft.

EMS = *Eesti murrete sõnaraamat* [Estonian Dialect Dictionary] (1994–2014). I–V (Fascicles 1–25), Eds. A. Haak, E. Juhkam, M.-L. Kalvik, M. Kendla, T. Laansalu, V. Lonn, H. Neetar, E. Niit, P. Norvik, V. Oja, V. Pall, E. Ross, A. Sepp, M.-E. Tirkkonen, J. Viikberg. Tallinn: Eesti Teaduste Akadeemia, Eesti Keele Instituut.

ERL = Troska, G., Viires, A., Karu, E., Vahtre, L., Tõnurist, I. (2007). *Eesti rahvakultuuri leksikon.* Ed. A. Viires. 3., corrected and supplemented edition. [1. edition 1995, 2. edition 2000.] Tallinn: Eesti entsüklopeediakirjastus.

Kask, A. (1984). Eesti murded ja kirjakeel [*Estonian dialects and literary language*]. Eesti NSV TA Emakeele Seltsi toimetised 16. Tallinn: Valgus.

Koponen, E. (1998). *Eteläviron murteen sanaston alkuperä. Itämerensuomalaista etymologiaa* [The Origin of the South-Estonian Dialect Vocabulary. Finnic etymologies]. Suomalais-Ugrilaisen Seuran Toimituksia 230. Helsinki: Suomalais-Ugrilainen Seura.

LÄGLOS = Kylstra, A. D., Hahmo, S.-L., Hofstra, T. & Nikkilä, O. (1991–2012). *Lexikon der älteren germanischen Lehnwörter in den ostseefinnischen Sprachen.* Amsterdam–Atlanta–New York: Rodopi.

Manninen, I. (1929). Übersicht der ethnographischen Sammelarbeit in Eesti in den Jahren 1923–1926. In *Õpetatud Eesti Seltsi Aastaraamat. Sitzungsberichte der Gelehrten Estnischen Gesellschaft* 1927. Tartu: C. Mattiesen, pp. 31–47.

Mark, J. (1932). Über das Roggendreschen bei den Esten. Festvortrag, gehalten am 19. Januar 1931, dem 93. Jahrestage der Gesellschaft. In *Õpetatud Eesti Seltsi Aastaraamat. Sitzungsberichte der Gelehrten Estnischen Gesellschaft* 1931. Tartu: Õpetatud Eesti Selts. Gelehrte Estnische Gesellschaft.

Metsmägi, I. (2010). Dialect materials in the Estonian Etymological Dictionary. In *Slavia Centralis* III, 1, pp. 196–204.

Must, M. (1968). Über die Arbeiten am estnischen Dialektwörterbuch. In *Congressus Secundus Internationalis Fenno-Ugristarum Helsingiae habitus 23.–28. VIII 1965.* Pars I, Acta Linguistica. Helsinki: Societas Fenno-Ugrica, pp. 348–351.

Must, M. (2000). *Vene laensõnad eesti murretes* [Russian Loanwords in Estonian Dialects]. Tallinn: Eesti Keele Sihtasutus.

Neetar, H. (1992). Etymologisches im estnischen Dialektwörterbuch (EDW). In *Euralex '92. Proceedings I–II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland.* Tampere, pp. 607–614.

Oja, V. (1996). Word Combinations in the Estonian Dialect Dictionary. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (eds.). *Euralex '96 Proceedings I–II. Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden.* Göteborg: Göteborg University, pp. 443–449.

Oja, V. (2004). Some colour words with restricted reference. In *Latvijas Zinātnu Akadēmijas Vēstis.* A daļa. Sociālās un humanitārās zinātnes 5, pp. 37–42.

Oja, V., Metsmägi, I. (2013). Laensõnade tähendussuhetest [Semantic relations of loanwords]. In H. Metslang, M. Langemets, M.-M. Sepper (eds.) *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics* 9. Tallinn: Eesti Rakenduslingvistika Ühing, pp. 181–194.

Ojansuu, H. (1922). Eesti etümoloogiad [Estonian etymologies]. In *Eesti Keel* 4–5, pp. 137–139.

Pajusalu, K. (2003). Estonian dialects. In *Estonian Language.* Ed. M. Erelt. Linguistica Uralica supplementary series 1. Tallinn: Estonian Academy Publishers, pp. 231–272.

Saareste, A. (1924). *Leksikaalseist vahekordadest eesti murretes. Du sectionnement lexicologique dans les patois estoniens* I. Acta et Commentationes Universitatis Dorpatensis B VI.1. Tartu.

Saari, H. (2004). *Keelehääling. Eesti Raadio "Keeleminutid" 1975–1999.* Tallinn: Eesti Keele Instituut, Eesti Keele Sihtasutus.

SKES = Toivonen, Y. H., Itkonen, E., Joki, A. J. & Peltola, R. (1955–1978). *Suomen kielen etymologinen sanakirja* [Etymological Dictionary of Finnish]. Lexica Societatis Fenno-Ugricae XII. Helsinki: Suomalais-Ugrilainen Seura.

SSA = *Suomen sanojen alkuperä. Etymologinen sanakirja* [The Origin of Finnish Words. An Etymological Dictionary] (1992–2000). Eds. E. Itkonen, U.-M. Kulonen. Suomalaisen Kirjallisuuden Seuran toimituksia 556. Kotimaisten kielten tutkimuskeskuksen julkaisuja 62. Helsinki: Suomalaisen Kirjallisuuden Seura, Kotimaisten kielten tutkimuskeskus.

Toivonen, Y. H. (1928). Zur Geschichte der finnisch-ugrischen inlautenden Affrikaten. In *Finnisch-Ugrische Forschungen* XIX, 1–270.

UEW = Rédei, K. (1986–1988). *Uralisches etymologisches Wörterbuch.* Budapest: Akadémiai Kiadó.

Vaba, L. (1997). *Uurimusi läti-eesti keelesuhetest* [Studies on Latvian-Estonian Linguistic Relations]. Tallinn–Tampere: Eesti Keele Instituut, Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitos.

Vaba, L. (2009). *Rend* ja *laud*. Kisklauast söögilauaks [*Rend* and *laud*. From a Split Log to a Dining Table]. In *Keel ja Kirjandus* LII (10), pp. 779–784.

Viires, A. (1960). *Hundilaut* ja *tengelpung*. In *Keel ja Kirjandus* III (3), pp. 157–158.

VMS = V. Pall (ed.) *Väike murdesõnastik* [Concise dialect dictionary] (1982, 1989). I, II. Tallinn: Valgus.

ÕS 2006 = Erelt, T., Leemets, T., Mäearu, S. & Raadik, M. *Eesti õigekeelsussõnaraamat 2006* [Estonian Normative Dictionary 2006]. Ed. T. Erelt. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.

Wiedemann, F. J. (1973 [1869]). *Eesti-saksa sõnaraamat. Estnisch-deutsches Wörterbuch.* Vierter unveränderter Druck nach der von Jakob Hurt redigierten [2.] Auflage [1893]. Tallinn: Valgus.

Acknowledgements

The study has been funded by the Estonian Ministry of Education and Research (SF0050037s10) and Estonian Science Foundation (ETF9367).

Research on Dictionary Use

Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernern

Carolina Flinz
Universität Pisa
c.flinz@ec.unipi.it

Abstract

Die vorliegende empirische Untersuchung befasst sich mit einer Umfrage zur Wörterbuchbenutzung bei 41 Studentinnen und Studenten des *Dipartimento di Filologia, Letteratura e Linguistica* der Universität Pisa, dasselbe Department, an dem auch das deutsch-italienische sprachwissenschaftliche Online-Wörterbuch DIL erarbeitet worden ist (vgl. Flinz: 2011). Die schriftliche Umfrage wurde in Anlehnung an Hartmanns 5. Hypothese *„An analysis of users´ needs should precede dictionary design"* (1989) durchgeführt. Die wichtigsten Ergebnisse waren von großer Bedeutung für die Gestaltung der makro- und mikrostrukturellen Eigenschaften des Fachwörterbuches. Die Ergebnisse der Untersuchung und die daraus folgenden Reflektionen werden in thematischen Kernblöcken vorgestellt.

Keywords: Wörterbuchbenutzung; Umfrage; Fachwörterbücher; Online-Fachwörterbücher

1 Einleitung

Ziel des Beitrags ist es, die Ergebnisse einer schriftlichen Umfrage zu präsentieren, welche der Frage nachgeht, in welchen Situationen und zu welchen Zwecken italienische DaF-Lernende zweisprachige Sprachwörterbücher und insbesondere sprachwissenschaftliche Fachwörterbücher verwenden. Ferner soll untersucht werden, welche Einstellungen die Testpersonen über Online- und Printprodukte haben, welche Eigenschaften der Printprodukte sie positiv beurteilen und welche nicht. Nach einem Exkurs über den theoretischen Hintergrund (§2) der Untersuchung, die Wörterbuchbenutzung, werden die Analyse und die Kernblöcke der Umfrage vorgestellt (§3). Die wichtigsten Ergebnisse werden anschließend vorgestellt und diskutiert (§4). Abschließende Schlussfolgerungen beenden den Beitrag.

2 Stand der Forschung

Wörterbuchbenutzung als wissenschaftlich fundierte Disziplin hat sich im Rahmen der Lexikographie in den 80er Jahren entwickelt. Unterschiedliche Artikel und Beiträge zu diesem Thema (Hausmann et al. 1989; Ripfel/Wiegand 1988; Rossenbeck 2005; Welker 2010; Wiegand 1987/1998/2008/2010) bestätigen dieses wachsende Interesse, obgleich es laut Wiegand trotz seiner Bedeutung für die Erschaffung einer größeren Benutzereffizienz (1987:179) der am wenigsten erforschte Bereich der Lexikographie bleibt (2008:1). Neue Wörterbücher oder Neuauflagen sollten aus der wissenschaftlichen Erkenntnis dieser Praxis entstehen und einen höheren Nutzungswert haben (Wiegand 1987:179).

Die Forschung zur Online-Wörterbuchbenutzung ist hingegen noch in ihren Anfängen (vgl. Nesi 2000:845; Simonsen 2011:7) und die im Abstand von drei Jahren am IDS durchgeführte Studie „User-adaptive access and cross-references in elexiko (BZV elexiko)“ setzt es sich zum Ziel, diese Lücke zu füllen. Online-Wörterbücher sollten sich empirischer Analysen bedienen, die aufzeigen, welche spezifischen Gebrauchssituationen und Bedürfnisse vorhanden sind, wie die Wörterbücher tatsächlich benutzt werden und wie sie benutzerfreundlicher gestaltet werden könnten (vgl. Atkins/Varantola 1998; Hartmann 2000; Spitzer/Koplenig/Töpel 2012).

Vereinzelt sind Untersuchungen zu unterschiedlichen Fragestellungen der Wörterbuchbenutzungsforschung veröffentlicht worden. Sie widmen sich der:

(1) Typologie der Wörterbücher nach Benutzungsmöglichkeiten (u.a. Engelberg/Lemnitzer 2009; Kühn 1989);

(2) Problemlösung und Optimierungsvorschlägen (u.a. Wiegand 1995; Ripfel 1989a; Domínguez 2006; Kemmer 2010);

(3) Erkundung typischer Benutzungssituationen (u.a. Kromann 1995; Tarp 2008) auch im Fachsprachenbereich (Engelberg/Lemnitzer 2009).

Der Mangel an empirischen Untersuchungen wird aber weiterhin von einigen Arbeiten hervorgehoben (Hartmann: 2001; Kromann: 1995; Wiegand 2008). Um diesem Umstand entgegenzuwirken, sind 2012 / 2013 äußerst interessante Studien durchgeführt worden, im deutschen Raum die obenerwähnte Studie am IDS, im spanischen Raum die UDALPE-Befragung . Sie bedienen sich unterschiedlicher empirischer Methoden, wie der Beobachtung, der Befragung, des Protokolls, des Tests, des Experiments, des interpretativen Verfahrens, des „Simultaneous Feedback“ (de Schryver/Prinsloo 2000). Auch die Verbindung mehrerer Methoden wird als positiv eingestuft. Die Umfrage erweist sich jedoch durch ihre Verwendung in mehreren Studien als eine der meistbenutzten Methoden, um die Benutzungssituationen sowohl seitens des Fremdsprachenlerners (de Schryver/Prinsloo 2011) als auch seitens des Muttersprachlers (Retti 2004; Ekwa Ebanéga/Tomba Moussavou 2008) zu erkunden und um die Verwendung von Spezialwörterbüchern (Wang 2001; Muráth 2005; Taljard / Prinsloo 2011) und von Online-Wörterbüchern (vgl. elexiko) zu erforschen.

3 Die Untersuchung

Die vorliegende Untersuchung wurde 2012/2013 durchgeführt und befasst sich mit der Auswertung der Ergebnisse einer Umfrage bei 41 italophonen Studierenden des *Dipartimento di Filologia, Letteratura e Linguistica* der Universität Pisa, die Deutsch als Fremdsprachestudieren. Die Umfrage, dessen Hauptanliegen die Erkundung der Wörterbuchbenutzung einer DaF-Lerngruppe einer sprachwissenschaftlichen Fakultät ist, wurde in Anlehnung an Hartmanns 5. Hypothese *„An analysis of users´ needs should precede dictionary design"* (1989) durchgeführt, da diese grundlegend für das Projekt *DIL* ist[1]. Ein Teil der Ergebnisse wird in diesem Beitrag vorgestellt und problematisiert, mit entsprechenden Zusammenfassungen und Schlussfolgerungen.

Der Fragebogen in italienischer Sprache wurde den Probanden aufgebreitet und hatte eine Ausfüllzeit von ca. 20 Minuten. Die 27 Fragen konzentrieren sich auf folgende Kernfrageblöcke: 1. Wörterbuchbenutzer (auch Soziodemografie), 2. Wörterbuchtyp, Wörterbuchbenutzungssituation, Bedürfnisse; 3. Wörterbuchformat; 4. Benutzung der Umtexte; 5. das sprachwissenschaftliche Fachwörterbuch und die mikrostrukturellen Eigenschaften. Die Fragen des ersten Blocks betreffen die persönlichen Daten der Befragten (Alter, Geschlecht, Muttersprache, Niveau der Deutschkenntnisse, Motivation zur Auswahl des Deutschen als Fremdsprache, das Erlenen anderer Fremdsprachen). Im zweiten Block werden die Gründe zur Benutzung des Wörterbuches, die typische Benutzungssituation, die Benutzungsfrequenz und die zu erfüllenden Bedürfnisse erforscht. Unterschiedliche Wörterbuchtypen (einsprachige, zweisprachige, fachliche, sprachwissenschaftliche und weitere) und ihre jeweilige Benutzungsfrequenz werden untersucht. Der dritte Block der Umfrage widmet sich dem Format, den Vor- und Nachteilen eines Online-Wörterbuchs, den verwendeten dynamischen Elementen und der Art der Verlinkung. Die Informationsstruktur, die vorkommenden Probleme und Optimierungsvorschläge werden ebenfalls analysiert. Es wird auch auf Copyright und Aktualisierung der Wörterbüchereingegangen. Der vierte Block betrifft die Umtexte und deren Benutzungsfrequenz: Ziele und Hauptmerkmale bestimmter Umtexte stehen im Mittelpunkt des Interesses. Im fünften Block wird spezifisch auf die Funktionen und Bedürfnisse des sprachwissenschaftlichen Fachwörterbuches eingegangen.

4 Die Untersuchung: Ergebnisse

Die Ergebnisse der Umfrage werden in den folgenden fünf vorgestellten Hauptfrageblöcken zusammengefasst: Wörterbuchbenutzer (4.1); Wörterbuchtyp und Bedürfnisse (4.2); Wörterbuchformat (4.3); Verwendung der Umtexte (4.4); Funktion und Bedürfnisse des sprachwissenschaftliches Fachwörterbuches (4.5).

[1] DIL ist ein Deutsch-Italienisches Online-Fachwörterbuch der Linguistik, das an der Universität Pisa entwickelt worden ist und in ständiger Bearbeitung ist.

4.1 Persönliche Angaben

Die ersten sieben Fragen betreffen die persönlichen Daten der Befragten. Die Analyse der Daten hat Folgendes ergeben:

(1) Die Gruppe besteht aus 23% Männer und 77% Frauen;

(2) Die Altersgruppe kann wie folgt dargestellt werden:

	Männer	Frauen
19-21 Jahre	67%	74%
22-24 Jahre	22%	23%
Über 25 Jahre	11%	3%

Tabelle 1: Altersaufteilung der Testpersonen.

(3) Die Muttersprache ist gemäß dem Umfrageziel hauptsächlich Italienisch, auch wenn 2 Studentinnen Albanisch und eine Studentin Russisch als Muttersprache haben;

(4) Die befragten Studenten besuchen den ersten, den zweiten oder den dritten Jahrgang des Bachelorstudienganges;

(5) Die Deutschkenntnisse der Testpersonen sind sehr unterschiedlich: die Mehrheit weist das Sprachniveau A1 (49%) auf; während nur 8% A2 hat; 38% der Studenten und Studentinnen besitzen das Mittelstufenniveau B1, während nur 5% B2erreicht haben . Kein einziger Student verfügt über ein Hochstufenniveau (C1 oder C2). Nur 18% der Versuchspersonen haben eine Zertifizierung der Fremdsprachenkenntnisse (Zertifikat B1).

(6) Die Gründe zum Erlernen des Deutschen als Fremdsprache sind ziemlich homogen: 50% der befragten Personen haben Deutsch wegen möglicher Chancen auf dem Arbeitsmarkt gewählt, 23% wegen persönlicher Interessen (Liebe zur deutschen Sprache und Kultur; Bedeutung in der Europäischen Union etc.); 23% sind von der Fremdsprache fasziniert; 4% aus Neugierde, da sie schon weitere Fremdsprachen kennen;

(7) Deutsch wird von fast allen Studenten als zweite (12%) oder dritte Fremdsprache (88%) gelernt. 98% sprechen Englisch als erste Fremdsprache und nur 2% Französisch. Die Aufteilung der zweiten erlernten Fremdsprache kann aus folgender Tabelle entnommen werden:

Französisch	Spanisch	Russisch	Polnisch	Keine weitere Fremdsprache nach Englisch
54%	28%	3%	3%	12%

Tabelle 2: Zweite erlernte Fremdsprache nach Englisch vor Deutsch.

Zusammenfassend kann festgestellt werden, dass die Testgruppe in großem Maße aus Frauen besteht, die ein Durchschnittsalter zwischen 19 und 21 Jahren aufweisen. Sie besitzen entweder das Sprachniveau A1 oder B1, haben aber selten Prüfungen zur Bestätigung des erlangten Fremdsprachenniveaus abgelegt. Die Motivation zur Wahl des Deutschen als Fremdsprache ist stark mit der Arbeitswelt verbunden; die zentrale Rolle Deutschlands in der EU ist einer der oft genannten Gründe. Deutsch wird meistens als dritte Fremdsprache nach Englisch und Französisch gewählt.

4.2 Wörterbuchtyp und Wörterbuchbenutzungssituation

Aus der Analyse der Items 7, 8, 9, 10, 11, 12, 13 kann ein Profil der verwendeten Wörter- und Fachwörterbücher seitens der Testpersonen geschaffen werden:

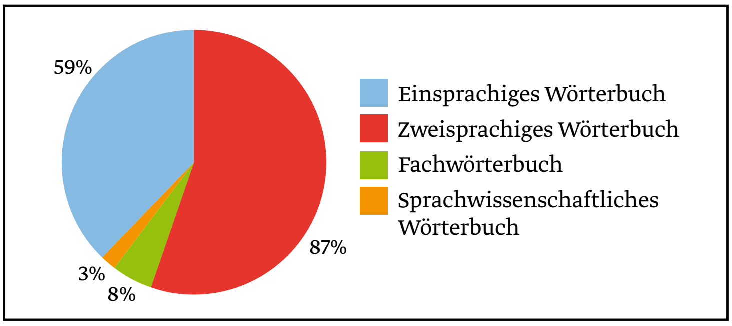


Abbildung 1: Diagramm der von den Probanden benutzten Wörterbuchtypen.

Die meisten Studentinnen und Studenten benutzen ein zweisprachiges Wörterbuch, Fachwörterbücher und Sprachwissenschaftliche Wörterbücherwerden kaum verwendet; dies obwohl die Testpersonen auch sprachwissenschaftliche Kurse besuchen. Es wäre interessant nachzuforschen, welche Gründe es dafür gibt.

Die Frequenz der benutzten Wörterbuchtypen variiert sehr stark und kann der nachfolgenden Tabelle entnommen werden:

	oft	manchmal	selten	keine Angabe zur Frequenz
Einsprachiges Wörterbuch	26%	52%	22%	-
Zweisprachiges Wörterbuch	65%	20%	-	15%
Fachwörterbuch	-	33%	33%	34%
Sprachwissenschaftliches Wörterbuch	-	-	-	100%

Tabelle 3: Darstellung der Verwendungsfrequenz der Wörterbuchtypen.

Das zweisprachige Wörterbuch ist der meistbenutzte Wörterbuchtyp, wie man es bei Studenten, die Fremdsprachen lernen, erwartet hätte. Auch die Benutzungsfrequenz bestätigt dies, denn 65% der Probanden benutzen es oft. Negativ ist das Ergebnis zum sprachwissenschaftlichen Wörterbuch; aufgrund der Präsenz von Fächern wie Linguistik und Fremdsprachendidaktik im Studienplan der Studenten hätte man eine höhere Frequenz erwartet.

Als weitere verwendete Wörterbuchtypen werden von den Versuchspersonen Synonymwörterbücher (31%), Wörterbücher zur Rechtschreibung (3%) und zur Phraseologie (3%) genannt. Die Benutzung von Valenzwörterbüchern enthält keine Treffer.

Interessant ist die Beobachtung zu den Gründen oder Motivatoren, welche die Testpersonen dazu veranlasst haben sich der Wörterbücher zu bedienen: 51% nennen die Schule oder universitäre Institution (eine große Zahl von Probanden nennt die Deutschlehrerin oder Deutschprofessorin); 6% die Familie und 18% das persönliche Interesse mehr zu verstehen und sich selbstständiger zu entwickeln.

4.3 Wörterbuchformate

Die Ergebnisse zu den Fragen, die sich auf das Wörterbuchformat (Fragen 14, 15, 16) konzentrieren, sind in einem Zeitraum der Technologie und Multimedialität erstaunlich, denn noch 43% der Probanden benutzen Printwörterbücher, 8% CD-Rom Wörterbücher und nur 18% Online-Produkte. 20% der Versuchspersonen entscheiden sich für zwei Typen: Print- und Onlinewörterbücher (20%), CD-Rom und Onlinewörterbücher (8%) und Print- und CD-Romwörterbücher (3%).

Die Fragen (15 und 16) haben versucht, ein Bild über die möglichen Vorteile und Nachteile der obengenannten Formate zu schaffen. Als Vorteile werden folgende Aspekte genannt (von links nach rechts in einer absteigenden Skala):

	1.	2.	3.	4.	5.
Printformat	Zweckmäßigkeit	Vollständigkeit	Präzision	Zuverlässigkeit	Größere Anzahl von Einträgen und Beispielen
CD-Rom Format	Schnelligkeit	Vollständigkeit	-	-	-
Online-Format	Schnelligkeit	Jederzeit benutzbar	Einfachheit	Suche auch ohne Nennform / keine Bezahlung	Forum

Tabelle 4: Vorteile der unterschiedlichen Formate.

Als Nachteile:

	1.	2.	3.	4.
Printformat	Zeitverschwendung	keine Zweckmäßigkeit (Größe, Schwere)	Kosten	Niedrige Erfolgsquote
CD-Rom Format	Fehler	Genauigkeit / Präzision	Lücken	Keine automatische Aktualisierung
Online-Format	Genauigkeit / Präzision	Fehler	Internetverbindung	Vollständigkeit

Tabelle 5: Nachteile der unterschiedlichen Formate.

Wie man der Analyse entnehmen kann, widersprechen die mehrmals genannten Vorteile, wie Schnelligkeit und Einfachheit von der negativen Annahme, dass Online-Wörterbücher fehlerhafter, lückenhafter und oberflächlicher als die Printprodukte seien. Das traditionelle Printformat gilt trotz Zeitverlust und Problemen mit der Größe und Schwere nach wie vor als der zuverlässigste Typus.
Es wurden auch spezifische Items zu den Online-Wörterbüchern erarbeitet, wie Items 17, 18, 19, 20, 21, 22, 23, 24. Laut der Befragten sind die folgenden dynamischen Instrumente am wichtigsten: fortgeschrittene Suchmaschine (mit Schreibhilfen, Suche auch in den Definitionen, Suche durch Platzhalter etc.) mit 86%; einfache Suchmaschine mit 49%, alphabetische Leiste mit 13% und das Vorhandensein eines Forums (3%). Verlinkte Verweise werden von 74% der Probanden als positiv eingestuft; während nur 29% sie als negativ beurteilen und die Verlustgefahr (vgl. der Begriff „lost in hyperspace“), Langsamkeit, Unvollständigkeit als Gründe nennen. Interne Links (Verweise zu verbundenen Einträge etc.) werden, wie die nachfolgende Graphik veranschaulicht, als positiv bewertet:

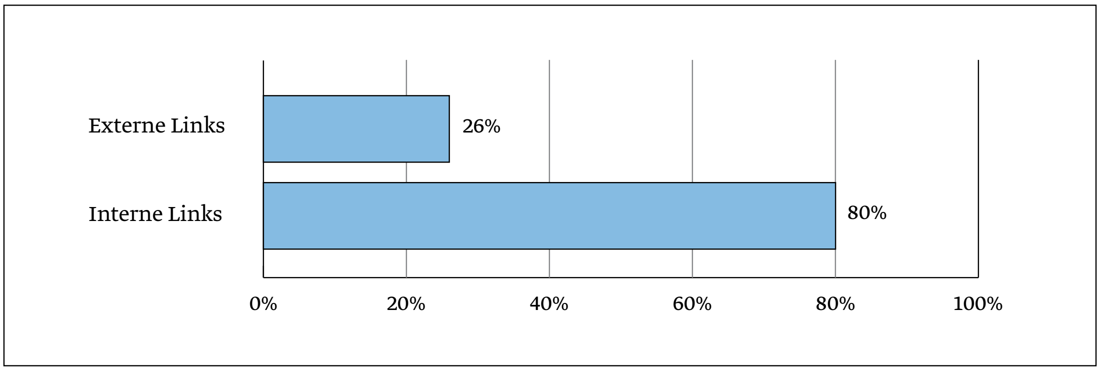


Bild 2: Graphik zu den bevorzugten Links.

Sehr interessant sind auch die Ergebnisse zu folgenden Instrumenten, die für Online-Produkte von großer Nützlichkeit sind, wie die Möglichkeit Kontakt mit den Autoren oder dem Wörterbuchteam aufzunehmen, um mögliche Fehler oder Lücken zu signalisieren, oder die Möglichkeit, den Eintrag zu drucken. Nur 10% der Befragten interessieren sich jeweils für beide Aspekte. Legitimierungsmerkmale haben nur für 28% der Versuchspersonen einen Wert, während die Aktualisierung der Wörterbücher (67%), die für viele Testpersonen einen ständigen Fortschritt gemäß der Entwicklung der Sprache bedeutet, von besonderer Wichtigkeit ist.

4.4 Verwendung der Umtexte

Die Frage 25 betrifft die Untersuchung zur Verwendung von Umtexten durch die Testpersonen, mit Berücksichtigung der Frequenzangaben (oft, manchmal, selten, nie). Die Ergebnisse bestätigen die Vorerwartungen, nämlich, dass wenige Umtexte verwendet werden:

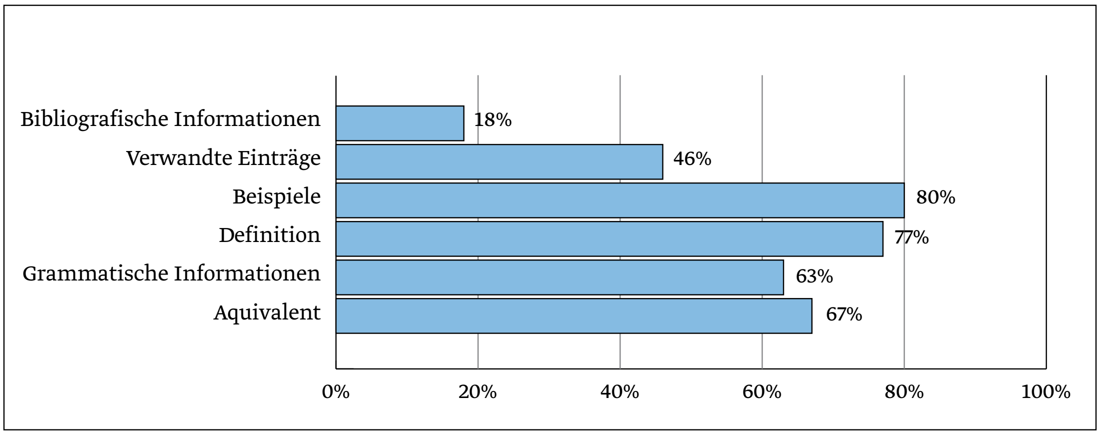


Bild 3: Graphik zu den verwendeten Umtexten.

Die am meisten verwendeten Umtexte sind das Abkürzungsverzeichnis und das Register. Selten und kaum werden die Einleitung und die Benutzungshinweise gelesen, was auch zu der obengenannten Einschätzung der Online-Wörterbücher führen kann. Durch eine genauere Lektüre dieser Texte könnten eine irrtümliche Suche und eine falsche Beurteilung der Online-Produkte vermieden werden.

5 Funktion und Bedürfnisse des sprachwissenschaftlichen Fachwörterbuches

Der fünfte Block konzentriert sich auf das sprachwissenschaftliche Wörterbuch. Trotz der negativen Ergebnisse bezüglich der Benutzung von sprachwissenschaftlichen Fachwörterbüchern würden die befragten Probanden, dieses Werkzeug zur Rate ziehen:

(1) um einen Begriff /Text zu verstehen (66%);
(2) um ein Wort oder einen Text zu übersetzen (33%);
(3) um einen Text zu produzieren (21%);
(4) um sich über ein Thema zu informieren (13%);
(5) um bibliographische Informationen zu suchen (3%);
(6) um ein Problem zu lösen (3%).

Folgende Angaben in der Mikrostruktur von sprachwissenschaftlichen Wörterbüchern werden von den Testpersonen signalisiert:

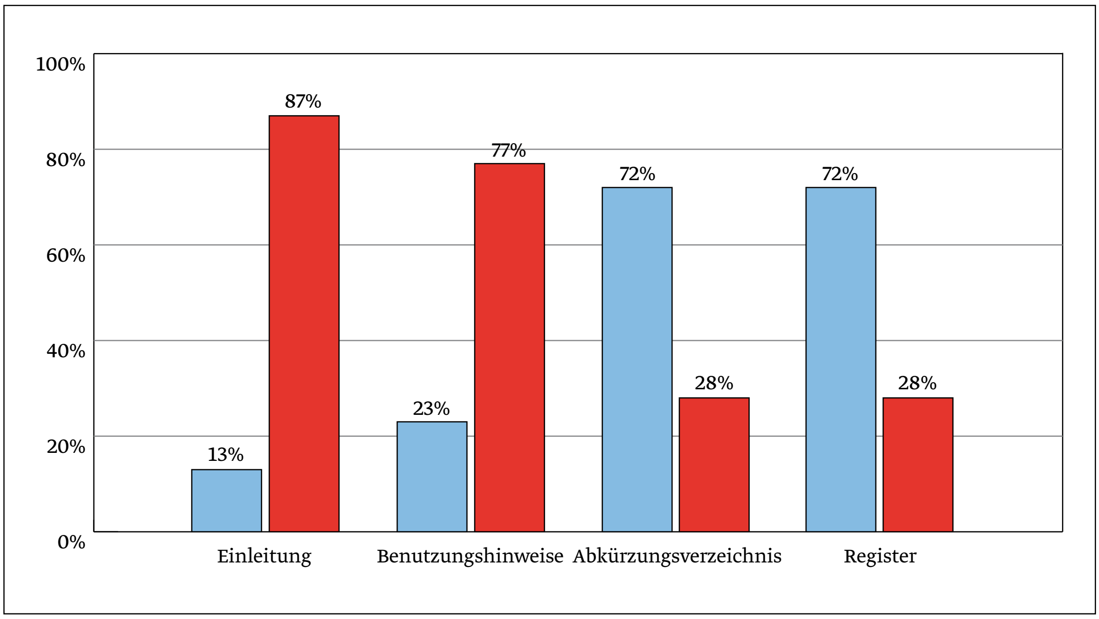


Bild 4: Graphik zu den mikrostrukturellen Angaben eines sprachwissenschaftlichen Wörterbuches.

Die Definition des Eintrages und das Vorhandensein von Beispielen werden als am wichtigsten eingeschätzt. Das Hinzufügen von grammatischen Informationen (wie Genus und Numerus), sowie des fremdsprachlichen Äquivalentes (eventuell auch mit grammatischen Informationen) sind auch in großen Maßen erwünscht. Die Angabe von verwandten Einträgen wird von 46% der Befragten hervorgehoben. Viel seltener wird das Hinzufügen von bibliographischen Informationen (18%) genannt.
Als weitere Angaben werden folgende Bereiche genannt: Phonetik (54%); Kollokationen (54%); Rechtschreibung (46%); Synonyme und Antonyme (41%); Etymologie (31%); Markierung des Fachbereiches (18%). Eine weitere Angabe ist die phonetische Transkription.

6 Schlussbemerkungen

Die empirische Untersuchung liefert ein Bild über die Wörterbuchbenutzung der Probanden und über ihre Einstellung zu den unterschiedlichen Typen und Formaten von Wörterbüchern. Die Tatsache, dass die meisten Testpersonen zweisprachige Wörterbücher verwenden, ist keine Überraschung, erstaunlich ist jedoch die Tatsache, dass sie kaum Fachwörterbücher und insbesondere sprachwissenschaftliche lexikographische Produkte verwenden.

Die Ergebnisse zu den bevorzugten Formaten hätte man auch nicht erwartet: in einem Zeitalter, in dem Smartphones, Tablets und Internet im Alltag dominieren, ist die Zahl der benutzen Printprodukte jedoch sehr hoch. Die Vorteile und Nachteile der jeweiligen Formate bezeugen weiterhin, dass die Internetlexikographie noch einiges machen muss, um bessere und zuverlässigere Produkte zu entwickeln. Trotz wichtiger Projekte und Wörterbücher sind noch viele Produkte verbesserungsfähig und können sich nicht der soliden Basis wissenschaftlicher und theoretischer Erkenntnisse entziehen.

Die Analyse hat auch ergeben, dass Umtexte im Online-Medium (außer Abkürzungsverzeichnisse und Register) noch zu wenig benutzt werden, was vermutlich mit ihrer „Wissenschaftlichkeit" in Zusammenhang gebracht werden kann. Das Augenmerk der Probanden auf das Copyright (oft mit Professionalität in Verbindung gesetzt) und insbesondere auf die zeitnahe Aktualisierung der Produkte deutet darauf hin, dass ein stärkerer Wunsch nach Produkten in ständiger Entwicklung besteht.

Interessant waren auch die Ergebnisse zu den mikrostrukturellen Eigenschaften eines sprachwissenschaftlichen Wörterbuchs: die Präsenz sowohl von grammatischen Informationen, Äquivalenzangaben (Sprachinformationen) als auch von Definitionen (Sachinformationen) deutet auf den Typ „Allbuch" (Wiegand 1998: 762) hin. Der Wunsch, dass auch grammatische Informationen, phonetische Angaben zu den Äquivalenten sowie Gebrauchsbeispiele vorhanden sein sollten, zeigt, dass Fachwörterbücher diese mikrostrukturellen Eigenschaften stärker berücksichtigen sollten.

Außerdem sollte der Frage nachgegangen werden, ob und in wie fern die Studenten des Masterstudiengangs von diesen Ergebnissen abweichen.

7 Literaturangaben

Atkins, B.T.S.,Varantola, K. (1998). Language learners using dictionaries: The final report on the EURALEX/AILA research project on dictionary use. In *Using dictionaries: Studies of dictionary use by language learners and translators*, S. 83-122.

De Schryver, G.M., Prinsloo, D.J. (2000). Dictionary-Making Process with 'Simultaneous Feedback' from the target Users to the Compilers. In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th, 2000.* Stuttgart: Niemeyer, S. 197–209.

De Schryver, G.M., Prinsloo, D.J. (2011). Do Dictionaries Define on the Level of Their Target Users? A Case Study for Three Dutch Dictionaries. In *International Journal of Lexicography* 24/1, S. 5-28.

Domínguez Vázquez, M.J. (2006). Von monolingualen Wörterbüchern zu kontrastiven Valenzwörterbüchern. Die Valenzwiedergabe unter der Lupe. In *Jahrbuch für Deutsch als Fremdsprache. Intercultural German Studies* 1/32, S. 231-241.

Domínguez Vázquez, M.J., Balsa, M.M., Vidal Pérez, V. (2013). *Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen.* http://www.academia.edu/4640188/Worterbuchbenutzung_Erwartungen_und_Bedurfnisse._Ergebnisse_einer_Umfrage_bei_Deutsch_lernenden_Hispanophonen

Ekwa Ebanéga, G-M., Tomba Moussavou, F. (2008). Survey of Dictionary Use: Case Studies of Gabonese Students at the University of Stellenbosch and the Cape Peninsula University of Technology. In *Lexikos*, 18, S. 349-365.

Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Flinz, C. (2011): DIL (Deutsch-Italienisches Wörterbuch der Linguistik). Vom Projekt zur Realität: Hinweise zum aktuellen Stand. In: *daf Werkstatt*, S. 185-200.

Hartmann, R.R.K. (1989). Sociology of The Dictionary User: Hypotheses and Empirical Studies. In F.J. Hausmann et al. (Hrsg.) *Wörterbücher: ein Internationales Handbuch zur Lexikographie*. Bd. 1. Berlin, New York: de Gruyter, S. 649-657.

Hartmann, R.R.K. (1999). Case Study: The Exeter University Survey of Dictionary Use [Thematic Report 2]. In R.R.K. Hartmann (Hrsg.) *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub- Project 9: Dictionaries*. Berlin: Thematic Network Project in the Area of Languages, S. 36-52. Accessed at: http://www.fu-berlin.de/elc/TNPproducts/SP9dossier.doc [07/03/2014].

Hartmann, R.R.K. (2000). European Dictionary Culture. The Exeter Case Study of Dictionary Use among University Students, against the Wider Context of the Reports and Recommendations of the Thematic Network Project in the Area of Languages (1996- 1999). In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart: Niemeyer, S. 385-391.

Hartmann, R.R.K. (2001). *Teaching and Researching Lexicography*. London: Pearson.

Hausmann, F.J., Reichmann, O., Wiegand, H.E., Zgusta, L. (Hg.) (1989). *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1. Erster Teilband. Berlin, New York: de Gruyter.

Kemmer, K. (2010). *Onlinewörterbücher in der Wörterbuchkritik* (OPAL 2-2010). Mannheim. (= Online publizierte Arbeiten zur Linguistik)

Kromann, H.P. (1995). Deutsche Wörterbücher aus der Perspektive eines fremdsprachigen Benutzers. In H. Popp (Hrsg.) *Deutsch als Fremdsprache. An den Quellen eines Faches. Festschrift für Gerhard Helbig zum 65. Geburtstag*. München: Iudicium, S. 501-512.

Kühn, P. (1989). Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In F.J. Hausmann et al. (Hrsg.) *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1. Erster Teilband. Berlin, New York: de Gruyter, S. 111-127.

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

Muráth, J. (2005). Wörterbuchbenutzung von Fachübersetzerstudenten Ihre Erwartungen an ein Fachwörterbuch. In *Lexicographica,* 115, S. 401-415.

Nesi, H. (2000). Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art. In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress Stuttgart August 8.-12. 2000*. 1. Halbbd. Stuttgart: Niemeyer, S. 839-847.

Retti, G. (2004). *Österreichisches Deutsch und Österreichisches Wörterbuch*. Accessed at: http://gregor.retti.info/oewb/ and http://gregor.retti.info/docs/ retti1991/4.pdf [07/03/2014].

Ripfel, M. (1989a). Wörterbuchkritik eine empirische Analyse von Wörterbuchrezensionen. Tübingen: Niemeyer.

Ripfel, M., Wiegand, H.E. (1988). Empirische Wörterbuchbenutzungsforschung. In *Studien zur neuhochdeutschen Lexikographie VI*. 2. Teilbd. Hildesheim: Olms, S. 91-520. (Germanistische Linguistik 87-90/1986)

Rossenbeck, K. (2005). Die zweisprachige Fachlexikographie in der neueren und neuesten Wörterbuchforschung". In *Lexicographica,* 21, S. 179-201.

Sánchez Ramos, M. del Mar (2005). Research on Dictionary Use by Trainee Translators. In *Translation Journal,* 9.2. Accessed at: http://www.proz.com/translation-articles/articles/227/1/Research-on-Dictionary-Use-by-Trainee-
Translators [19.03.2014].

Simonsen, H.K. (2011). User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. In *Hermes. Journal of Language and Communication Studies*, 46, S. 75-101.

Taljard, E., Prinsloo, D.J., Fricke, I. (2011). The use of LSP dictionaries in secondary schools? A South African case study. In *South African Journal of African Languages,* 31.1, S. 87-109.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Max Niemeyer.

Wang, W. (2001). Zweisprachige Fachlexikographie. Benutzungsforschung, Typologie und mikrostrukturelle Konzeption. Frankfurt a. M.: Lang.

Welker, H.A. (2010). Dictionary Use. A General Survey of Empirical Studies. Brasilia: Author's Edition.

Wiegand, H.E. (1987). Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. In *Lexicographica*, 3, S. 178-227.

Wiegand, H.E. (1995). Lexikographische Texte in einsprachigen Lernerwörterbuchern. Kritische Überlegungen anlässlich des Erscheinens von Langenscheidts „Grösswörterbuch Deutsch als Fremdsprache". In H. Popp (Hrsg.) *Deutsch als Fremdsprache. An der Quellen eines Faches. Festschrift für Gerhard Helbig zum 65. Geburtstag*. München: Iudicium, S. 463-499.

Wiegand, H.E. (1998). Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1 Teilband. Berlin, New York: de Gruyter.

Wiegand, H.E. (2008). Wörterbuchbenutzung bei der Übersetzung. Möglichkeiten ihrer Erforschung. In *Germanistische Linguistik*, 195-96, S. 1-43.

Wiegand, H.E. (2010). Zur Methodologie der Systematischen Wörterbuchforschung: Ausgewählte Untersuchungs- und Darstellungsmethoden für die Wörterbuchform. In *Lexicographica*, 26, S. 249-330.

Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of "Strategy Inventory for Dictionary Use" (S.I.D.U)

Gavriilidou Zoe
Democritus University of Thrace
zoegab@otenet.gr

Abstract

The present paper reports results regarding the adaptation in English of the Strategy Inventory for Dictionary Use (S.I.D.U.) and the preliminary psychometric evaluation of the English version. S.I.D.U is a 36-item self-report questionnaire for assessing dictionary use strategies specifying four main areas of interest: a) dictionary use awareness skills, b) strategies for dictionary selection and acquaintance with dictionary conventions, c) lemmatization strategies, and finally d) look-up strategies. The original scale was translated from Greek into English, back-translated and reviewed. Cross-cultural adaptation included the experts' revision followed by its administration to 52 participants. Its internal consistency was .89. Similarly all four subscales showed good to excellent internal consistency (dictionary use awareness: α=.83, dictionary selection and acquaintance: α=.76, lemmatization: α=.86 and dictionary search: α=.78). Test-retest reliability ranged from fair to good for the total scale and its six-subscales

Keywords: Assessment; validation; pedagogical Lexicography; strategic dictionary use; dictionary selection strategies; dictionary acquaintance strategies; lemmatization strategies; dictionary search

1 Introduction

There is in recent literature a growing interest in pedagogical lexicography and more precisely in the study of *use situations* of dictionaries (Béjoint 1989; Gavriilidou 2010; Petrylaité, Vezyté, Vaskeliené 2008; Prichard 2008; Scholfield 2002), the *dictionary-using skills and strategies* (Ard 1982; Bogaards 1994; Diab 1990; Gavriilidou 2014; Hartmann 1994) or *the role of dictionary in pedagogical process* (Bensoussan 1983; Bogaards 1998; Hulstjin 1993; Nesi 1996). Relevant research (Gu & Johnson 1996; Nation 2001; Schmitt 1997) has also shown that dictionary use is an important vocabulary strategy that a) occurs successfully in conjunction with guessing (or inferencing) and note-taking, b) provides information about a specific item, and c) has a positive influence on the learner's acquisition process (Hulstijn 1993; Luppescu & Day 1993; Knight 1994; Laufer & Hadar 1997; Laufer & Hill 2000; Bruton 2007). No previous research, however, has focused on the relationship between *strategic dictionary use* and successful reading comprehension, production or vocabulary acquisition even though current trends in language curricula design stress the importance of strategy use in language teaching. In the current study strategic dictionary use refers to a) the conscious awareness of when to use a dictionary and what type to use and b) the ability to employ efficient lemmatization and look-up strategies. Given that strategic dictionary use is beneficial in vocabulary acquisition, L2 Learning, reading comprehension (Scholfied 1982; Knight 1994; Hulstijn, Hollander and Greidanus 1996; Scholfield 1999; Prichard 2008) or text production (Nesi and Meara 1994; Fraser 1999; Elola, Rodríguez-García and Winfrey 2008), there is a major claim in pedagogical lexicography that strategic dictionary use should be taught and, as a result, dictionary users should develop a more effective strategic behavior while looking up words. However, strategic dictionary use instruction, while crucial, still remains a secondary concern in the relevant research. It is true, on the other hand, that the design of strategic dictionary use instruction programs has to be based on reliable data describing dictionary user's attitudes, preferences or strategies employed while they look up a word. In order to accurately explore the needs of dictionary users, measures that produce valid and reliable estimates of user's dictionary strategies must be identified. However, there was, until recently, no standardized instrument for profiling users' dictionary strategies in a valid and reliable way.

The purpose of this paper is to present the structure and characteristics of the English version of the newly developed self-report "Strategy Inventory for Dictionary Use" (S.I.D.U) (Gavriilidou 2011; 2013) which was first elaborated and standardized in Greek Language for profiling dictionary users in a valid and reliable manner. An assessment of strategic dictionary use which offers valid and reliable scores will further develop our understanding of the construct of strategic dictionary use and its relation to vocabulary acquisition, and text comprehension or production. The use of self-report instruments to investigate various aspects of individual learner differences is a common practice in the field of language learning research. However, although a given instrument may have been rigorously developed and subjected to various measures of reliability and validity, when it is translated into another language or used in a cultural setting different from the one originally intended, it must once again be rigorously examined. The cross-cultural adaptation of self-report questionnaires for use in a new country, culture or language necessitates use of a rigorous protocol in order to reach equivalence between the original source and target version. Furthermore, the items must not only be translated well linguistically, but also must be adapted culturally to maintain the content validity of the instrument at a conceptual level across different cultures (Beaton et al 2000). Thus, this paper also provides data about S.I.D.U's translation and cultural adaptation in English and focuses on following an appropriate adaptation protocol that would maximize the questionnaire's reliability and validity when used to compare the scores across cultures and languages. Finally the paper reports results regarding the instruments' reliability and validity.

2 The Strategy Inventory for Dictionary Use

The S.I.D.U (Gavriilidou 2011, 2013) is a standardized self-report instrument first elaborated in Greek for assessing the frequency with which the respondent uses different strategies or techniques during dictionary use. It can be administered since the age of 11 years old. It consists of 36 items with five Likert-scale responses of never or almost never true of me, generally not true of me, somewhat true of me, generally true of me, always or almost always true of me.

Given that strategic dictionary use is part of a larger construct of strategy use, we approached this instrument design following the development procedure of the Strategy Inventory for Language Learning (Oxford 1990). In order to develop the test's specification of the Greek version, all previous literature was consulted in detail and an exhaustive list including all reference skills cited in the literature was prepared. That list was used as a basis for item writing. More specifically two different kinds of research had been consulted: theoretical or empirical papers presenting detailed descriptions or taxonomies of the reference skills (or strategies) that dictionary users should demonstrate for a successful dictionary search (e.g. Béjoint 1981; Scholfield 1982 and 1999; Bogaards 1994; Roberts 1997; Hartmann 1999; Nesi 1999; Nation 2001; Hartmann and James 2002; Lew and Galas 2008) and empirical papers investigating the reference skills, misuse and errors of dictionary users during dictionary look-up (Béjoint and Moulin 1987; Maingay and Rundell 1987; Neubach and Cohen 1988; Nuccorini 1992 and 1994; Nesi and Meara 1994; Christianson 1997; Harvey and Yuill 1997; Wingate 2004; Elola, Rodríguez-Garcı´a and Winfrey 2008; Petrylaité , Vaškeliené, Vėžyté 2008). The method of multiple judges was adopted for the measurement of content validity of the pilot version of S.I.D.U. The measurement was carried out on a panel of 10 experts who were either University Professors specialized in Lexicology or Lexicography with a long experience in dictionary compilation or University Professors specialized in Language Teaching. The experts judged the relevance and usefulness of each one of the 52 candidate items of S.I.D.U and included 47 items in the pilot version.

To check the construct validity of the original S.I.D.U, principal component analysis with Varimax rotation on SPSS version 15 was adopted. A communality of .30 was set as a cut off for inclusion in the final analysis. Consequently, eleven items were excluded. The results showed that a total of 36 items loaded on four factors, accounting for the 51.7% of the variance. Based on these results, Gavriilidou (2013) organized the original S.I.D.U into four strategy subscales:

- Strategies which lead to a decision to use a dictionary in order to resolve a problem encountered inside or outside the class (dictionary use awareness) (items 1-14).
- Strategies which permit to select an appropriate dictionary type depending on the problem to be solved and guarantee the acquaintance with one's own dictionary (15-21).
- Lemmatization strategies, that is strategies which help finding the citation form of inflected forms found in the text. Users should be able to be based on morphological indices (stems, prefixes, suffixes, inflectional morphemes) of the unknown word that has been met in the text in order to make hypotheses about the look-up form of that word or should be acquainted with alphabetical sequencing otherwise lemmatization is not possible (22-29).

- Look up strategies, which control and facilitate the localization of the correct part of the entry where different meanings of the same word form are included (30-36)

Each of the four factors was considered according to previous literature and was named by the author. The test discriminated expert from non-expert users in all four categories of strategies (p<.001). Internal consistency of the four subscales (dictionary use awareness, dictionary selection, acquaintance and lemmatization and dictionary search) and the overall scale of the S.I.D.U. were found to be excellent.

3 The Translation and Adaptation Protocol of S.I.D.U in English

The process of adaptation of S.I.D.U into English was broken down into three steps: (a) the translation process, (b) cross-cultural verification and adaptation, and (c) verification of the psychometric properties of the instrument. The translation process consisted of the initial translations, synthesis of the translations and back translation. The second step included the expert committee review in the light of the focus group suggestions and other verification methods. Finally, in the third stage, the questionnaire was administered and its psychometric properties were verified.

The translation protocol was broken down into six stages: initial translation by two independent translators; synthesis of the translations during which any discrepancies between the two initial translations are resolved; back translation into the original language; expert committee review which should achieve semantic, idiomatic, experiential and conceptual equivalence; pretesting of the final version; and, finally, submission of final reports drown for all the five stages to the coordinating committee (Beaton et al., 2000).

3.1 The initial translation

The first stage in the adaptation was the forward translation of S.I.D.U from Greek into English. Two bilingual translators living in Greece, whose mother tongue was English, one naïve and one informed about the purpose of the study produced two independent translations (T1 and T2). They also composed two independent written reports in which they explained the rationale of their translation choices as well as dubious phrases, uncertainties or encountered translation problems.

3.2 The synthesis of the translations

The two translations were compared and discrepancies reflecting ambiguities in the original instrument were noted. Then the two translators and the creator of the instrument worked on the first translator's (T1) and the second translator's (T2) versions and produced a synthesis of the two versions (T12) by discussing the translation of each of the 36 items. They also wrote a report describing the synthesis procedure, all cases discussed and all solutions adopted.

3.3 Back translation

Two translators having Greek as mother tongue and ignoring the original version of S.I.D.U translated the T12 version back into Greek in order to verify that the translated T12 version in English reflects the same item content as the original Greek version. The two translators, who were not informed of the purpose of the study, handed in the two back translations (BT1 and BT2) as well as two independent reports documenting the back translation procedure.

3.4 Expert committee examination

The expert committee consisted of two linguists, two lexicographers and the four translators. Its role was to examine all the relevant material (initial instrument, T1, T2, T12, BT1, BT2, and the five reports) and to review all the translations for resolving any discrepancy. Its final goal was to arrive to the final English version of the S.I.D.U. To do so, the experts counter-examined the source and target version of S.I.D.U checking the following: a) the semantic equivalence, that is if the words meant the same in Greek and English and whether there was any grammatical difficulties in English translation, b) the idiomatic equivalence, in other words the correct translation of idioms or collocations c) the experiential equivalence, in other words if all items expressed tasks which are experienced in the target culture d) the conceptual equivalence, that is if all the words hold the same conceptual meaning in the two cultures.

The committee produced the final English version of S.I.D.U and wrote a final report which they handed to the author of the instrument. This version was then used for collecting data in order to measure the psychometric properties of the instrument.

4 Reliability

4.1 Sampling

52 under graduate and post graduate students as well as professors of the department of Linguistics of the University of Chicago filled in the questionnaire.

4.2 Statistics

To check the S.I.D.U's internal consistency a Cronbach's Alpha analysis was performed. To check the stability of S.I.D.U scores over time, test-retest data are reported and the intra-class correlation coefficient was computed.

5 Results

5.1 Internal Consistency

Based on the results of S.I.D.U, a total sum score of all 36 items was computed. Moreover, total scores in each subscale (dictionary use awareness, dictionary selection and acquaintance, lemmatization and dictionary search) were also computed. The total scale showed excellent reliability (Cronbach's α =.89). Similarly all four subscales showed good to excellent internal consistency (dictionary use awareness: α=.83, dictionary selection and acquaintance: α=.76, lemmatization: α=.86 and dictionary search: α=.78).

5.2 Test-retest reliability

Test-retest reliability for the total scale and the sub-scales ranged from fair to good (Total scale: r= .778, p<.001, dictionary use awareness: r= .831, p<.001, dictionary selection and acquaintance: r= .874, p<.001, lemmatization: r= .761, p<.001, dictionary search: r= .696,) indicating that at least within the time frame considered here scores of S.I.D.U mirror stable individual differences.

6 Discussion

The present article reports findings concerning the validity and reliability of the translated and culturally adapted in English version of S.I.D.U. Like the original Greek version of the instrument whose internal consistency was found to be excellent (total scale: α =.94, dictionary use awareness: α=.90, dictionary selection and acquaintance: α=.86, lemmatization: α=.83 and dictionary search: α=.84) (Gavriilidou 2013: 12), the translated version showed an excellent reliability for the total scale α =.89 and all four subscales (dictionary use awareness: α=.83, dictionary selection and acquaintance: α=.76, lemmatization: α=.86 and dictionary search: α=.78). Thus the paper, provides evidence for the English version of S.I.D.U as a useful and psychometrically sound measure of dictionary use strategies that may contribute to the scientific investigation of the strategies employed by dictionary users while choosing and using a dictionary, as well as for applied purposes such as the design of class interventions for raising strategic dictionary use. The purpose for developing the English version of S.I.D.U. was to provide a-simple-to administer and reliable instrument for assessing strategic dictionary use cross-linguistically. The fact that the S.I.D.U. was found to be valid and reliable both in the Greek and English version is very promising in that regard.

The paper also records an appropriate adaptation protocol that would maximize the questionnaire's reliability and validity when used to compare the scores across cultures and languages. There were attempts to reduce the potential biases that may occur during translation. Construct and item bias were recorded and were confronted appropriately in order to overcome the problem of measuring different constructs in different cultural groups or distorting the meaning of individual items. That is why "adaptation" and not "application" or "assembly" was selected as it allows for a solution to the afore-mentioned problems of bias. It can be concluded that the process of adapting the S.I.D.U from Greek into English recorded in this paper, however time consuming and costly, is the most effective way to produce an instrument for measuring the frequency of dictionary strategy use of dictionary users. It also allows for comparison of data and findings across nations as it provides the opportunity to examine dictionary strategies of those for whom there previously was no translated version of the S.I.D.U. The carefully planned and executed adaptation process ensures high instrument reliability and validity and offers other researchers interested in questionnaire adaptation a procedure that overcomes most of the problems entailed when instruments are used in different languages and cultures.

The major application of the English version of S.I.D.U. is to assess the dictionary use strategies employed by students or pupils in order to collect reliable data for the design of special curricula for dictionary use training. It can also be used to assess the improvement in dictionary use as a result of the application of these curricula in specific target groups. Furthermore, it can be used as an instrument of sample normalization in research focusing on the role of dictionary use in vocabulary acquisition and on the relationship between the dictionary use and successful reading comprehension or text production, ensuring that different samples of different researches include dictionary users with equivalent abilities in such a way that would yield comparable results. Finally, another possible use is for research purposes on pedagogical lexicography.

7 Conclusions-Limitations of the study

The main contribution of the present empirical study is data about the psychometric properties of the English version of S.I.D.U and the proposed model of questionnaire adaptation, which involves methodological and theoretical considerations necessary for researchers who will adapt or develop relevant tests for various constructs. The proposed model covers empirical, methodological, and theoretical issues. Theoretical issues were addressed in the stage of construct definition. Methodologically, an approach for construct validation was suggested. In short, this model includes different steps and procedures for adapting or developing tests for questionnaires, while still being able to produce instruments that are valid and reliable.

However, it needs to be pointed out that the cultural adaptation procedure carried out in this study has focused on adults, students or professors. This was in line with our need to develop a screening instrument for this particular population, since most of the relevant studies focus on that population. Therefore, unlike the original Greek version of S.I.D.U which can be administered since the age of 11, this particular translation may not be applicable to other age groups, and would need to be reviewed prior to generalized use.

Finally, the instruments' construct validity should be checked with a Factor Analysis using larger samples.

8 References

Ard, J. (1982). The use of bilingual dictionaries by ESL students while writing. In *Review of Applied Linguistics* 58, pp. 1-27.

Beaton, E. D., Bomardier, C., Guillemeni, F. & Bozi-Ferrz, M. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. In *Spine*, 25(24), pp. 3186–3191.

Béjoint, H. (1981). The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. In *Journal of Applied Linguistics*, II (3), pp. 207-222.

Béjoint, H. (1989). *The teaching of dictionary use: present state and future tasks. Dictionaries*. Berlin, New York: Walter de Gruyter. 1st vol. pp. 208-215.

Béjoint, H. and Moulin, A. (1987). The place of the dictionary in an EFL programme. In A. Cowie (ed.), *The dictionary and the language learner*: Papers from the EURALEX seminar at the University of Leeds, 1-3 April 1985. Tübingen: Max Niemeyer Verlag, pp. 97-114.

Bensoussan, M. (1983). Dictionaries and tests of EFL reading comprehension. In *English Language Teaching Journal* 37(4), pp. 341-345.

Bogaards, P. (1994). Tuning the dictionary to the skills of intermediate learners. In *Fremdsprachen Lehren und Lernen* 23, pp. 192-205.

Bogaards, P. (1998). Scanning long entries in Learner's Dictionaries. *Proceedings Euralex* 98, Liège, pp. 555-563.

Bruton, A. (2007). Vocabulary learning from dictionary referencing and language feedback in EFL Translational Writing. In *Language Teaching Research*, 11, pp. 413-431.

Christianson, K. (1997). Dictionary use by EFL writers: what really happens. In *Journal of second language writing*, 6, pp. 23-43.

Diab, T. (1990). Pedagogical Lexicography: a case study of Arab nurses as dictionary users. *Lexicographica Series Maior* 31. Tübingen: Niemeyer.

Elola, I., Rodríguez-García, V.and Winfrey, K. (2008). Dictionary use and vocabulary choices in L2 Writing. In *Estudios de linguistica Inglesa applicada*, 8, pp. 3-89.

Fraser, C. (1999). The Role of Consulting a Dictionary in Reading and Vocabulary Learning. In *Canadian Journal of Applied Linguistics*, 2(1-2), pp. 73-89.

Gavriilidou, Z. (2010). Profiling Greek adult dictionary users. In *Studies of Greek Linguistics* 31, pp. 166-172.

Gavriilidou, Z. (2011). Strategy inventory for dictionary use: Elaboration and Standardization. In Gavriilidou, Z, Efthymiou, A., Kambakis-Vougiouklis, P. and Thomadaki, E. (eds) *Proceedings of the 10th International Conference of Greek Linguistics* available at http://www.icgl.gr.

Gavriilidou, Z. (2013). Development and validation of the Strategy Inventory for Dictionary Use (S.I.D.U). In *International Journal of Lexicography*, 26 (2), pp. 135-153.

Gavriilidou, Z. (2014). User's abilities and performance in dictionary look up. In Lavidas, N., Alexiou, Th. and Sougari, A. (eds) *Major Trends in Theoretical and Applied Linguistics* vol. 2, De Gryuter Open, pp. 41-52 (available at http://www.degruyter.com/viewbooktoc/product/422023).

Gu, P. Y & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. In *Language Learning*, 46, pp. 643-679.

Hartman, R. R. K. (1987). Four perspectives on dictionary use: A critical review of research methods. In A. Cowie (ed) *The dictionary and the language learner*, Tubingen, Max Niemeyer Verlag, pp. 11-28.

Hartmann, R. R. K. (1994) Bilingualised versions of learners' dictionaries. *Fremdsprachen Lehren und Lernen* 23, pp. 206-220.

Hartmann, R. R. K. and James, J. (2002). *Dictionary of Lexicography*, Routledge.

Harvey, K. and Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. In *Applied Linguistics*, 18(3), pp. 253- 278.

Hulstjin, J. H. (1993). When do foreign language readers look up the meaning of unfamiliar words? The influence of task and learner variables. In *The Modern Language Journal* 7 (2), pp. 139-147.

Hulstijn, J-H., Hollander, M. and Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use and re-occurrence of unknown words. In *The Modern Language Journal*, 80(3), pp. 327-339.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. In *The Modern Language Journal*, 78, pp. 285-299.

Laufer, B., & Hadar, L. (1997). Assessing the effectiveness of monolingual, bilingual and 'bilingualized' dictionaries in the comprehension and production of new words. In *The Modern Language Journal*, 81, pp. 189-196.

Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? In *Language Learning and Technology*, 3, pp. 58-76.

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

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

Maingay, S. and Rundell, M. (1987). Anticipating learners' errors: Implications for dictionary writers. In A. Cowie (ed.), *The dictionary and the language learner*: Papers from the EURALEX seminar at the university of Leeds, 1-3 April 1985. Tubingen: Max Niemeyer Verlag, pp. 128-35.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press

Nesi, H. (1996). The role of illustrative examples in productive dictionary use. In *Dictionaries* 17, pp. 198-206.

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

Nesi, H. and Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. In *System* 22, pp. 1-15.

Neubach, A. and Cohen, A. D. (1988). Processing strategies and problems encountered in the use of dictionaries. Dictionaries: In *The Journal of the Dictionary Society of North America*, 10, pp. 1-19.

Nuccorini, S. (1992). Monitoring dictionary use. In H. Tommola (ed.), *EURALEX '92 Proceedings*, Studia Translatologica. Tampere, pp. 89-102.

Nuccorini, S. (1994). On dictionary misuse. In W. Martin (ed.) *EURALEX 1994 Proceedings*. Vrije Universiteit Amsterdam, pp. 586-597.

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House / Harper and Row. Now Boston: Heinle and Heinle.

Petrylaité, R., T. Véžyté, Vaškeliené, D. (2008). Changing skills of dictionary use. In *Studies about languages*, 12, pp. 77-82.

Prichard, C. (2008). Evaluating L2 readers' vocabulary strategies and dictionary use. In *Reading in a foreign language*, 20(2), pp. 216-231.

Roberts, R. (1997). Using Dictionaries Efficiently. In *38th Annual Conference of the American Translators Association*, San Francisco, California. 20 February 2012. http://www.dico. uottawa.ca/articles-en.htm.

Schmitt, N. (1997). Vocabulary learning strategies». In Schmitt & McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press. pp. 199-227.

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

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

Scholfield, P. (2002). Why Shouldn't Monolingual Dictionaries be as Easy to Use as Bilingual Ones? [Retrieved June 10, 2010, from http://www. longman.com/dictionaries/teachers/articles/p-scholfield-02. html].

Wingate, U. 2004. Dictionary use - the need to teach strategies. In *Language Learning Journal*, 29, pp. 5-11.

Acknowledgements

This study is part of the Thales project MIS 379335. It was held in the frame of the National Strategic Reference Frame (Ε.Σ.∏.Α) and was co-funded by resources of the European Union (European Social Fund) and national resources.

Appendix: English version of S.I.D.U

Name (not surname):
Gender:
Date of birth:
Mother Tongue:
Career orientation:

This questionnaire will be used for research purposes and your contribution is very significant. Thank you for your help. Please read the following statements carefully and circle 1, 2, 3, 4 or 5 according to what is most true for you.

(1) Never or almost never true of me.
(2) Generally not true of me.
(3) Somewhat true of me.
(4) Generally true of me.
(5) Always true of me.

I use a dictionary to find the meaning of a word	1	2	3	4	5
I use a dictionary to find the spelling of a word	1	2	3	4	5
I use a dictionary to find synonyms	1	2	3	4	5
I use a dictionary to find antonyms	1	2	3	4	5
I use a dictionary to check how a word is used	1	2	3	4	5
I use a dictionary to find the origin of a word	1	2	3	4	5
I use a dictionary to help myself in translation	1	2	3	4	5
I use a dictionary to find the syntax of a word	1	2	3	4	5
I use a dictionary to find the derivatives of a word	1	2	3	4	5
I use a dictionary to find word families	1	2	3	4	5
I use a dictionary to find the meaning of an expression	1	2	3	4	5

I use a dictionary at home	1	2	3	4	5
I use a dictionary when I read a text	1	2	3	4	5
I use a dictionary when I write a text	1	2	3	4	5
Before I buy a dictionary, I know the reason why I need it	1	2	3	4	5
Before I buy a dictionary at the bookshop, I glance through it to see what information it provides.	1	2	3	4	5
I choose a dictionary because it has a lot of entries and a lot of information in each entry.	1	2	3	4	5
I know what an etymological dictionary is and what it is used for	1	2	3	4	5
I know what a general dictionary is and what it is used for	1	2	3	4	5
I know what a bilingual dictionary is and what it is used for	1	2	3	4	5
I know what a dictionary of technical terms is and what it is used for	1	2	3	4	5
Before I use my new dictionary, I carefully read the introduction	1	2	3	4	5
Before I use my new dictionary, I carefully study the list of abbreviations	1	2	3	4	5
When I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary.	1	2	3	4	5
When I can't locate a proverb or a set phrase in the entry where I thought I would find it, I begin a new search	1	2	3	4	5
When I hear a word I don't know, I consider various spelling possibilities and look it up accordingly	1	2	3	4	5
When I can't find a word where I thought I would find it, I begin a new search until I find it	1	2	3	4	5
To see how a word is used in spoken language, I use the usage labels provided in the entry	1	2	3	4	5
When I look up a word beginning with E, I search in the first quarter pages as E is one of the first letters of the alphabet	1	2	3	4	5
When I look up a word beginning with L, I open my dictionary in the middle	1	2	3	4	5
When I look up a word, I bear in mind its initial letter and then search where I believe this initial letter is in the dictionary.	1	2	3	4	5
When I look up a word, I simply open the dictionary and see if I am near the specific initial letter	1	2	3	4	5
When I look up a word, I constantly bear it in my mind during the search	1	2	3	4	5
When I realize that the word I am looking for has various different meanings, I go through them all one by one, assisted by the example sentences	1	2	3	4	5
When I find the word that I was searching for, I return to the text to confirm that the word matches the context	1	2	3	4	5
Before I use a word I found in the dictionary when writing a text, I read all the information on the grammar of that word (conjugation, syntax) to be sure of the correct usage.	1	2	3	4	5

The Authentic Voices of Dictionary Users – Viewing Comments on an Online Learner's Dictionary Before and After Revision

Ann-Kristin Hult
Department of Swedish, University of Gothenburg
ann-kristin.hult@svenska.gu.se

Abstract

This paper deals with comment field data from two web questionnaire surveys, performed in 2007 and 2011, about the use and users of the *Swedish Lexin Dictionary* (SLD), an online monolingual learner's dictionary. The SLD underwent comprehensive revision in between the two studies. In order to evaluate the update I compare the respondents' comments on SLD before and after the revision. The two sets of comment field data are categorised by the types of comments expressed and by the information categories mentioned. As it turns out, respondents do seem happier with the new version of SLD. Generally, there are more positive comments after the revision than before and the information categories mentioned are more often set in positive contexts in the latter data. In addition, the majority of respondents in 2011 belong to the dictionary's target group, which was not the case in 2007. Although respondents seem more satisfied with the new version of SLD, the two sets of comment field data contain largely the same kind of criticism, as for example in comments concerning (usually the lack of) lemmas and examples.

Keywords: dictionary use; learner´s dictionary; user comments

1 Introduction

Comment fields are often provided in questionnaires to give respondents the opportunity to expand upon what they have already reported. This article deals with comment field data from two web questionnaire surveys, performed in 2007 and 2011, about the use and users of the online monolingual *Swedish Lexin Dictionary*, SLD (http://lexin2.nada.kth.se/lexin/) (Hult 2008a; 2012). In the interval between the two studies the SLD underwent a comprehensive revision. One way of evaluating the success of the updating project is to compare respondents' comments on SLD before and after the revision. Considering the improvements made, are the respondents in the 2011 study happier dictionary users than those in the 2007 study? In an attempt to answer this question I will first briefly describe the updating project. I will then categorise the two sets of comment field data according to the types of comments expressed. Thereafter, I will examine what information categories are mentioned in the comments, and what is said about them. Finally, I will give voice to the users by quoting from the comment field data, focusing on the "Lemma", "Meaning" and "Example" categories.

In the age of electronic dictionaries, users can and do play a much more active role in the process of dictionary making. Lexicographers may disagree on exactly how active users should be in this process (cf. Lew, in press). Be that as it may, since users are the consumers of our products, the least we can do is listen to what they have to say about them.

2 The Swedish Lexin Dictionary and the Updating Project

Lexin, short for "lexicon for immigrants", was originally a series of paper dictionaries at the advanced beginner's level for users with Swedish as a second language. Lexin includes one monolingual Swedish part, the SLD, which is the focus of this article, and about twenty bilingual dictionaries between Swedish and non-Scandinavian immigrant languages in Sweden such as Arabic, Russian and Somali. The Swedish material served as a basis for the bilingual dictionaries. The dictionaries are primarily aimed at recent immigrants to Sweden who are just beginning to learn Swedish. They are intended to be easy to use even by people with limited reading ability and little or no experience using dictionaries. Consequently, in comparison with many other dictionaries, a user-friendly layout is extremely important (Gellerstam 1999: 5). The SLD (and the entire Lexin series) is primarily intended for reception and secondarily for production. The lemma selection has been carefully adjusted to the needs of the dictionary's target group. All lemmas have been augmented with comprehensive information typical of a general dictionary. In addition, SLD includes specific words covering special issues and institutions in Swedish society, along with factual information. About 1,800 elementary words are illustrated with pictures in a special section that covers 31 themes, such as *The Human Body (external and internal)* and *Cooking and Meals*.

The free online version of SLD and the Lexin series have been available since 1994. In addition to pictures, the online version features animations that demonstrate the meanings of 700 verbs. These are classified in fifteen main sections, such as *The Kitchen and Cooking, Emergency and Medical Service* and *Travel and Transport (City and Traffic)*. The animations make verbs otherwise difficult to make comprehensible through pictures – let alone definitions – much easier to understand. There are links from the dictionary entries to the pictures and animations. Use of the online version has steadily increased from the start. The site presently has about 20 million searches per month and the SLD has slightly more than one million searches per month (http://lexin2.nada.kth.se/statistik/html). The recent revision of the SLD encompassed both dictionary content and the website interface. Eventually, the bilingual dictionaries will also be revised.

The revision focused on strengthening the nature of the SLD as a learner's dictionary, especially its function as a reception dictionary (Malmgren 2012: 456). It is the dictionary's first comprehensive revision since it was released some thirty years ago. Naturally, one important task was to update the lemma selection, which increased by about 3,000 lemmas. A great many synonyms and antonyms were also added. This was done mainly to improve users' chances of understanding the entry words: "one single synonym – even if not a perfect one – can make the difference between understanding and not understanding" (Hult et al. 2010: 807). The synonyms and antonyms also serve the purpose of increasing the users' vocabularies. Moreover, the update included a thorough adjustment subdividing the senses of each lemma and placing the examples in the immediate vicinity of the sense being illustrated. A great deal of effort was also dedicated to more explicitly describing particle verbs, which are often stumbling blocks for learners of Swedish. The structure of the entry after the definition was changed slightly. Compounds now come first, followed by self-explanatory syntactic examples and then idioms. Idioms are now given in bold, emphasising their status as sub-lemmas of a sort. Valency information – now somewhat more transparent – is found at the end of each numbered sense. Furthermore, although most examples in older editions of SLD are good, many infinitival phrases have been replaced by full sentences. In a learner's dictionary mainly intended for reception purposes, examples are extremely important. They must be self-explanatory, contain no difficult words and "[I]deally, they should evoke a little scene from everyday life, with prototypical actants" (Hult et al. 2010: 807). In the entry *väntar* ("wait(s)"), for example, instead of the infinitival phrase *vänta på bussen* ("wait for the bus"), the full sentence *de fick vänta på bussen i tio minuter* ("they had to wait for the bus for ten minutes") is given. In addition, all words in the dictionary are clickable.

The changes accomplished in the updating project are consistent with the conclusions of Pálfi & Tarp (2009) on the subject of learner's dictionaries. For instance, they recommend more synonyms and antonyms, more explicit valency information and lemmatisation of idioms. In his specific theories concerning learner's dictionaries, Tarp (2009: 199-200) also emphasises the importance of describing semantic relationships. This is realised in SLD through the copious links from the dictionary entries to the picture themes and, for verbs, the animations.

Let us now turn to the comment field data of the two web questionnaire surveys.

3 Types of Comments Expressed in the Two Sets of Comment Field Data

In the 2007 study, almost one third of the respondents, 110 out of 360, wrote something in the questionnaire's comment field. Interestingly, considering that the dictionary's target group is (relatively) recently arrived immigrants, almost 60% of respondents reported Swedish as their native language.

In terms of length, the shortest comment consists of one word and the longest of 130 words, with a median value of 14 words. In Table 1 the comments have been categorised into five types, each accompanied by a translated example. Comments belonging to more than one category were divided and placed into all appropriate types, which explains how the 110 comments in total turned into 152 parts.

Type of comment	**Frequency**		**Examples (my translations)**
	Absolute	Relative	
Criticism	53	35	*You only have easy words like "cat" and "dog", but people know those words...I need explanations for more difficult words.*
Suggestions for improvement	48	31,5	*It would be fun if there were more technical terms or words related to certain subjects like biology, etc.*
Praise	30	20	*I think Lexin and [SLD] is a very good site, I use it often and it is the best look-up site I have ever found.*
Other	17	11	*I think all the items in question 8 are important to have full understanding in a context.*
Nonsense	4	2,5	*ROCK THE WORLD*
Total	**152**	**100**	

Table 1: The 2007 study: comment types, frequency and an example of each type, in descending order.

Comments containing criticism are nearly as numerous as comments giving suggestions for improvement. Combined, they represent 66.5% of the comments. One fifth of the comments contain praise. The "Other" category covers comments of a more neutral nature or comments that did not directly relate to SLD. With a few exceptions, these comments contributed very little substantive information. The nonsense comments speak for themselves; they are few in number and contain no information of interest.

Let us now turn to the 2011 study. Nearly half the respondents, 371 out of 802, wrote something in the questionnaire's comment field. Gratifyingly, a full 67% of the respondents belong to the dictionary's target group: people whose native language is not Swedish and who arrived in Sweden fairly recently (during the 21st century).

In terms of length, the shortest comment consists of one word and the longest of 168 words, with a median of 17 words. In Table 2 the comments have been categorised into six types, each accompanied by a translated example. One column was added, called "Temporary change noticed".[1] Again, comments belonging to more than one category were divided and placed in all appropriate types. In total, the 371 comments turned into 565 parts.

[1] The same week the questionnaire was displayed on the SLD website, the settings had been manipulated. Five categories had been removed from the dictionary articles with a view to indicating which categories users most preferred. Users were informed of the ongoing survey and encouraged to change the settings back to default if they wished. This change was mentioned in the comment field in about 15 questionnaires.

Type of comment	**Frequency**		**Example (my translations)**
	Absolute	Relative	
Praise	239	42	*Free service. Simple translation. Good examples on how to use words. (Originally written in English)*
Criticism	123	22	*You don't have enough words in Lexin.*
Suggestions for improvement	115	20	*Add more compounds, please!*
Other	69	12	*Didn't know about the pictures and videos.*
Temporary change noticed	15	3	*It's a pity you took away the examples and verb forms.*
Nonsense	4	1	*Huh.*
Total	**565**	**100**	

Table 2: The 2011 study: comment types, frequency and an example of each type, in descending order.

Compared to the 2007 study, the figures are distributed somewhat differently. Here, praise is the most frequent comment type, followed by criticism and suggestions for improvement. The positive comments (praise) are almost exactly equal in number to the suggestions for improvement and critical comments put together. Moreover, in relative figures there are twice as many positive comments compared with the 2007 study.

Notably, many comments ended up in the "Other" category. These are often part of longer comments; for instance, some respondents write something about themselves or their background ("I have dyslexia"*)*, others say what other dictionaries they use ("I use Tyda instead"), and still others simply add a comment like "examples are important for immigrants" or "hope it remains free of charge". Compared with the "Other" comments in the 2007 study, these comments often contribute more or less substantive information.

The data reported so far indicate that respondents are happier with the revised version of SLD than with its unrevised counterpart. Comparing the figures from before and after the updating project, we see that praise has increased by 13.2 percentage points, whereas criticism has decreased by 12.4 percentage points. Comments in the "Other" category have declined by 0.8 percentage points. Let us now move on to see what information categories users comment on and what they say about them.

4 Information Categories Commented On

In the 2007 version of SLD, before the revision, a dictionary article could have up to nine information categories. These are mentioned 137 times in the comment field data of the 2007 study, distributed as shown in Table 3.

Information category	**Positive context**	**Negative context**	**Other**	**Total**	
				Absolute	Relative
1. Lemma	9	91	4	104	76
2. Pronunciation	1	7	4	12	9
3. Part of speech	1	-	1	2	1,5
4. Inflection	-	1	1	2	1,5
5. Meaning	-	8	-	8	6
6. Compound	1	1	1	3	2
7. Example	-	3	-	3	2
8. Phrase	-	2	-	2	1,5
9. Picture theme	-	1	-	1	0,5
Total	**12**	**114**	**11**	**137**	**100**

Table 3: Frequencies of SLD information categories in the 2007 study.

"Lemma" is by far the information category most frequently commented on (76%). "Pronunciation" and "Meaning" are in second and third place with 12% and 8% of the comments, respectively. The remaining categories are all rarely commented on. As shown, occurrences in a negative context are most common.

In the 2011 revised version of SLD, a dictionary article may have up to thirteen information categories. These are mentioned 332 times in the comment field data of the 2011 study, distributed as shown in Table 4.

Information category	**Positive context**	**Negative context**	**Other**	**Total**	
				Absolute	Relative
1. Lemma	36	126	10	172	51,5
2. Pronunciation	4	12	2	18	5
3. Part of speech	-	1	-	1	0
4. Inflection	5	11	1	17	5
5. Abbreviation	1	1	-	2	0,5
6. Meaning	8	13	2	23	7,5
7. Antonym	1	5	-	6	2
8. Compound	-	13	-	13	4
9. Example	9	30	3	42	13
10. Valency	-	4	1	5	1,5
11. Phrase	4	15	1	20	6
12. Picture theme	4	3	2	9	3
13. Video sequence	1	1	2	4	1
Total	**73**	**235**	**24**	**332**	**100**

Table 4: Frequencies of SLD information categories in the 2011 study.

Again, "Lemma" is by far most the most frequently mentioned category in the comments. "Example" is in second place, followed by "Phrases", "Pronunciation", "Inflection" and "Meaning". The least frequently mentioned are "Part of speech" and "Abbreviation". The remaining categories received between four and thirteen comments. Again, occurrences in a negative context are the most common, but are significantly fewer than in the 2007 study, with a difference of 12.4 percentage points. Thus, analogue to that indicated in the previous section, respondents seem happier with SLD after the revision than before.

Now let us look at more authentic examples from the comment field data. I will focus on what the respondents said about the information categories "Lemma", "Meaning" and "Example".

4.1 Voices on Lemma

The typical praise related to lemma simply states that the respondent often finds the words they searched for in SLD, and are satisfied with what can be found. Practically all the critical comments concerning lemma in both studies complain about the dearth of them. Many respondents are critical of the absence of more difficult words. Here are a few examples concerning lemma taken from the two sets of data (my translations):

Study 2007

(1) Some words you search for aren't there, even though they are real words.

(2) Quite a few psychological or political terms/words aren't there.

(3) It is bad that there are no compounds or particle verbs, which are what we immigrants need the most help with.

(4) I want the difficult words; the words included are mostly for beginners, or basic level.

Study 2011

(5) Many compounds are missing as well as special terms like *phonology* and *autism*.

(6) Hard to find the meaning of long words.

(7) Have a feeling only the most basic words are there, not other words.

(8) Sometimes important words are missing.

Approximately 3,000 new lemmas were added in the revision, including many particle verbs, and many synonyms and antonyms were added to the dictionary articles. In this respect, we have fulfilled the respondents' requests for more lemmas of different kinds. The figures speak for themselves: in 2007, 87.5% of the comments on lemma were negative, compared with 73% in 2011, while 8.7% of comments were positive in 2007 compared with 21% in 2011. Respondents are clearly less negative and more positive after the updating project. They remain, however, quite critical of the lemma selection and express more or less the same type of criticism in 2011 as in 2007. Indisputably, no dictionary will ever succeed in fully satisfying users' need for more lemmas. A simple comment like "all words aren't there" is very telling in respect to users' expectations of dictionary's lemma selection. Along with the comments quoted above, this also reveals that many respondents do not clearly understand what type of dictionary the SLD is.

4.2 Voices on Meaning

In the 2007 study, there were no positive or neutral comments and only eight negative ones about the information category "Meaning". In the 2011 study, there are eight positive comments, thirteen negative comments and two neutral ones. Here are a few examples (my translations):

Study 2007

(9) More examples of the meaning of the word and how you can use the word.
(10) A definition of *naturopathic practitioner* is missing and an understandable explanation of *recruit*.
(11) Hard to find explanations of words.
(12) You only have easy words like *cat* and *dog*, but people know those words. I need explanations for more difficult words.

Study 2011

(13) Would like to have more examples of possible useful prepositions with a word and what meaning you get if you use a particular preposition.
(14) It should be possible to search for meaning/translation of expressions, not only single words.
(15) Can you try to give the exact meaning of words or an exact synonym?
(16) The definitions are good, short and concise and easy to understand.

The update included a thorough revision subdividing the senses of each lemma. Moreover, a great deal of effort was dedicated to more explicitly describing particle verbs which, as mentioned, often cause difficulties for learners of Swedish. Again, users' expectations are high, as they should be. In some respects, the respondents' views have been taken into account, for example in (9) and (13), and in other respects they have not, such as in (10). In a perfect world, all users would agree with the respondent quoted in (16).

4.3 Voices on Example

There were only three comments concerning "Example" in the 2007 study, all negative. Respondents want more examples, as expressed in the following comment: "have at least 2-3 examples for every word so you understand 100%". In the 2011 study, there are nine positive comments, thirty negative ones and three of a neutral nature. Here are a few examples (my translations):

(17) Too few examples on how to use words, hard to find synonyms, antonyms and proverbs.

(18) You should have more phrases and maybe a few more examples of each word or preposition.

(19) Good examples where you often find exactly the inflection you are looking for.

(20) Better if you could add more examples, particle verbs, past participle, proverbs, slang, which you have to learn and use every day.

As thE senses of each lemma were subdivided, so were the examples, placing them in the immediate vicinity of the sense being illustrated. The update also added many morphological examples in terms of transparent compounds and several more, and more extensive, syntactical examples. In the updated version there is at least one example per lemma. This may not be enough in the opinion of the respondent quoted above, but examples are nonetheless much more numerous than was the case before the revision. What this respondent, and hopefully many others, acknowledges and appreciates is that the examples have been expanded to full sentences to more clearly illustrate the meaning of the lemma.

5 Conclusion

Taken as a whole, the users' comments on SLD can offer lexicographers valuable information about the needs of the dictionary's target group and might provide them with relatively concrete ideas on how to improve the dictionary. The question, however, was whether the respondents are happier with the SLD now than before the updating project. Comment field data from a web questionnaire survey before the update were compared with corresponding data from after the update. And, yes, respondents appear to be happier with the SLD now than before the revision. Firstly, there are significantly more positive comments after the revision than before. Secondly, the information categories commented on are more often set in positive contexts in the latter data. We also know that the majority of the respondents in 2011 belong to the dictionary's target group, in contrast to 2007. This suggests that the users who are happier with the SLD are also more representative of the intended users of the dictionary.

Even though respondents seem more satisfied with the updated version of SLD, the two sets of comment field data contain largely the same kind of criticism. Not unusually, many respondents have both good and bad things to say about the SLD. These circumstances indicate that one cannot have too much of a good thing. Or, as one respondent simply writes: "I think you can do it better, but it's not bad". Moreover, "Lemma" was the information category most frequently commented on, while the other categories were much less frequently commented on. Presumably, there is more information to be extracted from the comment field data.

It should be noted that the two sets of data were not fully comparable. The comments in 2007 add up to not quite one third of the number of comments in 2011. The first questionnaire contained ten questions and the second was extended to twenty questions. In addition, the texts preceding the comment field box were not identically worded.

6 Discussion

There are comments where respondents more or less explicitly demonstrate their awareness of what type of dictionary they are presently using, but there are also comments which clearly reveal the respondents' lack of understanding of what type of dictionary SLD is. This may not be a problem, since the dictionary is free and users are also free to go elsewhere if they are not satisfied with what they find. On the other hand, perhaps the purpose of SLD and all the Lexin dictionaries could be made more explicit on the website and on the internet as a whole. If not, at least the pictures and animations should be marketed since "for non-native speakers of the language, definitions, however skilfully written, are not usually the best way to convey meaning" (Lew, in press), and pictures and animations can concretise the meaning of a word in a very enlightening way. Surprisingly and somewhat disappointingly neither pictures nor animations are mentioned more than a handful of times in the comment field data. Admittedly, this does not necessarily mean that the respondents are unaware of these features, but unfortunately that might just be the case.

In the practice of dictionary making, professional lexicographers are undoubtedly the linguistic experts and users' potential contributions in this area may be limited. I would argue, however, that there are other equally important areas where users may very well have useful suggestions for improvement, particularly in relation to web-related issues like the user interface.

7 References

Hult, A.-K. (2008a). Användarna bakom loggfilerna. Redovisning av en webbenkät i Lexin online Svenska ord. In: *LexicoNordica*, 15, pp. 73-91.

Hult, A.-K. (2012). Old and New Study Methods Combined. Linking Web Questionnaire with Log Files from the Swedish Lexin Dictionary. In: R. Vatvedt Fjeld, J. M. Torjusen (eds.) *Proceedings of the 15th Euralex International Congress, Oslo, 7-11 August 2012*, pp. 922-928.

Hult, A.-K., Malmgren, S.-G., Sköldberg, E. (2010). Lexin – a Report from a Recycling Lexicographic Project of the North. In: A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*, pp. 800-809.

Gellerstam, M. (1999): LEXIN – lexikon för invandrare. In: *LexicoNordica 6*, pp. 3-17.

Malmgren, S.-G. (2012). Från Svenska ord (Lexin) 3 till Svenska ord 4. In: B. Eaker, L. Larsson and A. Mattisson (eds.) *Nordiska studier i lexikografi 11. Rapport från Konferensen om lexikografi i Norden, Lund 24-27 May 2011,* pp. 454-465.

Lew, Robert in press. User-Generated Content (UGC) in Online English Dictionaries. In: A. Klosa, A. Abel (eds.) *OPAL – Online publizierte Arbeiten zur Linguistik* 2013. (Preprint).

Lexin Online, Statistics. Accessed at: http://lexin2.nada.kth.se/statistik/html [29/03/2014].

Pálfi, L.-L., S. Tarp (2009). Lernerlexikographie in Skandinavien – Entwicklung, Kritik und Vorschläge. In: *Lexicographica. International Annual for Lexicography 25/2009*. Tübingen, pp. 135-154.

Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner´s Lexicography. Tübingen: Max Niemeyer.

Swedish Lexin Dictionary Online. Accessed at: http://lexin2.nada.kth.se/lexin [29/03/2014].

Mobile Lexicography:
A Survey of the Mobile User Situation

Henrik Køhler Simonsen
Copenhagen Business School
hks.ibc@cbs.dk

Abstract

Users are already mobile, but the question is to which extent knowledge-based dictionary apps are designed for the mobile user situation. The objective of this article is to analyse the characteristics of the mobile user situation and to look further into the stationary user situation and the mobile user situation. The analysis is based on an empirical survey involving ten medical doctors and a monolingual app designed to support cognitive lexicographic functions, cf. (Tarp 2006:61-64). In test A the doctors looked up five medical terms while sitting down at a desk and in test B the doctors looked up the same five medical terms while walking around a hospital bed. The data collected during the two tests include external and internal recordings, think-aloud data and interview data. The data were analysed by means of the information scientific star model, cf. (Simonsen 2011:565), and it was found that the information access success of the mobile user situation is lower than that of the stationary user situation, primarily because users navigate in the physical world and in the mobile device at the same time. The data also suggest that the mobile user situation is not fully compatible with for example knowledge acquisition.

Keywords: Mobile lexicography; mobile user situation; mobile user

1 Introduction and Problem

Today, most users are always on and always connected, cf. (Google 2013:2), which reports that 84% of us use smartphones while they do other things, and users today in fact use their mobile devices in a large number of situations.

Users are already mobile, but the question is to which extent knowledge-based dictionary apps are designed for the mobile user situation. The objective of this article is to analyse and discuss the characteristics of the mobile user situation with a view to putting the user back in focus in dictionary apps.

2 Methodology and Empirical Basis

Ten medical doctors were asked to look up five medical terms by means of the dictionary app Medicin.dk, which is a knowledge-based medical resource used by most health care persons (HCPs) in the Danish health care system. As many as 15,000 users regularly update the medical app Medicin.dk, which indicates that the app is widely used by a variety of users. According to (Dolan 2012) everything in medicine is going mobile and both patients and physicians are changing behaviour in line with developments in health technology.

The test persons accessed the medical terms by means of the app Medicin.dk on an iPhone 4S, which was wirelessly connected to a PC by means of Reflector, cf. http://www.airsquirrels.com/reflector/. The medical doctors were asked to participate in two tests. In test A the test persons were asked to look up five medical terms while sitting down at a desk. In test B the ten test subjects were asked to look up the same five terms while slowly walking around a hospital bed.

The two tests were designed to imitate two typical user situations for many doctors: knowledge acquisition and knowledge checking prior to patient consultation and knowledge checking during a patient consultation.

The five tasks included looking up the five proper names Terbasmin (asthma), Tamoxifen (breast cancer), Antepsin (ulcer), Tredaptive (cholesterol) and Fludara (leukaemia) and can be summarized as follows:

- Task 1: Look up Terbasmin – to find information
- Task 2: Look up Tamoxifen – to find and extract information about side effects to be able to inform patient
- Task 3: Look up Antepsin – to find and extract information about dosage to be able to check prescription
- Task 4: Look up Tredaptive – to find and extract information about dosage to be able to inform patient
- Task 5: Look up Fludara – to find and check spelling of term to be able to write a text.

Both tests were recorded from the "inside" by means of Reflector, and at the same time the user activities were recorded from the "outside" by means of a digital camera. In addition to the recordings from the "inside" and the "outside", the empirical basis also includes think-aloud-data, as the test persons were asked to think aloud and verbalize what they did and saw etc. To deduce additional qualitative comments, the empirical basis also includes interview data as the test persons were interviewed before and after the tests.

3 Theory

Related work with direct relevance for this survey includes a number of studies of how users interact with mobile devices, such as (Pedersen & Engrob 2008), (Church et al. 2009) and (Ehrler et al. 2013). In addition to theoretical considerations on user interaction and mobile devices, this work also includes selected theoretical considerations on lexicography such as (Tarp 2006), (Verlinde et al. 2010) and (Simonsen 2011).

Pedersen & Engrob (2008) discusses a number of interesting usability tests with mobile devices. The objective of their work was to discuss which interaction technique was most suitable for mobile users. Pedersen & Engrob (2008) asked eight students to walk on a running machine while interacting with a PDA. The focus of their tests is not completely comparable with this survey, but it is highly relevant. Pedersen & Engrob (2008) found that the test persons used different interaction techniques in different user situations.

Another highly relevant contribution in this area is Church & Smyth (2009). Church & Smyth (2009) conducted a number of surveys of the mobile information needs of different users. Church & Smyth (2009) asked 20 test persons to participate in a four week long diary survey during which the test persons made notes on their mobile information needs and user situations. Church & Smyth found (2009) that user situations can be categorized in five overall categories such as Navigational, Informational, Transactional, Geographical and Personal Information Management. The informational need, cf. (Church & Smyth 2009:251) is the most important need and is focused on the goal of obtaining information about a topic.

Ehrler et al. (2013) reports on an evidence-based survey of user-interface design on handheld devices in health care. The usability test discussed by Ehrler et al. (2009) aimed at acquiring evidence about the quality of data recorded through interfaces on mobile devices and the test showed that the majority of test persons preferred the simpler models for data entry geared to the actual healthcare environment – a finding which was also clear on the basis of this survey.

Finally, a number of contributions on lexicography and information science are also relevant for this analysis. First of all the many contributions on the user and the lexicographic functions as discussed by for example (Tarp 2006) are necessary to understand the characteristics of the user. Furthermore, (Simonsen 2009), (Simonsen 2011a) and (Simonsen 2011b), who builds on (Verlinde et al. 2010), is relevant for the understanding of user research, the mobile user and mobile lexicography. Simonsen (2011a:565) makes the case for the information scientific star model as shown in figure 1 below.

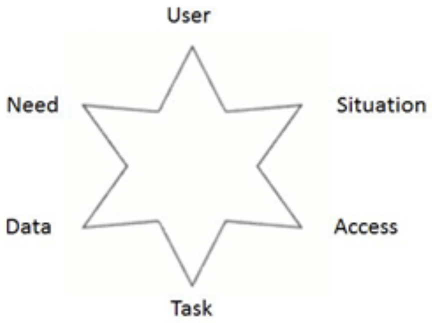


Figure 1: Information Scientific Star Model.

The information scientific star model proposed by Simonsen (2011:565) is applied on the analysis of the mobile user situations below, and it is argued that modern dictionary app development should be based on these six factors. The above model builds on (Verlinde et al. 2010:5), who argues that "relevant data should be retrieved and processed according to the external situation that motivated consultation in the first place, and the information needed to change a state of affairs in the outside world should be operationalized". Verlinde et al. (2010:5) make the case for a "lexicographic triangle" consisting of user, data and access, but it is argued that the analysis and design of information tools should be based on much more than that. Consequently, the information scientific star model was developed.

The star model includes the six dimensions: user, situation, access, task, data and need. Explicit considerations are required on the competency profile of the user, the user situation, the way the user accesses information, the type and complexity of the task, the type and complexity of data and the inherent need of the user. As the data suggest a number of these dimensions have been neglected during the design and development of the app Medicin.dk.

4 Results and Discussion

Ten medical doctors were asked to look up five medical terms by means of the dictionary app Medicin.dk and two tests were carried out. As already mentioned, the empirical data include ten recordings from the inside, ten from the outside, twenty think-aloud data recordings and interview data from all ten test persons.

An overview of some of the comparable answers offered by the test persons during the interviews is shown below in figure 2.

As will appear from figure 2 below, all ten medical doctors in fact prefer the website version of Medicin.dk when asked the question "Which platform and which user situation do you prefer"? This finding is in fact very much in line with (Ehrler et al. 2013), who argue that test persons seem to prefer the most simple data entry and data access model. The finding also seems to support the overall theoretical approach proposed by (Simonsen 2011), who makes the case for a balanced approach and focus in lexicography and information science. It may be argued that the finding is not that surprising, because medical doctors in public hospitals are not issued with a mobile device nor do their working conditions match knowledge acquisition by means of a mobile device. Furthermore, doctors often look for complex data and documents from many different sources and again the mobile device is not an obvious tool to use.

However, as will appear from figure 2 below seven out of ten doctors state that they in fact use their mobile devices professionally and nine out of ten doctors say that they use their mobile phone while moving around, so researching the mobile user situation is highly relevant.

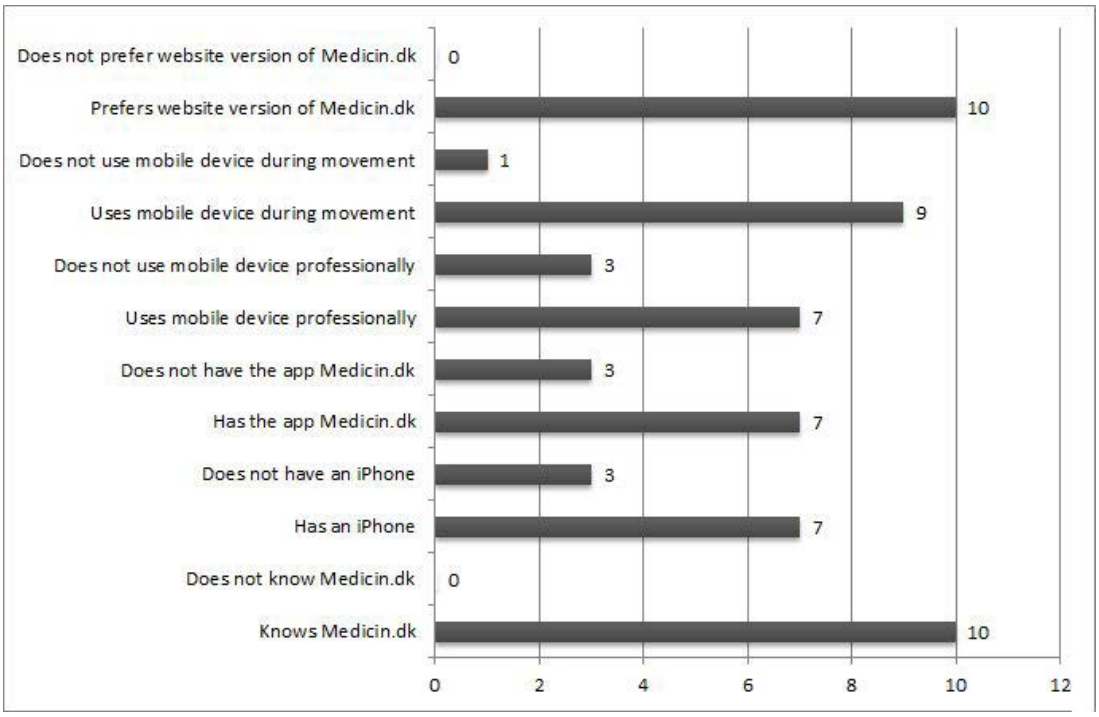


Figure 2.: Overview of Interview Data.

The six dimensions described in the information scientific star model, cf. (Simonsen 2011:565) are encapsulated by the answer given by test person 4, who is a 52-year old female medical doctor. Test person 4 states "I prefer the website version of Medicin.dk, if my problem is complex. The app and the iPhone are handy, if I suddenly have a problem that I know can be solved by the app. However, if I need more knowledge I would rather use the website".

What this statement seems to indicate is that the concrete task at hand more or less dictates the actual user situation and vice versa. Furthermore the task also dictates the amount and type of data sought by the user and in fact also the data access method needed. This correlation is in fact observable in most of the statements offered by the ten test persons, and the interview data show that the mobile user situation and cognitive lexicographic functions is not a perfect match. What is needed is an adaptable, dynamic and situational tool, which features seamless adaptation of data based on location-based services (LBS) and dynamic and situational presentation of data designed for the concrete task at hand and the competence profile of the user.

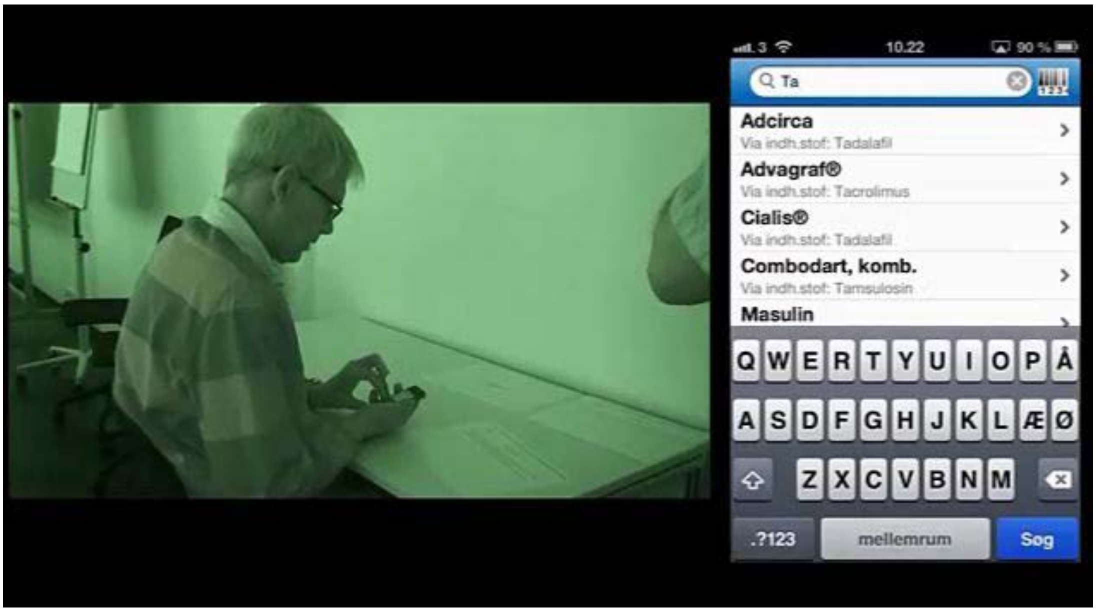
The quantitative test data from Test A and Test B also support these arguments, see figure 3 and 4 below. Figures 3 and 4 below show how the two tests were carried out and illustrate the stationary user situation and the mobile user situation.

Figure 3: Stationary Test.

Figure 4: Mobile Test.

The numbers in figure 8 below are numerical representations of a systematic evaluation of each test person's information access success in each situation. As will appear from figure 5 the five columns list the five tasks that the ten doctors were asked to do during the stationary test and the mobile test. The term information access success covers an evaluation of the search speed, search quality, focus ability, device interaction ability of each test person on a scale from 1 to 10, where 1 is low information access success and 10 is high information access success. Each number thus represents an overall evaluation of each situation based on the many internal and external recordings. The interview data and think-aloud data also substantiate the numerical evaluations made.

Figure 5 below shows how test person 3 (TP3), who is a 62-year old medical doctor, solves the task "Look up Tamoxifen and extract information about side effects to be able to inform a patient about the most common side effects" in Test A – that is while he is sitting down at a desk.



Adcirca
Via indh.stof: Tadalafil

Advagraf®
Via indh.stof: Tacrolimus

Cialis®
Via indh.stof: Tadalafil

Combodart, komb.
Via indh.stof: Tamsulosin

Masulin

Figure 5: TP3 solving Task 2 during Test A - Outside vs. Inside.

Figure 5 is a snap shot of two video recordings, which originally were recorded at the same time, but they have been edited by means of a video editing tool so that the two recordings can be shown at the same time as a picture-in-picture video.

The entire edited recording shows how TP3 sits at the table in the left hand side of the picture interacting with the mobile device in the physical world, and in the right hand side of the picture TP3's search behaviour is shown from the inside. The left hand side video was recorded by means of a standard digital camera and the right hand side of the video was wirelessly recorded by means of Reflector, cf. http://www.airsquirrels.com/reflector/.

The edited picture-in-picture video, which is based on aligned time codes to show a time-aligned video of the user situation seen from both the inside and the outside, gives a detailed picture of how TP3 solves a concrete task and it shows how a medical doctor uses a mobile phone while sitting down at a desk to look up complex medical information.

In comparison with the stationary user situation, Figure 6 below shows the mobile user situation, that is TP3 solving the same task (Task 2) during Test A. Again the actual user situation and user behaviour are recorded from the outside and the inside and Figure 11 is also a snap shot of an edited time-aligned, picture-in-picture video.

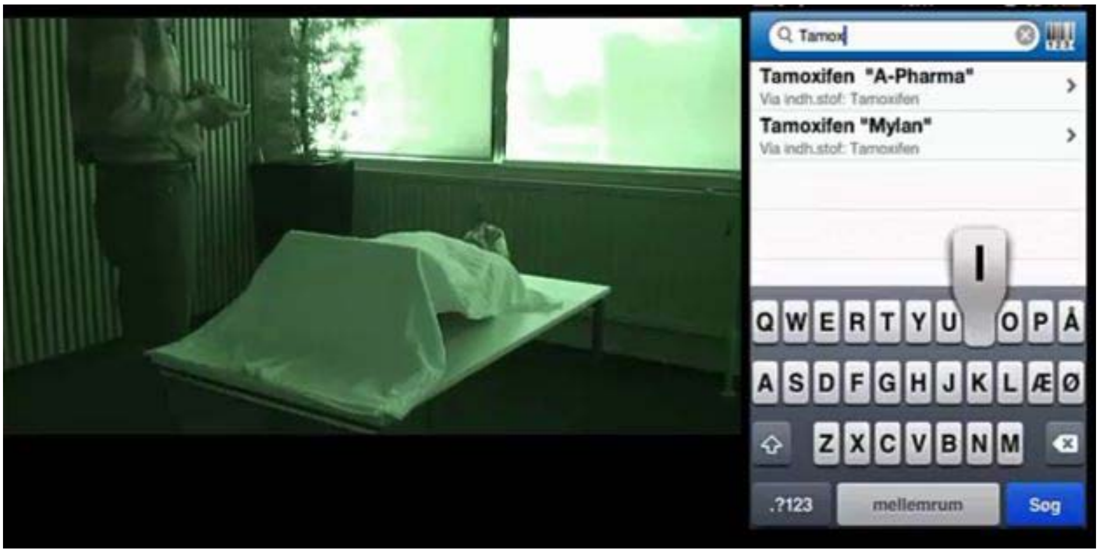
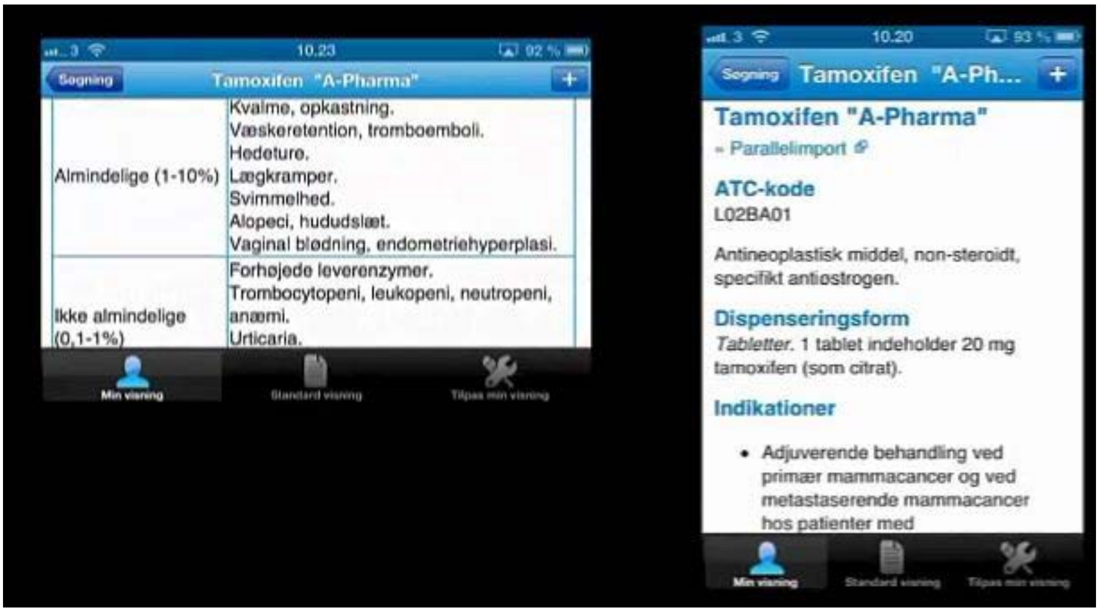


Figure 6: TP3 solving Task 2 during Test B – Outside vs. Inside.

Figure 6 above shows how TP3 walks around a "hospital bed" while solving task 2. The video gives a detailed picture of how TP3 uses a mobile phone to look up complex medical information while moving around at the same time.

A comparison of the two user situations shows that the access speed, that is from the moment the test person started the information access operation to the moment he ended the search operation, is higher during Test A than during Test B. That is in fact not surprising, because users can focus on the search operation and the mobile device while sitting down, which is in contrast to the mobile user situation where users also have to allocate cognitive effort on navigating in the physical world.

The differences between the two user situations become clearer when the two recordings from the inside are edited and contrasted. Figure 7 below shows a snap shot of the time-aligned edited picture-in-picture video of how TP7 solved Task 2 while sitting on the left hand side (Test A) and walking around (Test B) on the right hand side.



Almindelige (1-10%)	Kvalme, opkastning. Væskeretention, tromboemboli. Hedeture. Lægkramper. Svimmelhed. Alopeci, hududslæt. Vaginal blødning, endometriehyperplasi.
Ikke almindelige (0,1-1%)	Forhøjede leverenzymer. Trombocytopeni, leukopeni, neutropeni, anæmi. Urticaria.

Tamoxifen "A-Pharma"

- Parallelimport

ATC-kode

L02BA01

Antineoplastisk middel, non-steroidt, specifikt antiøstrogen.

Dispenseringsform

Tabletter. 1 tablet indeholder 20 mg tamoxifen (som citrat).

Indikationer

- Adjuverende behandling ved primær mammacancer og ved metastaserende mammacancer hos patienter med

Figure 7: TP7 solving Task 2 during Tests A and B – Inside.

The video shows that TP7 is much faster at locating the section on side effects while sitting down than while moving around. The information access speed is clearly higher when sitting down than when moving around. Another interesting fact is that TP7, just as three other test persons, chose to use the mobile device horizontally allowing the screen to show more text. This result also appeared for TP8, who also chose to use the mobile device in horizontal position. On the basis of these results it may be argued that users tend to use mobile devices like small computers while sitting down (the horizontal position), which in fact the video recordings from the outside also seem to document.

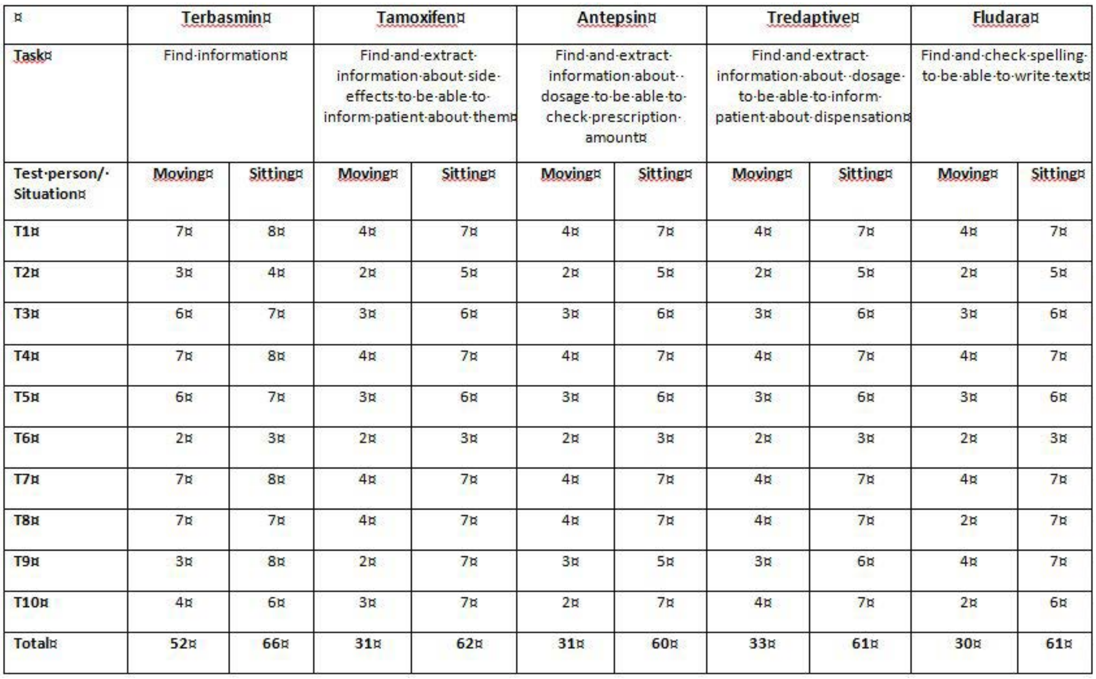
The many recordings from the inside and the outside are systematized and tabulated in Figure 8 below. The many numbers in figure 8 are numerical representations of a systematic evaluation of each test person's information access success in each situation. The term information access success covers an evaluation of three factors: search speed, search quality and device interaction ability.

The search speed was relatively easy to measure and is based on the time recorder in the many recordings. The measure used here was time.

It was far more difficult to precisely measure the search quality and device interaction ability. The evaluation of the search quality was partly based on an assessment of the quality of the search result found by the test person. The evaluation was based on an analysis of the think-aloud data where the test person described what he did and found and an analysis of the video recordings. The most important measure in this analysis was the test person's ability to quickly find the right information and verbalize it as think-aloud data. The measure used here was the ability to find the right information.

The device interaction ability was equally difficult to accurately measure. The evaluation of the test person's ability to use the device effectively was partly based on an analysis of the video recordings and the think-aloud data, which together made it possible to describe each test person's ability to use the device effectively. The measure used here was the ability to use the device effectively.

The data show that the information access success of the ten medical doctors was higher when they sat down at a desk than when they walked around a hospital bed. The data also seem to suggest that the task itself and the cognitive complexity of the information dictate the degree of information access success. In other words, simple and easy-to-find information correlates with high information access success while on the other hand complex and hard-to-find information yields lower information access success. This is clear when the two tasks "Find information" and "Find and extract information about dosage to be able to check prescription amount" are compared.



	Terbasmin		Tamoxifen		Antepsin		Tredaptive		Fludara	
Task	Find information		Find and extract information about side effects to be able to inform patient about them		Find and extract information about dosage to be able to check prescription amount		Find and extract information about dosage to be able to inform patient about dispensation		Find and check spelling to be able to write text	
Test person/ Situation	Moving	Sitting	Moving	Sitting	Moving	Sitting	Moving	Sitting	Moving	Sitting
T1	7	8	4	7	4	7	4	7	4	7
T2	3	4	2	5	2	5	2	5	2	5
T3	6	7	3	6	3	6	3	6	3	6
T4	7	8	4	7	4	7	4	7	4	7
T5	6	7	3	6	3	6	3	6	3	6
T6	2	3	2	3	2	3	2	3	2	3
T7	7	8	4	7	4	7	4	7	4	7
T8	7	7	4	7	4	7	4	7	2	7
T9	3	8	2	7	3	5	3	6	4	7
T10	4	6	3	7	2	7	4	7	2	6
Total	52	66	31	62	31	60	33	61	30	61

Figure 8: Overview of Test Data.

When asked the question "Do you use your mobile device while moving?" test person 3 states "Yes – when I suddenly think of a medical question that I would like to look up, but I also use my mobile phone in other situations". This answer is somewhat in contrast to the answer provided by test person 5, who states "No – not really. I mostly use my mobile phone when I am sitting down because I think the screen is too small and my fingers are too big for the key pad screen".

Interestingly, test person 5 is a 61-year old male medical doctor, and is the oldest test person, which seems to indicate that age plays a role in mobile information access behaviour as does age in the discussion of digital natives, who are tech-savvy young people to whom digital technology is an integrated part of their lives, cf. (Prensky 2001) for a detailed discussion of digital natives.

It was also found that the information access speed and quality of the mobile, punctual user situation is somewhat lower than the stationary, punctual user situation. The many recordings from both the inside and the outside clearly show that the test persons need to navigate both in the physical world and in the mobile device user interface. They stop walking during the interaction with the mobile device, because they also need to look up and navigate in the room.

When asked the question "What do you think of the mobile user situation?" test person 1 states"I do not think that there is a big difference between moving around and sitting. Okay – maybe you spend more time on the search operations when you walk around, because you have to look up and see where you are" and test person 1 states "As long as I stop up and stand still I actually think it works fine".

To substantiate the argument about information access speed, test person 7 is much faster at locating the section on side effects while sitting down than while moving around. The information access speed is clearly higher when sitting down than when moving around. Another interesting fact is that test person 7, just as three other test persons, chose to use the mobile device horizontally allowing the screen to show more text. This result also appeared for test person 8, who also chose to use the mobile device in horizontal position. On the basis of these results it may be argued that users tend to use mobile devices like small computers while sitting down (the horizontal position), which in fact the video recordings from the outside also seem to document.

The 5-inch screen size of a standard smartphone such as the iPhone is simply not enough. Size does matter when it comes to successful information access and the layout and design of dictionaries has always been relevant for lexicography, cf. for example (Almind 2005) and (Almind & Bergenholtz 2007). This is very much still the case as the data presented above suggest. The problem is that the human-mobile interaction is not optimal. The input device (the finger) and the small letters shown on the 5-inch screen are not a perfect match as one of the test persons surveyed actually also verbalize. Obviously, it would be logical just to call for bigger screens, but that would be naïve because smartphones are in fact supposed to be small. However, HUD technology may at some point allow us to display dictionary data in HUD format (Head-up Display) where information is visually size enhanced and relayed to the user surroundings, but it will be some years before that technology becomes commoditized. A number of theoretical contributions discuss mobile design and mobile usability, for example (Budiu & Nielsen 2013), who make a very strong case for more usability research in mobile design, (Cerejo 2012), who discuss the many elements of the mobile user experience, including the more social and personal elements of mobile user personas and of course also (Nielsen 2011), who offers a myriad of practical and easy-to-use instructions on mobile design.

However, what we can do at this point is to design the actual dictionary app in such a way that intelligent search engines and easy-to-use interfaces facilitate easy information access. As the data of the survey suggest, simple search engines with a simple search field and a simple TOC-like display of the dictionary data are preferred by most users. Scrolling through large text blocks reduces the information access success of the ten medical doctors surveyed.

It was also found that the information access success of the ten medical doctors was drastically reduced in cognitive user situations, that is when they were asked to solve cognitively-based problems like task 2, task 3 and task 4, which were all about locating complex information with a view to making decisions as to side effects, dosage and how to take the medicine etc.

This finding is also expressed by test person 7 who states "If I have to look a little bit deeper into a question then I clearly prefer the computer. I would definitely use the computer if I were to prescribe medicine that I have never used before". In other words, it was found that the mobile user situation and cognitive lexicographic functions is not a perfect match.

All this in fact seems to suggest that mobile lexicography needs to reinvent itself and take into account the six dimensions proposed above by (Simonsen 2009). This contention seems to be supported by (Church & Smyth 2009:255-256), who state that: "…mobile users are on-the-move and as such are interested in locating different types of content. We found context to be a very influential factor in many mobile information scenarios and as such argued for the need for new types of context-sensitive mobile interfaces that take full advantage of temporal, location, and preference-based contexts".

A similar argument was made by (Leroyer & Kruse 2011: 411-415), who describe a pragmatic data presentation and user interface in a French/Danish Real Estate e-Dictionary. Leroyer & Kruse (2011) make the case for a situational user interface, which definitely is the way forward in mobile lexicography.

However, mobile lexicography should not only be based on temporal and situational dimensions. Mobile lexicography is different from Internet lexicography and very much different from paper lexicography. Mobile lexicography is unique, because the user very often is mobile and on the move when using his device.

That very fact calls for new theoretical considerations and on the basis of the empirical data and the discussion above the following mobile lexicography principles can be identified.

Mobile user principle

The mobile user is on the move and needs and accesses information while on the go. This makes the mobile user punctual, impatient, imprecise and preoccupied with other things.

Mobile situation principle

The mobile user situation is characterized by being volatile, punctual and by often taking place while the user does other things. The mobile user typically checks knowledge and performs simple searches. The mobile user situation primarily supports simple, punctual, communicative lexicographic functions, and is not suited to support complex, cognitive lexicographic functions.

Mobile data access principle

The mobile user navigates in both the physical world and in the user interface of the mobile device at the same time. This calls for a very simple and easy-to-use data access method for example a very intelligent semasiological search engine or even better a voice-activated search engine like Siri in an iPhone.

Mobile data principle

The mobile user situation also dictates the type and complexity of the mobile data. The size of the user interface and the punctuality of the user situation mean that complex data and long text segments are not optimum mobile data.

5 Conclusion

This article discussed the mobile user situation, and it was demonstrated that medical doctors prefer the website version instead of the app version. It was also found that the information access success of the mobile user situation is lower than that of the stationary user situation, primarily because users are required to navigate in both the physical room and in the mobile device. It was also found that the mobile user situation is not at all suitable for solving cognitive lexicographic problems such as for example knowledge acquisition etc.

On the basis of the survey it is argued that classic lexicographic virtues such as attention to the characteristics of the user situation, the task, the type of user and the presentation of data seem to be in demand in app development. The data provided in a dictionary app must be adapted to the mobile user situation and the data access structure of the app should take into account the limitations of the mobile user situation and should be task-dependent. The empirical data and the discussion led to the formulation of four principles on mobile lexicography.

Users are already mobile, but lexicography does not seem to be up to speed with the users. A dictionary app should satisfy concrete and potential lexicographic needs. Consequently, further research in mobile lexicography is needed – to put the user back in focus.

6 References

Almind, Richard (2005): Designing Internet Dictionaries. In: *Hermes, Journal of Linguistics no 34-2005*, pp. 37-54.

Almind, Richard & Bergenholtz, Henning (2007): Klæder skaber folk: Om layout i ordbøger. In: *Hermes – Journal of Language and Communication Studies no 39-2007*, pp. 31-47.

Budiu, R., & Nielsen, J. (2013). In Rimerman S., Walker A. M. (Eds.), Mobile usability. Berkeley: The Nielsen Norman Group.

Church, Karen & Smyth, Barry (2009): Understanding the intent behind mobile information needs. In: *IUI 2009 International Conference on Intelligent User Interfaces*, pp. 247-256.

Cerejo, L. (2012). The elements of the mobile user experience. Mobile design patterns (1st ed., pp. 5-20). Freiburg, Germany: Smashing Media GmbH.

Dolan, Pamela Lewis (2012): Everything in medicine is going mobile. In: amednews.com. Accessed at http://www.ama-assn.org/amednews/m/2012/03/26/bsa0326.htm [04/07/2012].

Ehrler, Frederich, Walesa, Magali, Sarrey, Evelyne, Lovis, Christian (2013): Evidence-based User-Interface Design. In: *Studies in health technology and informatics*, pp. 57-61.

Google (2013): *Our Mobile Planet*: Accessed at: http://services.google.com/fh/files/misc/omp-2013-dk-en.pdf [01/09/2013].

Leroyer, Patrick & Kruse, Liselotte (2011): Ejendomsordbogen fransk/dansk: ny integreret e-ordbog. In. Nordiska studier i lexikografi 11 - 2011, pp. 405-417.

Nielsen, Jakob (2011): Mobile Usability Update. In: useit.com. Accessed at: http://www.useit.com/alertbox/mobile-usability.html [04/07/2012].

Pedersen, Anders Fritz & Engrob, Jan Hyldgaard (2008): Interaktion under bevægelse: Et komparativt studie af interaktionsteknikker til arbejde med komplekse data på håndholdte enheder, AAU, 60 sider.

Prensky, M. (2001): Digital natives, digital immigrants part 1. On the Horizon, 9(5), 1–6: Accessed at http://www.emeraldinsight.com/journals.htm?issn=1074-8121 1 [01/04/2014].

Reflectorapp.com (2012): Accessed at: http://www.airsquirrels.com/reflector/ [02/01/2012].

Simonsen, Henrik Køhler (2009): Se - og du skal finde: en eyetrack-undersøgelse med særlig fokus på de leksikografiske funktioner. In: *Nordiske studier i leksikografi 11. Rapport fra Konference om leksikografi i Norden. Finland 3.-5. juni 2009.* Tampere: Nordisk forening for leksikografi 2009, pp. 274-288.

Simonsen, Henrik Køhler (2011a): User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. In. *Hermes - Journal of Language and Communication in Business, 46-2011, pp. 75-102.*

Simonsen, Henrik Køhler (2011b): Et informationsvidenskabeligt serviceeftersyn af Medicin.dk. In. *Nordiska studier i lexikografi 11 - 2011*, pp. 563-574.

Tarp, Sven (2006): *Leksikografien i grænselandet mellem viden og ikke-viden: Generel leksikografisk teori med særlig henblik på lørnerleksikografi.* Doktorafhandling. ASB.

Verlinde, Serge, Leroyer, Patrick, Binon, Jean, (2010): Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats. In: *International Journal of Lexicography*, Vol. 23, Issue (1) 2010. S. 1–17.

Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht

Martina Nied Curcio
Università degli Studi Roma Tre
martina.nied@uniroma3.it

Abstract

Die Benutzung von Smartphones, iPhones, iPads und Tablets im Fremdsprachenunterricht scheint den Studierenden den schnellen und immediaten Gebrauch von Online-Wörterbüchern zu garantieren, sowie unbegrenzte Recherchemöglichkeiten zu bieten, so dass sprachliche Schwierigkeiten direkt überwunden werden können. Doch sieht es danach aus, als würden die Fremdsprachenstudierenden das Potential nicht ausnützen und sich auf zweisprachige Wörterbücher und Übersetzungsprogramme konzentrieren. Auch ihr Benutzerverhalten in Bezug auf Online-Wörterbücher scheint dem der Verwendung von Print-Wörterbüchern ähnlich zu sein. Eine empirische Untersuchung mit italienischen Studierenden der Germanistik zur Smartphone-Benutzung zeigt, wie und wofür sie hinsichtlich lexikalischer Fragen das Smartphone verwenden, welches Benutzerverhalten sie generell, aber auch in spezifischen Übersetzungsaufgaben, an den Tag legen und wie sie selbst über ihre Benutzung mit dem Smartphone bei bestimmten Schwierigkeiten reflektieren. Die Ergebnisse vermitteln erste Eindrücke und zeigen Trends auf, die Ausgangspunkt für weitere umfangreichere Forschungen sein können, die ihrerseits der Fremdsprachendidaktik und der lexikographischen Praxis wichtige Impulse geben können. Interessant in diesem Zusammenhang sind die „ricerche incrociate“ (‚cross research‘) und die Konsultation von multilingualen Online-Wörterbüchern mit der Konsultation des Englischen als „Sandwichsprache“.

Keywords: Benutzung von Smartphones; Online-Wörterbücher; Recherchekompetenz; Lexikographie und Fremdsprachendidaktik

1 Einleitung

Noch bis vor wenigen Jahren haben die Studierenden einer Fremdsprache im Fremdsprachenunterricht spezifische lexikalische Lücken beim Umkodieren in die Fremdsprache mit einem bilingualen Wörterbuch (oft sogar in Taschenbuchformat (sic!), vgl. Nied Curcio 2011) versucht zu schließen. Heute holen sie während des Fremdsprachenunterrichts spontan ihr Smartphone[1] hervor, um sprachli-

[1] Der Begriff *Smartphone* wird hier als Hyperonym für *iPhone*, *smartphon*e, *iPad* und *Tablet-Computer* verwendet, d.h. sämtliche Produkte, die in Form eines Mobiltelephons oder Mini-Computers die Funktionalität und Konnektivität eines Computers übernehmen.

che Schwierigkeiten anhand von Online-Informationen oder mit Hilfe von Apps zu überwinden.[2] Es ist offensichtlich, dass die Verwendung von Smartphones den Fremdsprachenlernenden die verschiedensten, fast unbegrenzten Möglichkeiten bietet, um bestehende sprachliche Schwierigkeiten in kürzester Zeit überwinden zu können. Auch die Möglichkeit, die Struktur des Online-Wörterbuchs durch Notizen, Lesezeichen, Verweise, Schlagwörter u.a. den individuellen Bedürfnissen anzupassen, die Wörterbucheinheiten zu aktualisieren, zu erweitern, zu diskutieren, sind verlockend. Im Bereich der Wörterbuchbenutzungsforschung liegen bisher – in Bezug auf die Übersetzung und die Fremdsprachendidaktik – noch relativ wenige empirische Forschungen vor (vgl. Mackintosh 1998; Nied Curcio 2011), auch wenn sich die Publikationen, die die Nutzerperspektive in den Vordergrund stellen, in den letzten Jahren vermehrten (Tarp 2011; Taljard, Prinsloo & Fricke 2011; Boonmoh 2012; de Schryver, Prinsloo 2011; Domínguez Vázquez, Mirazo &Vidal 2013, Müller-Spitzer 2013). Es hat sich mittlerweile bestätigt, dass Online-Wörterbücher häufiger als Printwörterbücher verwendet werden; die ersten Beiträge zu empirischen, nutzerorientierten, Untersuchungen sind erschienen.[3] Publikationen zur spezifischen Verwendung von Smartphones hinsichtlich lexikographischer Fragen in der Fremdsprachendidaktik sind mir derzeit nicht bekannt. Interessant ist m.E. eine spezifische Analyse zur Benutzung von Smartphones gerade deshalb, weil dessen Verwendung über eine „reine" Wörterbuchbenutzung hinausgeht, weitere Informationsquellen einschließt und einen Einblick in die generelle Recherchekompetenz der Fremdsprachenlernenden gibt, denn die Lernenden sind in ihrer Wahl der Recherchetools und der Informationsquelle (z.B. Online-Wörterbücher bzw. Wörterbuch-Portale, Enzyklopädien, Suchmaschinen, Foren, usw.) völlig frei. Meiner Beobachtung nach benutzen die Fremdsprachenlernenden zurzeit wie selbstverständlich das Smartphone und es scheint, als würden sie es auch während des Fremdsprachenunterrichts verwenden,[4] und sich dabei auf den Gebrauch von zweisprachigen Internet-Wörterbüchern – und dort auf die immediate Suche nach dem passenden Äquivalent in der anderen Sprache – zu konzentrieren; das technische Potential und seine Recherchemöglichkeiten werden anscheinend nicht ausgenutzt. Es sieht demnach aus, als wäre die Benutzung des Smartphones ähnlich der Verwendung von zweisprachigen Print-Wörterbüchern, d.h. die Lernenden verwendeten insbesondere bilinguale Wörterbücher (vgl. Engelberg & Lemnitzer 2004; Albrecht 2005; Corda & Marello 2004: 82), sie würden meist nur das 1. Übersetzungsäquivalent in Be-

[2] Das Nachschlagen in Printwörterbüchern wurde immer als Last empfunden (vgl. Hulstijn, Hollander & Greidanus 1996 in Engelberg &Lemnitzer ^{2}2004: 81) und ihre Verwendung scheint nun durch Online-Wörterbücher (vgl. Domínguez, Mirazo & Vidal 2013) abgelöst worden zu sein. Das Smartphone und die Möglichkeit, schnell und (fast) überall auf online-Informationen zurückgriffen zu können, scheint diesen Trend voranzutreiben.

[3] Für einen Überblick über den Gebrauch von Online-Wörterbüchern s. Möhrs & Müller-Spitzer (2013), spezifische empirische Untersuchungen wurden von Domínguez, Mirazo &Vidal (2013) und Dóminguez Vázquez, Mollica & Nied Curcio (2014) durchgeführt.

[4] Diese Tatsache ist nicht selbstverständlich, war doch der Gebrauch von Wörterbüchern während des Fremdsprachenunterrichts lange Zeit untersagt, da das Wörterbuch zu sehr in Verbindung mit der Grammatik-Übersetzungsmethode gebracht wurde und viele Lehrpersonen zudem der Meinung waren, die Kenntnisse der Fremdsprache ohne Wörterbuch abprüfen zu müssen. Außerdem gaben sie den einsprachigen Wörterbüchern den Vorrang und die Verwendung von Smartphones an pädagogischen Institutionen ist zudem oft verboten.

tracht (vgl. Atkins & Rundell 2008) ziehen und schenkten eventuellen metasprachlichen Notationen keine Beachtung (vgl. Nied Curcio 2011: 204; Dóminguez Vázquez, Mollica & Nied Curcio 2014). Um genauere Informationen zum Gebrauch des Smartphones während des Fremdsprachenunterrichts zu erhalten, wurde von mir im März 2014 eine empirische Untersuchung im akademischen DaF-Unterricht in Italien durchgeführt. Diese erste Untersuchung verfolgt ein wichtiges Ziel, nämlich auszuloten, welche Forschungsfragen in Bezug auf den Gebrauch von Online-Wörterbüchern, Smartphones und der Recherchekompetenz von Fremdsprachenlernenden noch offen sind und welche spezifischen empirischen Untersuchungen im Bereich der Wörterbuchbenutzungsforschung in Zukunft angegangen werden sollten, um daraufhin wichtige Impulse sowohl der lexikographischen Praxis als auch der Fremdsprachendidaktik zu geben. Im nächsten Kapitel (2.) wird auf das Experiment näher eingegangen; danach (Kapitel 3) werden die Ergebnisse exemplarisch präsentiert, um schließlich (Kap. 4) die Forschungsmöglichkeiten aufzuzeigen.

2 Ziel und Design der empirischen Untersuchung zum Gebrauch des Smartphones

Die empirische Untersuchung, die aus einer Kombination von Umfrage und spezifischen Aufgaben mit retrospektiven Fragen[5] bestand, mit Studierenden des Deutschen als Fremdsprache an der Universität Roma Tre durchgeführt wurde, sollte (neben den oben genannten Zielen) auch Auskunft darüber geben, in welchen kommunikativen und Lernsituationen sie das Smartphone für lexikographische Fragestellungen nutzen, bei welchen sprachlichen Schwierigkeiten sie auf Online-Informationen zurückgreifen, welche Internetseiten oder/und Online-Wörterbücher sie verwenden und ob sie den dort gefundenen Informationen vertrauen. In Bezug auf zweisprachige deutsch-italienische bzw. mehrsprachige Internetwörterbücher sollte herausgefunden werden, welche sie verwenden[6] und warum? Nutzen Sie auch Übersetzungstools/-programme oder Suchmaschinen?[7] Werden auch einsprachige Internetwörterbücher des Deutschen wie Das digitale Wörterbuch

[5] *Retrospektive Fragen* (vgl. Faerch & Kasper 1987; Flick 2007) eignen sich gut, um herauszufinden, welche Überlegungen die Probanden bei einer bestimmten auszuführenden Handlung anstellen. Die Auswertung kann sowohl quantitativ als auch qualitativ erfolgen, da sie die klassische methodologische Dichotomie zwischen qualitativen und quantitativen Methoden überwindet, was vor allem im Bereich des Fremdspracherwerbs und der Fremdsprachendidaktik wünschenswert ist (vgl. Aguado 2009).
Für präzisere Ergebnisse zu den einzelnen Aufgaben wäre es besser, bei jeder einzelnen Aufgabe retrospektive Fragen zu stellen, was jedoch den Fragebogen beträchtlich verlängert und die Konzentration der Probanden strapaziert. Das *Think-Aloud-Protocol*, vgl. Flick 2007) wäre m.E. für weitere Forschungen die beste Möglichkeit, um genauere Daten zu erhalten.

[6] Bekannte deutsch-italienische Internetwörterbücher sind: *Leo* (dict.leo.org/itde/index_de), *Pons* (de.pons.eu/italienisch-deutsch/), *bab.la* (de.bab.la/woerterbuch/deutsch-italienisch/), *dict.cc* (deit.dict.cc/), *dicios* (www.it/dicios.com) oder ELDIT, ein lexikographisches Großprojekt der EURAC in Bozen (www.eurac.edu/ELDIT) [4.4.2014].

[7] bspw. *Google*-Übersetzer translate.google.de oder die Kombination eines mehrsprachigen Wörterbuchs und Suchmaschinen wie z.B. *linguee* (www.linguee.de) [4.4.2014]

der deutschen Sprache des 20. Jahrhunderts,[8] die deutsche Seite von free dictionary[9] oder das einsprachige Duden-Online-[10] oder das italienische einsprachige italienische *Treccani*-Wörterbuch[11] spezifische Synonym-Wörterbücher[12] oder spezielle DaF-Wörterbücher[13] konsultiert? Nutzen sie Wortschatzportale[14] und Wortschatz-Glossare, wie bspw. das Glossar vom Goethe-Institut[15] oder Internet-Seiten zur deutschen Sprache wie canoo.net, oder Online-Enzyklopädien, Abbildungen, Videos, bestimmte Online-Texte oder Korpora? All diese Möglichkeiten stehen den Lernenden frei zu Verfügung.

Hierzu wurden folgende konkrete Fragen gestellt, die von den Probanden nacheinander (ohne vor- oder zurückzublättern) beantwortet wurden.[16] Die Fragen beinhalteten Entscheidungsfragen (1., 6., 12., 16., 18., 22., 24., 26.), Multiple-Choice- (1., 11,) und offene Fragen (2., 3., 4., : 5., 6., 7., 8., 9., 10., 13., 14., 15., 17., 19., 20., 21., 22., 23., 25.):

(1) Benutzt du das Smartphone während des Fremdsprachenunterrichts, um bestimmte Informationen zu suchen? JA/ Nein. Wenn ja: häufig?/manchmal/ selten?

(2) Warum nutzt du es/ nutzt du es nicht?

(3) In welcher Situation suchst Informationen (z.B. während einer Grammatik-Übung, eines Lesetextes, Schreibaufgabe, um Vokabeln nachzuschlagen, um Vokabeln zu lernen,...)?

(4) Suchst du im Internet oder benutzt du eine bestimmte App? Wenn du eine App verwendest, welche? Wenn du im Internet suchst, auf welchen Seiten suchst du normalerweise?

(5) Benutzt du auch Online-Wörterbücher? Wenn ja, welche?

(6) Benutzt du Google-Translator oder ein anderes Übersetzungsprogramm? JA/ Nein.

(7) Welche Internetseite oder App ist deiner Meinung nach die beste, um nach der *Bedeutung eines deutschen Wortes* zu suchen? Warum?

(8) Welche Internetseite oder App ist deiner Meinung nach die beste, um *ein italienisches Wort* ins Deutsche zu *übersetzen*?[17] Warum?

(9) Traust du den Informationen, die du findest? Warum ja/nein?

(10) Gibt es auch Probleme/ Schwierigkeiten? Wenn ja, welche?

(11) Benutzt du das Smartphone auch in authentischen Kommunikationssituationen? Zum Beispiel
- wenn du mit deutschsprachigen Personen zusammen bist?
- wenn du ein Video in deutscher Sprache anschaust?

[8] www.dwds.de [4.4.2014]

[9] de.thefreedictionary.com/ [10.11.2013]

[10] www.duden.de (10.11.2013]

[11] www.treccani.it/vocabolario/dizionario/ [10.04.2014]

[12] www.duden.de/rechtschreibung/Synonymwoerterbuch oder synonyme.woxikon.de/ [10.11.2013]

[13] de.pons.eu/deutsch-als-fremdsprache/ [10.11.2013]

[14] wortschatz.uni-leipzig.de/[10.11.2013]

[15] www.goethe.de/z/jetzt/dejwort/dejwort.htm [10.11.2013]

[16] Die Fragen wurden in italienischer Sprache formuliert. Sie werden hier direkt in deutscher Sprache wiedergegeben. Es wurde bewusst die *Du*-Form gewählt, da im italienischen Kontext die Form des *Lei* (‚Sie') eine zu große psychologische Distanz schafft. Die Form des *voi* (‚ihr') schien mir zu unpersönlich und generell.

[17] Während es bei Frage 7 um die Konsultation in Bezug auf die sprachliche Rezeption geht, fokussiert die Frage 8 eine Handlung in Bezug auf die sprachliche Produktion.

- wenn du deutsches Radio hörst?
- um Vokabeln zu lernen?
- wenn du einen Text/ Buch/ online-Informationen auf Deutsch liest?
- Ich benutze ein Online-Wörterbuch auch, um nur darin zu „lesen"/"herumzublättern" ohne ein bestimmtes Ziel
- Anderes: ....

Während die ersten 11 Fragen allgemein gehalten wurden und auf der Erinnerung der Lernenden basierten, wurden für die nächsten Fragen konkrete Aufgaben formuliert, um herauszufinden, welche Schwierigkeiten die Probanden bei einer bestimmten Handlungssituation im Fremdsprachenunterricht haben und wie sie dabei mit dem Smartphone umgehen. Zuerst sollten die Probanden 5 Sätze mit polysemen italienischen Verben ins Deutsche übersetzen:[18]

A. Devo chiedere al mio capo (‚Ich muss meinen Chef *fragen.*'),
B. Mi ha chiesto di te (‚Er/Sie hat mich *nach* dir *gefragt*.'),
C. Mi ha chiesto un favore (‚Er/Sie hat mich *um* einen Gefallen *gebeten.*'),
D. Per Natale mia figlia mi ha chiesto un viaggio (‚Für Weihnachten hat sich meine Tochter von mir eine Reise *gewünscht*/hat mich um eine Reise *gebeten.*') und
E. Mi chiede sempre cose impossibili (‚Sie *verlangt* immer Unmögliches von mir'/'Sie fragte immer unmögliche Sachen'/‚Sie fragt (mich) immer nach unmöglichen Sachen').

Direkt anschließend wurden retrospektive Fragen gestellt:
(12) Hast du das Smartphone benutzt, um nach Informationen zu suchen? JA/ Nein.
(13) Wenn ja, für welche/n Satz/ Sätze?
(14) Warum/ Was hast du gesucht?
(15) Wo hast du gesucht (Internetseiten, Online-Wörterbücher,...)?
(16) Bist du auf Schwierigkeiten gestoßen bei den Informationen, die du gefunden hast bzw. bei dem Wörterbuch, das du benutzt hast?
(17) Wenn ja, welche?
(18) Danach sollten die Studierenden jeweils vier italienische bzw. deutsche Wörter und jeweils einen Satz in die andere Sprache übersetzen. Die Wörter waren: *F. merenda*, *G. bamboccioni*, *H. raccomandazioni*, *I. ISTAT* sowie *J. Wendehals*, K. *Quark*, *L. hartzen*, *M. hdl*. Zu den Wörtern gehörten Kulturspezifika, Neologismen und Abkürzungen, z.T. aus der gesprochenen Sprache, die bei einer Übersetzung in eine andere Sprache Schwierigkeiten bereiten, und zudem oft nicht in (Online-)

[18] Da die Studierenden mit der Übertragung polysemer italienischer Verben nachweislich Schwierigkeiten haben (vgl. Nied Curcio 2005), eignet sich diese Aufgabe besonders gut, um herauszufinden, ob das Wörterbuch, das die Studierenden verwenden, verschiedene Übersetzungsäquivalente angibt und ob metasprachliche Notationen zur Valenz und zum Kasus bei der Disambiguierung der Lesart des polysemen Verbs helfen können, bzw. ob die Studierenden diese Informationen – vorausgesetzt sie sind vorhanden – überhaupt berücksichtigen.

Wörterbüchern. Bspw. ist *hartzen* nicht in den bekanntesten deutsch-italienischen Wörterbüchern *Leo, Pons, canoo.net, dicios*,[19] vertreten, während es jedoch bspw. in der online-Enzyklädie *Wikipedia* erklärt wird und somit durch eine einfache Suche gefunden werden kann. In *duden.de* werden sogar Wortbedeutung (‚von Hartz IV leben'), Gebrauch (Jargon), Wortart, Worttrennung, Aussprache, Deklination und ein Beispiel mit der übertragenen Bedeutung im Jugendjargon (gestern Abend war ich nur am Hartzen (*konnte mich zu keiner Arbeit, Tätigkeit überwinden)*) angegeben. Bei den Sätzen ging es um umgangssprachliche Ausdrücke, die sehr häufig sind, aber in der Lexikographie oft vernachlässigt werden: N. *„Kommst du noch auf einen Absacker mit?"* und O. *„È stato stroncato da un infarto mentre lavorava.."* Die Schwierigkeit im Satz N. ist das Verständnis der Bedeutung von *Absacker*, im Satz O. geht es um den produktiven Transfer des Wortes *stroncare* und des gesamten Satzes mit einem Verb im Passiv und einem im Gerundio.[20] Die Online-Wörterbücher *pons, leo, bab.la*[21] und *Eldit* enthalten keinen Eintrag für *Absacker* und geben nicht das richtige Äquivalent für *stroncato* an (*Eldit* enthält keinen Eintrag für *stroncare*), *dicios* enthält in einer Liste die Äquivalente *abbrechen, dahinraffen, erliegen, zerstören, zunichtemachen* (das 2. und 3. wäre in diesem Kontext richtig), ohne jedoch weitere Zusatzinformationen zu geben, damit der Benutzer das richtige Äquivalent auswählen und den Satz korrekt produzieren könnte. *Google Übersetzer* gibt für den ersten Satz folgende Übersetzung an: **Sarà ancora con un drink*, was in die richtige Richtung geht, jedoch grammatisch nicht korrekt ist, während wenn man nur das Wort *Absacker* eingibt, das Wort *berretto da notte* (‚Schlafmütze') erscheint, was kontextuell selbstverständlich ebenfalls keinen Sinn ergibt.[22] Für den zweiten Satz gibt die Übersetzungsfunktion von *Google* im Deutschen **wurde von einem Herzinfarkt schlug während der Arbeit* an, sucht man nur nach dem Wort *stroncato*, so erscheint *verrissen*.[23]

Auch zu diesen Übersetzungsaufgaben wurde retrospektiv gefragt (wie 12-17) :

(19) Hast du das Smartphone benutzt, um nach Informationen zu suchen? JA/ Nein.

(20) Wenn ja, für welche/n Satz/ Sätze?

(21) Warum/ Was hast du gesucht?

(22) Wo hast du gesucht (Internetseiten, Online-Wörterbücher,...)?

(23) Bist du auf Schwierigkeiten gestoßen bei den Informationen, die du gefunden hast bzw. bei dem Wörterbuch, das du benutzt hast?

(24) Wenn ja, welche?

[19] it.dicios.com/ [10.04.2014]

[20] Das italienische *Gerundio* hat keine exakte Entsprechung im Deutschen und muss i.d.R. entweder durch eine Nominal- oder Verbalphrase im Deutschen wiedergegeben werden (s. Sattler 2008).

[21] www.bab.la [10.04.2014]

[22] Zu vermuten ist, dass es sich hier um eine inkorrekte, wörtliche – und nicht figurative – Übertragung des engl. „nightcap" (‚Schlummertrunk') geht.

[23] Aus offensichtlichen Platzgründen muss hier auf eine umfassende und detaillierte Präsentation der Ergebnisse verzichtet werden; die gewählten Beispiele sollen exemplarisch die Problematik aufzeigen.

Interessant in diesem Zusammenhang ist natürlich nicht nur die subjektive Reflexion der Probanden bezüglich ihrer eigenen Wörterbuchbenutzung, sondern auch das Ergebnis selbst. Wurden die Aufgaben korrekt ausgeführt bzw. welche Fehler wurden gemacht? War die Konsultation des Smartphones objektiv erfolgreich und wenn nicht, welche Benutzungsfehler gehen aus einem Vergleich der Ergebnisse und der Benutzerhandlung[24] hervor?

Zum Abschluss wurden von mir noch zwei allgemeine Fragen zur Lexikographie eingefügt:

(25) Möchtest du mehr über Wörterbücher erfahren? Ja/Nein.

(26) Wenn ja, was genau?

(27) Möchtest du mehr über Recherchemöglichkeiten im Internet (einsprachige deutsche Online-Wörterbücher, zweisprachige Online-Wörterbücher, Glossare, Korpora, Internetportale, spezifische Internetseiten,...) erfahren?

3 Die Ergebnisse der Untersuchung[25]

3.1 Die Benutzung des Smartphones im Fremdsprachenunterricht

An der Untersuchung nahmen 36 Germanistikstudierende[26] teil. 32 Probanden hatten ein Smartphone/I-phone mitgebracht, zwei ein Tablet und zwei Probanden waren nicht im Besitz eines Smartphones und haben das ihres Nachbarn benutzt. Die Auswertung erfolgte sowohl quantitativ als auch qualitativ; die Ergebnisse sollen im Folgenden mit ihrem prozentualen Anteil zusammenfassend beschrieben werden.[27]

Von den Probanden benutzen 80% ihr Smartphone generell während des Fremdsprachenunterrichts, um bestimmte Informationen hinsichtlich des Unterrichts zu suchen. 38% der Studierenden suchen nach Wörtern für die schriftliche Textrezeption (und 32% für die Textproduktion; 18% verwenden es, um Vokabeln zu lernen. Fast immer geht es dabei um Wörter: „per cercare il significato delle parole“

[24] Zur Klassifikation der Wörterbuchbenutzungshandlungen sowie der Benutzungsfehler vgl. Wiegand 1998.

[25] Die Ergebnisse können hier aus offensichtlichen Platzgründen nur stark gekürzt und ausschnitthaft präsentiert werden. Auch auf Graphiken, die für eine bessere Übersichtlichkeit dienen könnten, muss aus dem gleichen Grund verzichtet werden.

[26] Von den 36 Studierenden waren 34 Muttersprachler Italienisch, eine spanischsprechende Studierende und eine Studentin mit Rumänisch als Muttersprache. 28 der Studierenden (78%) lernten seit Beginn ihres Universitätsstudiums Deutsch, d.h. seit 2 Jahren und haben somit noch relativ geringe Deutschkenntnisse (Niveaustufe A2-B1 nach dem Gemeinsamen Europäischen Referenzrahmen); Die anderen 8 Studierenden (22%) hatten Deutsch als Fach in der Schule und lernen die Sprache seit 5-7 Jahren. Die Zeit war von mir nicht limitiert, das Design jedoch für eine Bearbeitung innerhalb 90 min. ausgelegt. Die Studierenden gaben nach 75-90 min. ab.

[27] Es ist selbstverständlich, dass eine Untersuchung mit 36 Studierenden nur eine erste Idee geben kann und in ihrer Quantität keinesfalls eine valide Aussagekraft erlangt. Trotzdem zeigt sie interessante Ergebnisse auf (vgl. 3) und weist m.E. auch auf Potential für weitere umfangreichere Forschungsmöglichkeiten hin (vgl. 4.).

('um die Bedeutung von Wörtern nachzuschlage'), „durante una produzione scritta, per cercare dei vocaboli" ('während der Textproduktion, um die Wörter nachzuschlagen') oder „per studiare dei vocaboli" ('beim/zum Vokabellernen'). Nur zwei Probanden nannten auch „espressioni" ('Ausdrücke') und „modi di dire" ('Redewendungen'), die sie nachschlagen.

Fast alle Studierenden gebrauchen sowohl das Internet, als auch verschiedene Apps. Sie geben den Apps den Vorrang, wenn sie die Bedeutung eines konkreten Wortes suchen und es schnell gehen soll. An 1. Stelle bei den Apps steht das *Pons*-Wörterbuch und an 2. Stelle *Google Übersetzer*, gemeinsam mit dem Portal *WordReference.*[28] Bei der Internet-Recherche geben Sie *Google Übersetzer* den Vorrang und an 2. Position stehen die *Google*-Suchfunktion und *Wikipedia*. Seltener genannt werden *Pons*, *Youtube* und *WordReference*. Es wird deutlich, dass das wichtigste Kriterium die schnelle und direkte Suche ist.

Bei der Frage 5, in der konkret nach der Verwendung von Online-Wörterbüchern gefragt wird, geben 34 Studierende (94%) an, dass sie welche benutzen: Die drei am häufigsten genannten sind: *WordReference* (12), *pons* (11) und *Eldit* (11). Außerdem wurde 4 Mal *bab.la* sowie je 3 Mal *Leo* und *Larousse*[29] angeführt, die auch eine zweisprachige Wörterbuchversion Italienisch↔Deutsch beinhalten. Von drei Probanden wurde auch *Google Übersetzer* als Wörterbuch aufgezählt (sic!). Nur vereinzelt wurden einsprachige Wörterbücher (*Duden*, *Treccani* für das Italienische) oder Portale wie *Collins*[30], *Reverso*[31], *Urban dictionary*[32], *dictionary reference*[33], und das italienisch-deutsche Wörterbuch des *Corriere della Sera* erwähnt. Die spezifische Frage, ob sie *Google Übersetzer* verwenden, beantworteten 25 (69%) der Germanistikstudierenden mit Ja.

Die Antworten auf die Frage 7. *Welche Internetseite oder App ist deiner Meinung nach die beste, um nach der <u>Bedeutung eines deutschen Wortes</u> zu suchen? Warum?* zeigen folgende Ergebnisse.

Zehn der Studierende (28%) geben *Eldit* den Vorrang, gefolgt von *Pons* (19%). An dritter Stelle der Präferenzen stehen *Leo* und *Larousse*. Zweimal wurde auch *Google Übersetzer* angeführt. Interessant ist, dass die Studierenden zwar verschiedene Internetseiten/Apps aufgezählt haben, aber trotzdem die gleichen Gründe für ihre Wahl nennen, und zwar:

(1) Zufriedenstellende Beispielsätze (30%)
(2) Das Wort wird im Kontext angeführt (19%)
(3) Erwähnung von typischen Ausdrücken, Idiomen (11%)
(4) Grammatische Informationen, z.B. Konjugation, Valenz (8%)
(5) Leichter und schneller Zugriff (8%)
(6) Präzise und komplett (5%)
(7) Liste von Übersetzungsäquivalenten (5%)

[28] www.wordreference.com/ [10.04.2014]
[29] www.larousse.fr/dictionnaires/allemand-italien [10.04.2014]
[30] www.collinsdictionary.com/dictionary/italian-english/dizionario [10.04.2014]
[31] dizionario.reverso.net/ [10.04.2014]
[32] www.urbandictionary.com/ [10.04.2014]
[33] dictionary.reference.com/ [10.04.2014]

Acht Probanden (22%) beantworteten die Frage nicht oder gaben an, dass sie keine Antwort wussten. Dieser Anteil liegt bei Frage 8., in der nach der besten Internetseite/ App *zum Übersetzen* gefragt wurde, sogar bei 12 (33%). Auch die Ergebnisse im Vergleich zu 7. sind unterschiedlich. Als die beste Seite/App wird *Pons* zitiert (22%), da der Zugriff leicht sei, die Beispiele zufriedenstellend seien, das Wort im Gebrauchskontext stehe, es viele Kollokationen und (idiomatische) Ausdrücke gebe und grammatische Zusatzinformationen (auch die syntaktische Struktur) vorhanden seien. Interessant ist, dass an 2. Stelle das Print-Wörterbuch liegt (14%) – obwohl sich die Frage eigentlich nur auf Online-Wörterbücher bezog. An 3. Stelle liegt *Leo* (11%). Zwei Studierende waren der Meinung, dass kein Wörterbuch besser als das andere sei, sondern dass die „ricerche incrociate" bzw. cross research oder „überkreuzte" Recherchen, bei denen man hin- und herspringt, vergleicht und abwägt, was das Beste wäre. Eine weitere interessante Aussage war die einer Studentin, die meinte, dass für sie *WordReference* die beste Internetseite sei, da sie bei der Übersetzung vom Italienischen ins Deutsche (was ja ihre Fremdsprache ist) via Englisch übersetze, und nicht direkt Italienisch>Deutsch.

Bei Frage 9., ob die Studierenden den Informationen, die sie finden, trauen, gab es folgende Ergebnisse:

Ja (absolut): 19%
Ja, aber ich suche noch an anderer Stelle: 22%
Ja, meistens: 15%
Ja, ziemlich: 8%
Nicht immer: 14%
Nein, ich suche noch an anderer Stelle: 22%

Der Prozentsatz der Studierenden, die sich auf die Informationen im Internet mehr oder weniger verlassen – wenn auch mit Einschränkung – beträgt doch 64%. Gründe dafür waren, dass ihnen die Seite von der Lehrperson empfohlen wurde, dass sie die Seite kennen oder dass sie gute Erfahrungen damit gemacht hatten. 44% der Probanden (mehr oder weniger von den Online-Informationen überzeugt) suchen noch an anderer Stelle und nannten als Grund ihre Zweifel aufgrund von negativen Erfahrungen. Drei Probanden, die sich auf die Informationen im Internet nicht verlassen, trauen nur dem Print-Wörterbuch.

67% der Probanden gaben – in Bezug auf die Schwierigkeiten, die sie mit Online-Informationen haben – an, dass sie sich unsicher fühlen und Zweifel haben, da sie nie wissen, *welches* Äquivalent in einem bestimmten Kontext das richtige ist[34]. Diese Unsicherheit zeige sich auch, wenn sie verschiedene Internetseiten konsultierten und die gefundenen Informationen nicht übereinstimmten. Das Problem ist demnach die Auswahl des richtigen Übersetzungsäquivalents, v.a. gerade dann, wenn zu wenige Übersetzungsäquivalente/ Bedeutungsvarianten angegeben würden und die Informationen zu vage seien. Ihrer Meinung nach mangele es an Kriterien, die ihnen bei der Orientierung helfen, sowohl auf der Mikro- als auch auf der Makrostruktur. Mangel an Informationen wurde von ihnen auch

[34] Dies gilt sicherlich insbesondere für die sprachliche Produktion.

in Bezug auf die metasprachliche Notation angegeben. Zudem fehlten grammatische Informationen wie z.B. Präpositionen, die von Nomen und Verben regiert werden („mancano delle informazioni grammaticali (es. le preposizioni che reggono certi nomi/ verbi)"). Weitere Mängel wurden auch in Bezug auf Phraseologismen, Wortkompositionen[35] und Fachbegriffe genannt.

Auch in authentischen Situationen, in denen sie mit Deutsch in Kontakt sind, wird das Smartphone von ihnen verwendet: 75% der Probanden nutzen es v.a. beim Lesen eines Textes/Buches oder von Online-Informationen auf Deutsch (d.h. Rezeption), 69% der Studierenden verwenden es während sie ein Video/einen Film in deutscher Sprache anschauen (d.h. Rezeption) und 64% lernen damit Vokabeln. Nur selten wird darin ohne Ziel „gelesen"/"herumgeblättert" (22%). In Anwesenheit von Muttersprachlern wenden sich die Studierenden i.d.R. an diese (11%), da sie als Experten betrachtet werden und als Autorität über dem Wörterbuch stehen. Dass sie von dieser Möglichkeit Gebrauch machen, gab eine Probandin auch als Antwort bei Frage 9 an, in der es um vertrauensvolle Informationen ging: „Non sempre mi fido quindi provo a consultare più fonti o chiedo a persone più competenti (amici madrelingua)." ('nicht immer trau ich [den Informationen], deshalb versuche ich mehrere Quellen heranzuziehen oder frage Leute, die kompetenter sind (Muttersprachler)'). Am wenigsten benutzen sie es beim Radiohören (5%).[36]

3.2 Wie Wörterbuchnutzung bei spezifischen Übersetzungsaufgaben

In diesem Kapitel werden die Ergebnisse vorgestellt, die sich aus der Analyse der Übersetzung der fünf Sätze A.-E. mit dem polysemen italienischen Verb *chiedere* (s. 2.), der Wörter (F.-M.) und der beiden Sätze N. und O. Sie werden zudem in Bezug auf ihre Korrektheit und die Konsultation des Wörterbuchs von Seiten der Germanistikstudierenden, bzw. die Reflexion der Studierenden gebracht.

3.2.1 Die Übersetzungen des italienischen Verbs chiedere und die Benutzung des Smartphones

Auch wenn man berücksichtigt, dass die Studierenden noch kein hohes Sprachniveau haben, machen die Ergebnisse sehr nachdenklich, denn der höchste Prozentsatz an Korrektheit bei einem Satz lag nur bei 31% für Satz A, was auf große Schwierigkeiten von Seiten der Studierenden hinweist.

31% A. *Devo chiedere al mio capo* (‚Ich muss meinen Chef *fragen.*')

[35] Dieser Mangel kann nicht unbedingt den Wörterbüchern zugeschrieben werden; die Komposition, auch ad-hoc, ist charakteristisch für die deutsche Sprache, und nicht jede Komposition kann in einem Wörterbuch aufgeführt werden, auch wenn Online-Wörterbücher weniger als Print-Wörterbücher das Platzproblem haben und „auf dem Laufenden" sein könnten. Hier würde ein zusätzliche Suche mit Suchmaschinen weiter helfen, um die Bedeutung des Wortes zumindest im Kontext auflösen zu können. Aber auch eine sprachliche Dekomposition mit einer erneuten Suche im Wörterbuch kann zu einem positiven Ergebnis führen.

[36] Diese Frage war zu sehr auf das Radio limitiert, und müsste bei einer zukünftigen Untersuchung generell auf Hörtexte erweitert werden, so dass auch das Musikhören einbezogen wird. Eine Studentin gab von sich aus an, dass sie beim Musikhören das Smartphone zur Konsultation von unbekannten Wörtern verwende. (Bei einer spezifischeren Fragestellung wäre diese Möglichkeit sicher öfters in Betracht gezogen worden.)

22% B. *Mi ha chiesto di te* (‚Er/Sie hat mich *nach* dir *gefragt*.‘)

28% C. *Mi ha chiesto un favore* (‚Er/Sie hat mich *um* einen Gefallen *gebeten*.‘)

11% D. *Per Natale mia figlia mi ha chiesto un viaggio* (‚Für Weihnachten hat sich meine Tochter von mir eine Reise *gewünscht*/hat mich um eine Reise *gebeten*.)

20% E. *Mi chiede sempre cose impossibili* (‚Sie *verlangt* immer Unmögliches von mir‘/‘Sie fragte immer unmögliche Sachen‘/‚Sie fragt (mich) immer nach unmöglichen Sachen‘).

Für die Entscheidung, ob der Satz als richtig interpretiert wurde, war die Tatsache, ob

(1) Die Probanden das richtige Übersetzungsäquivalent gefunden haben,

(2) ob sie auf die grammatischen Zusatzinformationen geachtet hatten (falls diese vorhanden waren). Es ging hier insbesondere um die Valenz und den Kasus.[37]

Die Schwierigkeit der Übersetzung lag bei Satz A. insbesondere in der kontrastiven Valenz: Im Italienischen braucht das Verb *chiedere* ein Subjekt und ein indirektes Objekt, während das deutsche Verb *fragen* ein Subjekt und eine Akkusativergänzung realisieren muss. Die Fehler lagen zu 80% bei der Valenz, z.B. **Ich muss meinem Chef fragen*, aber auch in der Wahl des falschen Verbs (20%), wie **Ich muss meinen Chef anfragen* oder **Ich muss meinen Chef nachfragen*. Satz A war der Satz, zu dem das Smartphone am wenigsten verwendet wurde. Vermutlich waren sich einige Probanden der kontrastiven Valenz nicht bewusst, haben sich nur auf die Bedeutungsentsprechung konzentriert, und nicht nachgeprüft. Sie haben parallel zur italienischen Valenz den deutschen Satz konstruiert. Die Konsultation des Smartphones lag von allen recherchierten Sätzen bei A. am niedrigsten: nur 44% der Studierenden haben das Smartphone für die Übersetzung dieses Satzes verwendet.

Bei Satz B. war es wiederum die Valenz. Nicht nur die italienische Präposition *di*, sondern auch das direkte Objekt im Italienischen wurde oft nicht adäquat im Deutschen wiedergegeben. Sätze wie **Sie fragt an mich über dich./ *Sie hat mich über dich gefragt./ *Er hat von dir mir gefragt./ *Er fragte mich nach dich.* waren häufig. Der Satz * **Er fragte mich über dich.* wurde von denjenigen geschrieben, die *google Übersetzer* verwendet hatten und – da sie auf Fehler gestoßen waren – in Anlehnung an den dort aufzufinden Satz **Er fragt mich über Sie*. umformuliert haben: „Ho cercato l'intera frase da Google Translator ma nella traduzione ho rilevato alcuni errori.“ (‚Ich habe den ganzen Satz in Google Translator gesucht aber in der Übersetzung habe ich einige Fehler bemerkt.). 77% der Probanden haben sich des Smartphones hier bedient und es ist zu bemerken, dass auch diejenigen, die in einem Online-Wörterbuch wie z.B. *Leo*, nachgeschaut haben, Fehler gemacht haben, was zeigt, dass sie sich bei den verschiedenen Übersetzungsäquivalenten nicht orientieren konnten, denn *Leo* gibt für *chiedere di qu./qc.* vier Äquivalente an (*sich*^{Akk} *nach jmdm/etw. erkundigen*; *nach jmdm./ etw. fragen*; *nach jmd. Fragen* und *nach jmd. verlangen*). Doch auch eine Recherche in *Pons*, wo *chiedere notizie di qu – sich nach jmdm erkundigen* steht, sind die

[37] Dabei muss erwähnt werden, dass morphologische Fehler wie bspw. eine falsche Konjugation (**Er bitt mich um unmögliche Dinge*) sowie Fehler in der Satzstellung (**Er bat um einen Gefallen mich*) nicht berücksichtig wurden, da der Schwerpunkt der Untersuchung darauf lag, ob die Probanden das richtige Übersetzungsäquivalent finden und auf die grammatischen Zusatzinformationen achten.

Sätze nicht korrekt, da die Probanden nicht auf *jmdm* geachtet haben und dadurch keine Dativergänzung realisiert haben: **Er hat sich nach dich erkundigt*.

Obwohl der Satz C. ein Funktionsverb bzw. Phraseologismus in beiden Sprachen enthielt (*chiedere un favore – um einen Gefallen bitten*), war die Fehlerquote etwas geringer als bei B, lag aber immer noch hoch. 83% der Probanden haben diesen Satz im Smartphone recherchiert. Wenn sie nach dem gesamten Ausdruck in einem Online-Wörterbuch (oder den gesamten Satz in *Google* Übersetzer gesucht hatten), dann kamen sie zum richtigen Satz. Einige suchten jedoch nur nach dem Wort *favore* und schrieben deshalb **Sie/Er hat mir einen Gefallen gefragt*.

Satz D war der Satz, bei dem die Fehlerquote am höchsten war (72%). 64% der Probanden haben dafür das Smartphone verwendet, d.h. weniger als bei Satz C. Die Einträge sowohl in den gängigen Online-Wörterbüchern als auch in Übersetzungsprogrammen weisen die Studierenden nicht darauf hin, dass es sich hier um den Kontext eines *Geschenks* geht. Der Satz wurde nur von denjenigen richtig übersetzt, die entweder schon mehrere Jahre Deutsch lernen oder die einen Ausdruck für *chiedere un regalo* (‚um ein Geschenk bitten') gesucht hatten.

Der fünfte Satz (E.) wurde ebenfalls von 80% der Studierenden nicht korrekt übersetzt. 66% konsultierten das Smartphone hierzu. Auch wenn man gestehen muss, dass die Übersetzung des Satzes ins Deutsche für Lernende, die in der Mehrheit erst seit 2 Jahren Deutsch lernen, sehr schwierig war, muss doch unterstrichen werden, dass es – auch wenn sie *fragen* als Verb verwendeten, meist die Valenzangaben (wie bei A.) nicht richtig realisiert wurden.

Generell kann gesagt werden, dass 35 von 36 Probanden (!) das Smartphone für fast alle Sätze verwendet haben! – und trotzdem lag die Fehlerquote zwischen 69-89%! Bei der Suche nach Informationen haben sie – wie auch unter 3.1. beschrieben – nur die bekanntesten Online-Wörterbücher zu Rate gezogen oder mit Übersetzungsprogrammen gearbeitet; es wurden keine weiteren Hilfsmittel und Recherchemöglichkeiten – wie unter 2. beschrieben – in Betracht gezogen. Hier liegt sicher schon der erste Benutzerfehler[38], nämlich der Wörterbuchwahlfehler. Wie man jedoch an der Relation zwischen Häufigkeit der Benutzung des Smartphones und Fehlerquote ablesen, begehen die Studierenden auch eine Menge Handlungsfehler, v.a. sind es die Nichtbeachtung von grammatischen Zusatzinformationen (Verbergänzung (z.B. Akkusativ- oder Dativergänzung), regiertem Kasus einer Präpositivergänzung (z.B. einen Gefallen bitten + um+*Akk*), aber auch die Suche nur nach einzelnen *Wörtern*, und nicht nach dem gesamten Phraseologismus. Diese Fehler führten i.d.R. immer zu einem nicht korrekten Satz. Die Umkehrung bedeutet jedoch nicht immer, dass bei Beachtung dieser Informationen der Satz korrekt übersetzt wird, denn das Online-Wörterbuch selbst bietet, wie wir gesehen haben, nicht immer eine wirkliche Unterstützung, denn manchmal gibt es nur bedingt grammatische Zusatzinformationen oder/und es gibt mehrere Übersetzungsäquivalente, so dass die Benutzer keine Orientierungshilfe für die Auswahl des richtigen Äquivalents finden.

[38] Zu Wörterbuchbenutzerfehler s. Wiegand 1998: 519.

3.2.2 Die Übersetzung von Neologismen und die Benutzung des Smartphones

Alle Probanden (100%) gaben an, dass sie ihr Smartphone bei den Übersetzungsaufgaben F.-O. benutzt hatten, um nach bestimmten Informationen in Bezug zu suchen. 33 davon (92%) haben zugegeben, dass sie bei der Aufgabenbewältigung gleichzeitig Schwierigkeiten mit den Internetseiten oder Online-Wörterbüchern hatten, die sie für diese Aufgabenstellung konsultiert hatten. Diese Schwierigkeiten, Zweifel und Unsicherheiten werden von ihnen oft auch geäußert (s. Frage 23): Die häufigste Antwort war: „non ho trovato alcune parole" (‚einige Wörter habe ich nicht gefunden'). Weitere Kommentare waren bspw. „risultava strana la traduzione" (‚die Übersetzung war seltsam'), „più traduzioni discordanti" (‚mehrere gegensätzliche Übersetzungen') und „sempre quella di riuscire a capire/trovare la giusta traduzione" (‚immer das gleiche [Problem], nämlich zu verstehen, welche die richtige Übersetzung ist').

Bei der Beschreibung der Ergebnisse in Bezug auf die Übersetzungsäquivalente der italienischen und deutschen Wörter in F. – M. möchte ich mich auf die deutschen Neologismen *hartzen*, *Absacker* und die Abkürzung *hdl*, sowie den umgangssprachlichen italienischen Ausdruck *stroncare* (‚sterben') in Zusammenhang mit *Herzinfarkt* im Satz O. konzentrieren. Der Grund für die Auswahl der deutschen Wörter liegt darin, dass für *hartzen* und *hdl* in den Online-Wörterbüchern *Pons* und *Leo* kein Eintrag zu finden ist (für *Absacker* gibt es zwar die Entsprechung *bicchiere della staffa*, doch meinten die Studierenden, dass sie nicht wüssten, worum es ginge, was darauf hinweist, dass es ein weniger gebräuchlicher Ausdruck als im Deutschen ist). Hätten die Probanden jedoch in der *Pons*-Textübersetzung geschaut, so hätten sie den korrekt übersetzten Satz *unitevi a me per un bicchierino?* gefunden, was der deutschen Bedeutung ziemlich nahe kommt. Um das richtige Äquivalent zu finden, waren die Studierenden darauf angewiesen, nicht nur das Online-Wörterbuch zu verwenden, sondern weitere (auch sehr simple) Recherchen zu tätigen. Zu einer erfolgreichen Recherche führten hier die Konsultation der Suchmaschine von *Google* im Allgemeinen, da man bei *hartzen* auf die *Wikipedia*-Seite geführt wird und dort eine Bedeutungserklärung findet. Bei der Google-Suche nach *Absacker* erscheint an oberster Stelle die Webseite *duden.de* mit der Erklärung „am Ende eines Zusammenseins oder vor dem Schlafengehen getrunkenes letztes Glas eines alkoholischen Getränks", was für einen Deutschlernenden auch mit keinem hohen Deutschniveau m.E. verständlich ist. Für eine erfolgreiche Recherche zu *hdl* muss man zuerst verstehen, dass es sich um eine Abkürzung handelt und dann in *Google* z.B. „Abkürzung hdl" eingeben, um *hab dich lieb* zu finden und dann das entsprechende *tvb* zu realisieren.[39] Auch wenn die weiteren Recherchen von den Probanden keine komplexen Recherchekompetenzen abverlangt haben, so war es für eine erfolgreiche Übersetzung doch wichtig, verschiedene Informationsquellen zu konsultieren, evtl. zwischen Wörterbüchern und Internetseiten zu vergleichen. Sie konnten sich nicht nur auf zweisprachige Wörterbücher verlassen, wie sie das gerne tun, sondern mussten auch

[39] Bei einer einfachen Recherche „hdl" in der Suchmaschine gelangt man zu „high density lipoprotein". Die Abkürzung sollte bei einer weiteren Untersuchung in einen Kontext eingebaut werden, auch wenn von den Studierenden bemerkt wurde, dass es ihnen klar war, dass es sich nicht um das Protein handeln könne.

mit Informationen im einsprachigen umgehen können und gezielt – im Fall von *hdl* – nach einer Abkürzung suchen. Der Prozentsatz an richtigen Lösungen war extrem niedrig: nur 4 Studierende (11%) haben die Bedeutung von *hartzen* gefunden. Drei der Probanden schrieben „essere disoccupato" (‚arbeitslos' sein) und eine Studierende, die jedoch schon seit 7 Jahren Deutsch lernt, gab an: „essere disoccupato (arbeitslos sein) parola gergale, introdotto nel 2009, neologismo" (‚arbeitslos sein (...) umgangssprachlicher Ausdruck/Jargon, 2009 eingeführt, Neologismus). Die Abkürzung *hdl* wurde von sieben Probanden (19%) als *tvb/ti voglio bene/ti amo* wiedergegeben. Die „erfolgreichen" Studierenden haben meist gegoogelt und eine „ricerca incrociata" (‚cross research', überkreuzte Recherche') durchgeführt, d.h. auf verschiedenen Webseiten und Online-Wörterbüchern nachgeschaut, verglichen und sich für eine der Übersetzungsmöglichkeiten entschieden. Der deutsche Satz *Kommst du mit, einen Absacker trinken?* wurde von 19% der Probanden korrekt übertragen. Der italienische Satz, bei dem die Hauptschwierigkeit im Verb *stroncare* (hier ugs.: ‚sterben') lag, hatte eine Erfolgsquote von 16%. Als Exempel für eine erfolgreiche Übersetzung sollen folgende zwei Kommentare als Zitate dienen: „mi sono aiutata da google che mi ha dato indicazioni su neologismi" ('ich habe Google zu Hilfe genommen, wo ich Hinweise auf Neologismen gefunden habe') und „Il primo dizionario utilizzato non era in grado di tradurre alcune parole che, attraverso una ricerca sul web, sono riuscita a trovare tramite i collegamenti alle varie pagine" (‚Das erste Wörterbuch, das ich benutzt habe, konnte einige Wörter nicht übersetzen; über eine Webrecherche und den verschiedenen Verbindungen (Vergleichen) ist es mir dann gelunge." Eine weitere Strategie, die i.d.R. zu Erfolg führte, war die Suche mit Hilfe des Englischen als „Sandwichsprache",[40] d.h. die Lernenden verwenden das Englische als Zwischensprache, um zur anderen Spache zu gelangen, sowohl Deutsch>Englisch>Italienisch als auch viceversa: „Ho avuto difficoltà con „hartzen" la cui traduzione era disponibile solo in inglese sotto forma di slang. Per „Absacker" ho dovuto cercare da tedesco a inglese e poi da inglese a italiano." Interessant wäre eine noch präzisere Angabe zu *Google*, d.h. wo und was sie gegoogelt haben, ob sie auch z.B. Bilder verwendet haben – eine Strategie, die gerade für das Verstehen von Kulturspezifika sehr sinnvoll ist. Nur ein Student, der den Satz ?**Vieni ancora a bere un Absacker* geschrieben hat, hatte in *Google*-Bilder recherchiert und eine Flasche gefunden, auf deren Etikett der Name Absacker stand, so dass er dachte, das Getränke hieße so: „Sono andato su google e dalle immagini ho visto cosa era." (‚Ich bin auf Google gegangen und auf den Bildern habe ich gesehen was es war.') Eine eindeutige Relation besteht bei den Sätzen N. und O. zwischen dem Gebrauch von *Googler*-Übersetzer und nicht korrekten Übersetzungen, evident bei Sätzen wie *W*urde von einem Herzinfarkt schlug während der Arbeit.* Hatte der Proband jedoch Zweifel an der Übersetzung und wendete linguistische Strategie, wie die des Vereinfachens und Umformulierens an, so war die Chance auf Erfolg groß, wie bei den folgenden Fällen: „Absacker, però la traduzione non mi soddisfaceva perciò l'ho interpretata a modo mio, cercando altre

[40] Ich definiere diesen Begriff – in Anlehnung an die *Sandwich-Technik* als didaktische Methode (vgl. Butzkamm 2004) – als die Sprache, die bei der lexikographischen Recherche zwischen einem Sprachenpaar, bzw. zwischen einer Ausgangs- und einer Zielsprache eingeschoben wird, mittels deren Bedeutung das entsprechende Übersetzungsäquivalent gefunden wird.

combinazioni di parole per una traduzione più esatta.“ (‘A., aber die Übersetzung hat mich nicht überzeugt, also habe ich es auf meine Art interpretiert und andere Kombinationen von Wörtern für eine exaktere Übersetzung gesucht’) und „stroncato: è stato difficile trovare un corrispettivo tedesco e ho avuto bisogno di semplificare la frase in ambito di traduzione.“ (‚stroncato: es war schwierig ein deutsches Äquivalent zu finden und ich musste den Satz bei der Übersetzung vereinfachen‘). Es zeigt sich, dass positive Resultate auch bei relativ niedrigem Sprachniveau auch dann erreicht werden können, wenn die Recherchekompetenz gut ausgebildet ist, ein gesundes Misstrauen der dargebotenen Übersetzungsmöglichkeit gegenüber besteht, das Sprachbewusstsein hoch ist und so zum Einsatz von linguistischen Strategien führen kann.

4 Zusammenfassung und Forschungsausblick

Wenn man bedenkt, dass die Studierenden mit ihrem Smartphone unbegrenzte Recherchemöglichkeiten hatten und dass sie bequem und schnell an Informationen gelangen konnten, ist es doch verwunderlich, dass

(1) sie nur einige wenige Internetseiten (meist Übersetzungsprogramme) und Online-Wörterbücher für die konkrete Übersetzung verwendet haben. Es scheint, als würden sie nur wenige kennen. Interessant ist, dass sie zwar mehr Online-Seiten bzw. -wörterbücher aufgezählt haben (Fragen 1.-11. der Untersuchung), als sie an späterer Stelle konkret für die Aufgaben benutzt haben (was aber auch daran liegen kann, dass sie innerhalb des Seminars von 90 min. fertig sein wollten, auch wenn kein Zeit-Limit angegeben war).

(2) ihr Benutzungsverhalten dem des Print-Wörterbuchs ähnelt, bspw. dass nur wenige Studierende einsprachige Online-Wörterbucher wie *Duden* oder *Treccani* verwenden, dass sie zu schnellen Ergebnissen kommen wollen und das erstbeste Tool nutzen, das sie finden können (bei der *Google*-Recherche nehmen sie das an erster Stelle stehende Wörterbuch, im Wörterbuchartikel nehmen sie das 1. Äquivalent) und machen sich nicht die Mühe, weiter zu suchen – nach dem Motto: das Erste ist gut genug. Außerdem sind sie auf die Suche nach Wörtern fixiert; sie recherchieren nicht nach der Bedeutung eines Wortkomplexes, d.h. sie suchen keine Phraseologismen. Metasprachliche Notationen, wie bspw. Kasus und Valenz, werden „übersehen“ bzw. nicht berücksichtigt.

(3) viele der Studierenden (noch) kein Vertrauen in die Online-Recherche haben und sich noch recht orientierungslos in Bezug auf lexikalische Lernschwierigkeiten im Internet bewegen, andererseits aber auch Studierende gibt, die Übersetzungsprogramme verwenden, ohne sich darüber Gedanken zu machen.

All diese Ergebnisse zeigen wieder einmal – wie schon bei Forschungen zu Print-Wörterbüchern –, dass generell die Lexikographie und im Besonderen die Wörterbuchbenutzung in die Fremdsprachendidaktik integriert werden müssen. Wie auch die Antworten der Studierenden auf die Frage 26.

zeigen, wünschen sich 35 Studierende (97%), mehr über Recherchemöglichkeiten im Internet zu erfahren, welche Online-Wörterbücher es gibt, wie verschiedene Online-Wörterbücher konzipiert wurden, wie sie strukturiert sind, wie man sich orientiert und wie man zuverlässige Informationen erkennt, und v.a. wie man sie erfolgreich benutzt: „come si usa senza sbagliare" (‚wie man sie gebraucht ohne Fehler zu machen').

In Bezug auf zukünftige Forschungsmöglichkeiten geben m.E. insbesondere die „ricerce incrociate" und der Verwendung des Englischen als „Sandwichsprache" Anlass, neue Ideen für weitere umfassendere Untersuchungen zu entwickeln.

(1) Hinsichtlich der überkreuzten Recherche wäre es bspw. interessant, in einer umfassenden Studie herauszufinden, welche Online-Informationen die Smartphone-Benutzer zu Rate ziehen, wie sie konkret von einer Internseite/einem Wörterbuch zum anderen wechseln, wie sie vergleichen, welche Schwierigkeiten und Zweifel sie dabei haben und welche Informationen die Wahl eines bestimmten Übersetzungsäquivalentes beeinflussen. Als Methode wäre m.E. das *Think-Aloud-Protocol* (Lautes Denken-Protokoll) die beste Möglichkeit, mit der die Gedankengänge des Smartphone-Benutzers während dieser Recherche am besten zum Vorschein kämen und für detaillierte Analysen auch aufgezeichnet werden könnten.

(2) Meines Wissens noch nicht erforscht ist die Tatsache, dass sich Studierende für ein bestimmtes Sprachenpaar des Englischen als „Sandwichsprache" bedienen. Man könnte z.B. Studierende, die gewohnheitsmäßig mit *WordReference* arbeiten, bei ihrer Smartphone-Benutzung beobachten und ihr Lautes Denken gleichzeitig aufzeichnen. Man würde von ihren sprachlichen Schwierigkeiten Genaueres erfahren und wie sie mit ihrem Smartphone diesbezüglich umgehen, um Lösungen für ihre Schwierigkeiten zu suchen, d.h. wo suchen sie, was suchen sie, wie suchen sie, wie wägen sie ab, wie gelingt ihnen der Transfer zwischen den Sprachen, usw. Ich kann mir gut vorstellen, dass sich hier neue Erkenntnisse ergeben, die für die Konzeption und Ausarbeitung zukünftiger Online-Wörterbücher und Apps von großer Bedeutung sein werden. Außerdem bin ich überzeugt – nicht zuletzt wegen der zunehmenden Mehrsprachigkeit –, dass Englisch im Bereich der Wörterbuchbenutzung zwischen zwei Sprachenpaaren immer häufiger als „Sandwichsprache" fungieren und bewusst von den Studierenden als Strategie eingesetzt wird. Diese Ergebnisse aus der Wörterbuchbenutzungsforschung in der Fremdsprachendidaktik sollten in die zukünftige Schreibung von Online-Wörterbüchern unbedingt einfließen und gerade im Bereich der zweisprachigen Lexikographie überdacht werden.

5 Literatur

Aguado, K. (2009). Möglichkeiten und Grenzen mehrmethodischer empirischer Fremdsprachenlehr- und -lernforschung. In B. Baumann, S. Hoffmann, M. Nied Curcio (Hgg.). *Qualitative Forschung in Deutsch als Fremdsprache*. Frankfurt: Lang, S. 13-22.

Albrecht, Jörn (2005). *Übersetzung und Linguistik*. Tübingen: Narr.

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

Bimmel, P., Van de Ven, M. (2000). Man nehme ein Wörterbuch... *Fremdsprache Deutsch. Übersetzen im Deutschunterricht*, 23, S. 38-39.

Boonmoh, A. (2012). E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos. Journal of the African Association for Lexicography*, S. 43–68.

Butzkamm, W. (2004). Lust zum Lehren, Lust zum Lernen. Eine neue Methodik für den Fremdsprachenunterricht. Tübingen: Francke.

Corda, A., Marello, C. (2004). *Lessico. Insegnarlo e impararlo*. Perugia: Guerra.

de Schryver, G.-M., Prinsloo, D. J. (2011). Do Dictionaries Define on the Level of their Target Users? A Case Study for Three Dutch Dictionaries, *International Journey of Lexicography*, 21(1), S. 5–28.

Domínguez Vázquez, M. J., Mirazo Balsa, M. & Vidal Pérez, V. (2013). Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In M. J. Domínguez Vázquez (Hg.). *Trends in der deutsch-spanischen Lexikographie*, 135–172. Frankfurt: Lang.

Domínguez Vázquez, M.-J., Mollica F. & Nied Curcio, M. (2014, im Druck). Simplex-Verben im Italienischen und Spanischen vs. Präfix- und Partikelverben im Deutschen. Eine Untersuchung zum Gebrauch von Online-Wörterbüchern bei der Übersetzung. In M.-J. Domínguez Vázquez, F. Mollica & M. Nied Curcio. *Zweisprachige Lexikographie zwischen Translation und Didaktik*. Berlin, New York: de Gruyter (= Lexicographica: Series Maior).

Engelberg, S., Lemnitzer, L. (^{2}2004). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Faerch, C., Kasper, G. (Hgg.) (1978). *Introspection in Second Language Research*. Clevedon Philadelphia: Multilingual Matters LTD.

Flick, U. (2007). *Qualitative Sozialforschung. Eine Einführung*. Reinbek bei Hamburg: Rowohlt.

Mackintosh, K. (1998). An empirical study of dictionary use in L2-L1 translation. In S. Atkins (Hg.): *Using Dictionaries*. Tübingen: Niemeyer, S. 123–149.

Möhrs, Ch., Müller-Spitzer, C. (2013). *Elektronische Lexikografie*. Tübingen: Groos.

Müller-Spitzer, C. (2013). Contexts of dictionary use. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Hgg.). *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Institute for Applied Slovene Studies/Eesti Keele Instituut, S. 1–15. http://eki.ee/elex2013/proceedings/ eLex2013_01_Mueller-Spitzer.pdf [5.4.2014]

Müller-Spitzer, Carolin (Hg.) (2014): *Using Online Dictionaries*. Berlin, New York: de Gruyter. (= Lexicographica: Series Maior 145).

Müller-Spitzer, Carolin & Koplenig, Alexander (2014, im Druck). Requisitos y expectativas de un buen diccionario online. Resultados de estudios empíricos en la investigación sobre el uso de diccionarios con especial atención a los traductores. In M. J. Domínguez Vázquez, X. G.Guinovart & C. Valcárcel Riveiro (Hgg.). *Lexicografía románica, Vol. 2. Aproximaciones a la lexicografía moderna y contrastiva*. Berlin, New York: de Gruyter.

Müller-Spitzer, Koplenig & Töpel (2012). Online dictionary use. Key findings from an empirical research project. In: S. Granger, M. Paqot (Hgg.). *Electronic lexicography*. Oxford: Oxford University Press, S. 426-457.

Nied Curcio, M. (2005). Verbale Polysemie und ihre Schwierigkeiten im DaF-Erwerb. In D. Di Meola, A. Hornung & L. Rega. *Perspektiven Eins. Tagungsakten der Tagung ‚Deutsche Sprachwissenschaft in Italien' vom 6./7. Februar 2004*. Roma: Istituto Italiano di Studi Germanici, S. 195 – 211.

Nied Curcio, Martina (2011). Der Gebrauch von Wörterbüchern im DaF-Unterricht. Am Beispiel von Übersetzungsübungen. In P. Katelhön, J. Settinieri (Hgg.): *Wortschatz, Wörterbücher und L2-Erwerb*, Wien: Praesens, S. 181– 204.

Sattler, W. (2008). Rund ums Gerund. Das Gerundio und seine Wiedergabe im Deutschen. In M. Nied Curcio. Ausgewählte Phänomene zur Kontrastiven Linguistisch Italienisch – Deutsch. Ein Lehr- und Übungsbuch für italienische DaF-Studierende. Milano: Franco Angeli, S. 98-117.

Taljard, E., Prinsloo, D. & Fricke, I. (2011). The use of LSP dictionaries in secondary schools? a South African case study. In *South African Journal of African Languages*, 31(1), S. 87–109.

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

Wiegand, H. E. (1998). Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 2 Bände. Berlin, New York: de Gruyter.

Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses.

Sascha Wolfer, Alexander Koplenig, Peter Meyer, Carolin Müller-Spitzer
Institut für Deutsche Sprache, Mannheim, Germany
wolfer@ids-mannheim.de

Abstract

We start by trying to answer a question that has already been asked by de Schryver et al. (2006): Do dictionary users (frequently) look up words that are frequent in a corpus. Contrary to their results, our results that are based on the analysis of log files from two different online dictionaries indicate that users indeed look up frequent words frequently. When combining frequency information from the Mannheim German Reference Corpus and information about the number of visits in the Digital Dictionary of the German Language as well as the German language edition of Wiktionary, a clear connection between corpus and look-up frequencies can be observed. In a follow-up study, we show that another important factor for the look-up frequency of a word is its temporal social relevance. To make this effect visible, we propose a de-trending method where we control both frequency effects and overall look-up trends.

Keywords: research into dictionary use; frequency; corpus; social relevance; log file analysis

1 Introduction[1]

In this paper, we use the 2012 log-files of two German online dictionaries (Digital Dictionary of the German Language and the German language edition of Wiktionary) and the 100,000 most frequent words in the Mannheim German Reference Corpus (Deutsches Referenzkorpus, DEREKO) from 2009 (Kupietz et al., 2010) to answer the question of whether dictionary users really do look up frequent words, first asked by de Schryver et al. (2006). The research question standing behind is whether it actually makes sense to select words based on frequency, or, in other words, if it is a reasonable strategy to prefer words that are more frequent over words that are not so frequent. Answering this question is especially important when it comes to building up a completely new general dictionary from scratch and the lexicographer has to compile a headword list. By using an approach to the comparison of log-files and corpus data which is completely different from that of the aforementioned authors, we provide empirical evidence that indicates – contrary to the results of de Schryver et al. and Verlinde & Binon (2010) – that the corpus frequency of a word can indeed be an important factor in determining

[1] In Koplenig, Meyer, & Müller-Spitzer (2014) we present and discuss the results of this study in more detail.

what online dictionary users look up. In addition, we incorporate word class information readily available in Wiktionary into our analysis to improve our results considerably. In a follow-up study, we show that (temporal) social relevance of particular words can influence look-up behaviour considerably. For the latter study, we used the 2013 log files of the German language edition of Wiktionary.

2 Previous research

To understand whether including words based on frequency of usage considerations makes sense, it is a reasonable strategy to check whether dictionary users actually look up frequent words. Of course, in this specific case, it is not possible to design a survey (or an experiment) and ask potential users whether they prefer to look up frequent words or something like that. That is why de Schryver and his colleagues (2006) compared a corpus frequency list with a frequency list obtained from log-files. The aim of De Schryver et al.'s study was to find out whether dictionary users look up frequent words. Due to the nature of the statistical method they used, de Schryver et al. (2006) actually tried to answer two different questions: do dictionary users look up frequent words frequently? And, do dictionary users look up less frequent words less frequently? (cf. Koplenig, Meyer, & Müller-Spitzer, 2014: 232). The result of their study is part of the title of their paper: "On the Overestimation of the Value of Corpus-based Lexicography". Verlinde & Binon (2010: 1148) replicated the study of de Schryver et al. (2006) using the same methodological approach and essentially came to the same conclusion.[2]

In this paper, we will try to show why de Schryver et al.'s straightforward approach is rather problematic due to the distribution of the linguistic data that is used. In this context we suggest a completely different approach and show that dictionary users do indeed look up frequent words (sometimes even frequently). In a follow-up study, we present a case study that suggests that, as soon as frequency information is partialled out, analyses of log-files can also reveal information about the (sometimes very short-lived) social importance of particular words.

3 The Data

Corpus data

When we look at the DEREWO list, a word list compiled using the DEREKO corpus, and plot the relative frequency against the rank, we receive a typical Zipfian pattern. This means that we have a handful of word forms that have a very high frequency and an overwhelming majority of word forms that

[2] In contrast, **Henrik** Lorentzen, Nicolai H. Sørensen and Lars Trap-Jensen in their talk at the e-lexicography conference 2013 also came to the conclusion that frequent words in a corpus are also frequently looked up in a dictionary (Talk: "An odd couple – corpus frequency and look-up frequency: what relationship?" Video available at http://eki.ee/elex2013/videos/ [last access on 02/04/2014].

have a very low frequency. Or, in other words, our DEREWO list consists of 3,227,479,836 word form tokens. The 200 most frequent word form types in the list make exactly half of those tokens.

Log-files

The Wiktionary log file types are roughly 8 w times as many as the DWDS log file types. To make the results both comparable and more intuitive, we rescaled the data by multiplying the raw frequency of a query by 1,000,000, dividing it by the sum of all query tokens and rounding the resulting value. We then removed all queries with a value smaller than one. Thus, the resulting variable is measured in a unit that we would like to call *poms* (per one million searches). For example, a value of 8 means that the corresponding phrase is searched for 8 times per one million search requests. Table 1 summarizes the resulting distribution.

Category (*poms*)	**Wiktionary log-files (%)**	**DWDS log-files (%)**
1	57.94	57.30
2 - 10	33.71	31.15
11 - 49	6.69	9.09
50 - 500	1.63	2.44
500 +	0.03	0.02
Total	**100.00 (abs. 185,071)**	**100.00 (abs. 156,478)**

Table 1: Categorized relative frequency of the log file data.

Category	**X searches *poms***	**Wiktionary log-files (%)**	**DWDS log-files (%)**
regular	at least 1	100.00	100.00
frequent	at least 2	42.06	42.70
very frequent	at least 11	8.35	11.55

Table 2: Definition of the categories used in the subsequent analysis and relative log file distribution.

Table 1 shows two things: firstly, the Wiktionary and the DWDS log-files are quite comparable on the poms-scale; secondly, just like the corpus data, the log-files are heavily right skewed. More than half of all query types consist of phrases only searched for once poms. When we cumulate the first two categories, we can state for both the Wiktionary and the DWDS data that 90% of the queries are requested 1 up to 10 times poms. So there is only a small fraction of all phrases in the log-files that are searched for more frequently.

4 Data analysis

In the previous section, we described the data and presented a new unit of measurement called *poms*. If we think about our research question again – whether dictionary users look up frequent words (frequently) – it is necessary to find an appropriate method for analyzing the data using this unit. For example, we could regress the log file frequency (in poms) on the corpus frequency, but an ordinary least squares (OLS) regression implies a linear relationship between the explanatory and the response variable, which is clearly not given. (Log-)Transforming both variables does not solve our problem, either, and this is in any case seldom a good strategy (O'Hara & Kotze, 2010). We could use the appropriate models for count data such as Poisson regression or negative binomial regression, but, as Baayen (2001, 2008: 222-236) demonstrates at length, we still have to face the problem of a very large number of rare events (LNRE), which is typical for word frequency distributions. And even if we could fit such a model, it would remain far from clear what this would imply for our initial lexicographical question. Using the standard Pearson formula to correlate the corpus and the log-file data suffers from the same nonlinearity problem as the OLS approach. Therefore de Schryver et al. (2006) implicitly used the nonparametric Spearman rank correlation coefficient which is essentially just the Pearson correlation between ranked variables. We believe that this is still not the best solution, mainly because, on a conceptual level, ranking the corpus and log-file data implies that subsequent ranks are equidistant in frequency, which is clearly not the case. Again, the inherent Zipfian character of the distribution explains why the ranks are far from equidistant. For example, the difference in frequency between the first and the second rank is 251,480, whereas the difference between the 3000th and 3001th is only 5. Nevertheless the Spearman rank correlation coefficient treats the differences as equal[3].

In the last section, we grouped the log-files (cf. Table 1) into poms categories. As a possible solutions to the problems we just outlined, we now use this grouping again and stipulate the following categories: if a word form is searched for at least once poms, it is searched for regularly, if it is searched for at least twice, we call it frequent, and if it is searched for more than 10 times, it is very frequently searched for. Table 2 sums up the resulting values. Please keep in mind that according to this definition, a very frequent search term also belongs to the regular and the frequent categories. Our definition is, of course, rather arbitrary and mainly has an illustrative function, but due to the Zipfian distribution of the data, only a minority of the searches (roughly 4 out of 10) occur more than once poms and even fewer words (roughly 1 out of 10, roughly 8 percent for Wiktionary, roughly 12 percent for the DWDS) are searched for more than ten times poms (cf. Table 1). Therefore this definition at least approximates the distribution of the log file data. Nevertheless, instead of using the categories presented in the first column of Table 2, we could also use the second column to label the categories. So it must be borne in mind that the labels merely have an illustrative function.

[3] In principle, we could use another similarity metric, for example the cosine measure (i.e. the normalized dot product, cf. Jurafsky & Martin, 2009: 699), but as in the case of using a count regression model, we are not sure what the value of the coefficient would actually imply both theoretically and practically.

We then wrote a Stata program that starts with the first ten DEREKO ranks and then increases the included ranks one rank at a time. At every step, the program calculates how many of the included word forms appear in the DWDS and Wiktionary log-files regularly, frequently, and very frequently (scaled to percentage). Table 3 summarizes the results for 6 data points.

Included DEREKO ranks	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.0	87.5	99.5	99.5	86.5
2,000	96.9	91.0	67.6	98.4	96.0	64.9
10,000	85.5	72.9	47.5	86.3	75.3	40.2
15,000	80.3	66.5	41.8	77.4	66.1	33.7
30,000	69.4	54.6	31.3	62.7	50.9	23.4

Table 3: Relationship between corpus rank and log file data.

In this table, the relationship between the corpus rank and the log file data becomes obvious: the more DEREKO ranks we include, the smaller the percentage of those word forms appearing regularly/frequently/very frequently in both the DWDS and the Wiktionary log-files. Let us assume for example that we prepare a dictionary of the 2,000 most frequent DEREKO word forms; our analysis of the DWDS and the Wiktionary data tells us that 96.9 % of those word forms are searched for regularly in DWDS, 91.0 % are searched for frequently and 67.6 % are searched for very frequently. For Wiktionary, these figures are a bit higher.

Figure 2 plots this result for the DWDS and the Wiktionary log-files separately. It comes as no surprise that the curve is different for the three categories, being steepest for the very frequent category, since this type of log file data only makes up a small fraction of the data. To further improve our analysis, we looked at the word forms that are absent in both the DWDS and the Wiktionary log-files but that are present in the unlemmatised DEREKO corpus data. There is a roughly 60% overlap, which means that 6 out of ten word forms missing in the DWDS data are also missing in the Wiktionary data. To understand this remarkable figure, we tried to find out more about the words that are missing in the log-files but are present in the corpus data. In our talk, we can present how we used this data to improve our analyses.

We would like to provide an additional impression of our results by asking what proportion of all search requests (tokens) could be covered with such a corpus-based strategy. If we again use the example of the first 15,000 DEREKO most frequent word forms, then around half of all DWDS search requests that occur regularly or frequently (poms) are covered, while around two-thirds of all very frequent requests are successful. If we included the 30,000 most frequent DEREKO words, roughly two-thirds of the regular and frequent and 80.0% of the very frequent DWDS search requests would be covered in the dictionary. In other words, this means if we included the 30,000 most frequent DEREKO word forms, the vast majority of requests would be successful.

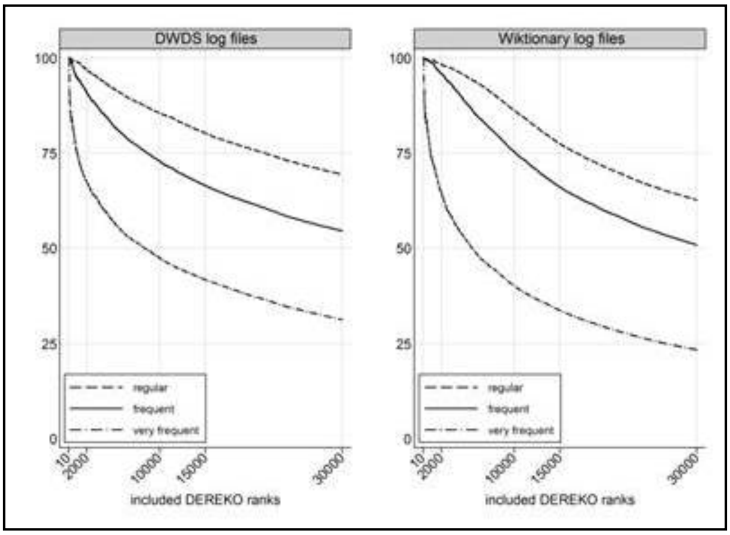


Figure 1: Percentage of search requests which appear in the DWDS/Wiktionary log-files as a function of the DEREKO rank.

5 Social relevance and dictionary usage

The previous sections showed why we think that it is a good idea to include frequently occurring words in dictionaries. Although it is clear that there is a strong and reliable relationship between frequency of occurrence and look-up frequency, it is also quite clear that the former is not the *only* predictor of the latter. To investigate further contributing factors, we consulted log files from the German Wiktionary in the year 2013 and aggregated the hourly Wiktionary log files to weekly datasets to keep the computations manageable. We excluded all pages with titles longer than 80 characters. We also excluded all pages that were visited fewer than one time in one million visits. The overall aim of this study is to identify points in time where certain words are looked up extraordinarily often. To achieve this, we need to control for the overall trend of look-up frequency of each word. It is no surprise that look-up frequency varies over time. Words are looked up more or less often during the course of a year. This variation can be captured by the overall trend *within* a word's visits. By controlling for these long-term trends, we also capture general look-up differences *between* words that stem i.a. from the frequency effects outlined above. What we are currently interested in are rather short-term 'peaks' in the number of visits a specific word receives. The number of visits a specific word receives is the sum of the overall trend for this word and 'noise' which is not captured by this trend (cf. Becketti, 2013:92-95,103,109). This noise, or – informally speaking – what is left over after the overall trend has been considered is exactly the kind of data we are interested in. To extract this variable, we fitted a Tukey smoother using running medians of length 3[4]. This smoother captures the trend. The variable we are going to use is the difference between this smoother and the actual visits, we call this the difference score or residual visits. Using this technique, we can look beyond the effect of frequency and overall tendencies of a specific word. In other words, this technique enables us to identify extraordinary look-up behavior for individual words in individual points in time[5]. To extract interesting words, we rank words by their smooth-difference score. All highly ranking words have especially high proportions of unexplained variance in visits per one million visits in the respective week.

We will now describe two headwords with noticeable difference scores to provide a first impression of our results. The word "Furor" (English "furor", "rage") takes rank 14 of the ordered list of difference scores. The differences between smoothed and observed visits per one million visits are constantly around zero. However, in week 10 of 2013 (04/03/2013 to 10/03/2013), its difference scores go up to 2,210 with a total of 4,687 visits (for all other weeks except week 10, the mean of raw visits for "Furor" is 60.7). In this week, German president Joachim Gauck used the word "Tugendfuror" (roughly: "furor/ rage of virtue") in a debate on sexism in Germany. His whole comment, and especially the word "Tugendfuror" was subject of public discussion in Germany throughout the media. Figure 3 shows the difference scores of "Furor" on a daily basis for one month (20/02/2013 to 20/03/2013). One can clearly see how residual searches *poms* rise for one critical day (06/03/2013) and then take one to two days to 'normalize' again to a residual value around zero.

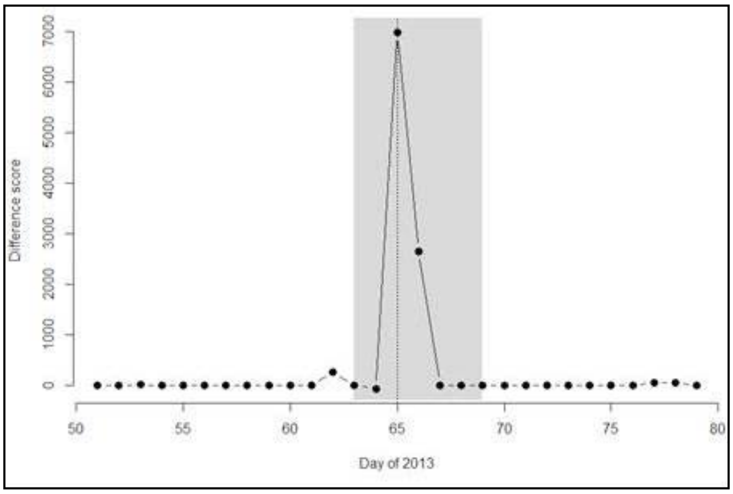


Figure 2: Difference between smoothed and observed visits per 1 million visitsfor "Furor" between 20/02/2013 and 20/03/2013. Week 10 is highlighted.

[4] To do this, we used the default behavior of the function *smooth()* provided by the 'stats' package of the statistical programming language R (R Core Team, 2013).

[5] Of course, these differences can also take negative values. Indeed, many of them do. This means that a word received less searches poms in a particular week than would be expected given the word's overall trend. If we normalize these differences by dividing the difference score by the smoothed visits per one million visits, we can also compare words between one another. In this paper, we did not apply this operation.

Interestingly, Joachim Gauck used the word in an interview[6] already published on 03/03/2014. Though one can see a minimal rise on that day, it took three days until other major newspapers picked up the debate because other voices alluded to potential problems of Gauck's choice of words[7]. Obviously, quite a lot people then wondered what the head of the compound "Tugendfuror" actually means and referred to Wiktionary during those days. "Furor" is a good example of a temporarily socially relevant word. Actually, subject of discussion was the lexical meaning of "Furor" and its connections to other, potentially pejorative words. So, the discussion was lexico-semantic in nature and it comes as no surprise that people tended to look up the word the German president used.

However, there are other noticeable words in certain periods of time, which are not directly related to discussions in society or politics that are lexical in nature. Figure 4 shows the residual searches poms of the word "Borussia" in time. "Borussia" is latin for "Prussia" and part of the name of several German sports clubs. The most prominent ones are football clubs.

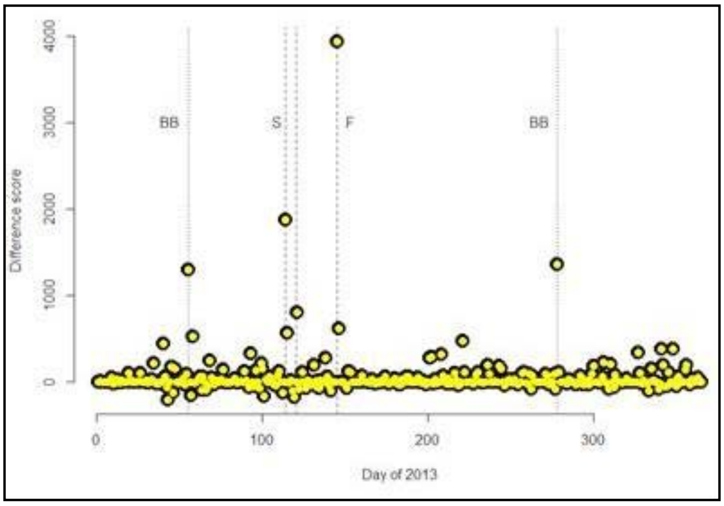


Figure 3: Difference between smoothed and observed visits per 1 million visits for "Borussia" for the whole year 2013. Vertical lines indicate football matches (BB: Borussia Mönchengladbach vs. Borussia Dortmund, S: UEFA Champions League semi-finals, F: Final).

Peaks are identifiable in the difference scores for "Borussia" over time; symbolizing temporarily increased look-ups for "Borussia" in Wiktionary that cannot be explained by frequency of occurrence or overall search preferences alone. Each dashed vertical line in Figure 4 represents one match in the knockout phase of the UEFA Champions League (CL) competition with the participation of Borussia Dortmund. Look-ups of "Borussia" sharply increase around match days. For the semi-finals ("S") and especially the all-important final match ("F"), residual searches poms increase sharply around match

[6] See http://www.spiegel.de/politik/deutschland/sexismus-debatte-gauck-beklagt-tugendfuror-im-fall-bruederle-a-886578.html [last access on 01/04/2014].

[7] See http://www.sueddeutsche.de/politik/sexismus-debatte-als-tugendfuror-aufschrei-wegen-gauck-1.1616310 [last access on 02/04/2014].

days. There are two other vertical lines ("BB") which do not mark a match day in the CL. BB marks 24/02/2013 and 05/10/2013, the days Borussia Dortmund competed against Borussia Mönchengladbach in the German first division. This match is associated with increased difference scores, too. In contrast, no other match in the German first division did lead to increased residual searches poms for "Borussia". Obviously, the popularity and importance of the CL competition led to repeated increases in the social relevance of the term "Borussia". Also, when both Borussias competed against each other in the national championship, public interest in the somewhat cryptic name part was also increased. In comparison to the "Furor" case presented above, the look-up behavior concerning "Borussia" is more surprising. There is no lexico-semantic debate involved that could persuade people to look up "Borussia". Increased media coverage and general public awareness concerning a football club alone seems sufficient to trigger noticeable increases in look-up behavior. This is a remarkable and important observation for research into dictionary use. Another example is the word "larmoyant" (English "lachrymose" meaning "tearfully sentimental") which was used in a sports commentary in an exhibition match between the French and German football national team. Here, the commentator described one specific German national as being too "larmoyant" which led to sharply increased lookups within the same hour (which is the minimum temporal resolution available for the Wiktionary log files).

There are several more interesting cases extractable from the Wiktionary log files that we cannot report here. Social relevance in other cases was induced by a variety of social contexts like TV game shows and even astronomical events like a solstice. Explaining why residual look-ups increased in a specific timeframe is interesting and it definitely points to the fact that the social context directly influences look-up frequencies in internet dictionaries – all in a very short time frame. In future research, however, we want to operationalize social relevance of words in a large-scale, automatized way. Such a measure would enable us to correlate these two measures not only over singular cases but many words. Furthermore, one could identify social contexts that are especially capable (or others that are not capable at all) to trigger look-up peaks in online dictionaries. This line of research could also contribute to the overall question of this paper: Which words should be included in dictionaries? Certainly, words that are socially highly relevant over long periods of time are good candidates.

6 Conclusion

In general, the use of a corpus for linguistic purposes is based on one assumption:

"It is common practice of corpus linguistics to assume that the frequency distributions of tokens and types of linguistic phenomena in corpora have – to put it as generally as possible – some kind of significance. Essentially more frequently occurring structures are believed to hold a more prominent place, not only in actual discourse but also in the linguistic system, than those occurring less often" (Schmid, 2010: 101).

We hope that we have provided evidence which shows that, based on this assumption, corpus information can also be used fruitfully when it comes to deciding which words to include in a dictionary. This corpus-based strategy is no “magic answer”. We simply think it is the best one there is, given that there are no other systematic alternatives.

Beyond that, social relevance of words or their (extraordinary) presence in social discourse seems to be a highly relevant factor in this context.

7 References

Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge, UK: Cambridge University Press.

Becketti, S. (2013). *Introduction to Time Series Using Stata*. College Station: Stata Press.

De Schryver, G.-M., Joffe, D., Joffe, P., & Hillewaert, S. (2006). Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos, 16*, 67–83.

Jurafsky, D. & Marti, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational Linguistics, and speech recognition*. Upper Saddle River: Pearson Education (US).

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, … K. Choukri (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)* (pp. 1848–1854). Valetta, Malta: European Language Resources Association (ELRA).

Koplenig, A., Mayer, P., & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In: C. Müller-Spitzer (Ed.). *Using online dictionaries* (pp. 229-250). Berlin, New York: de Gruyter. (Lexicografica: Series Maior 145).

O’Hara, R.B. & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution, 1*(2), 118-122.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org (last accessed 02/04/2014).

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101–133). Berlin, New York: de Gruyter.

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* (pp. 1144–1151). Ljouwert: Afûk.

Acknowledgements

We are very grateful to the DWDS team for providing us with their log-files.

La performance dell'utente apprendente di italiano LS/L2 e la microstruttura dei dizionari: sussidi per lo sviluppo della Lessicografia Pedagogica

Angela Maria Tenório Zucchi
Università di San Paolo, Brasile
angelazucchi@usp.br; angelatz@gmail.com

Abstract

Questo paper presenta dei risultati quantitativi e qualitativi di una ricerca empirica sulla comprensione di unità lessicali contestualizzate e l'uso dei dizionari, svolta con studenti brasiliani di lingua italiana come lingua straniera presso la Facoltà di Lettere dell'Università di San Paolo. Con una metodologia innovativa nell'usare immagini come risposta nelle alternative a scelta multipla e nell'organizzare i dati raccolti in particolari schede a partire dal comportamento dei tre gruppi di studenti (con dizionario monolingue, con dizionario bilingue e senza dizionario), questo esperimento ha reso possibile verificare statisticamente l'efficacia dell'uso dei dizionari, sia bilingui o monolingui, per la comprensione di parole di uso comune, ma a volte non presenti nei manuali didattici, presentate in testi scritti in italiano. Attraverso la scheda denominata FLICULAD, si è potuto controllare in modo sistematico quali sono state le informazioni contenenti la microstruttura di ogni unità lessicale analizzata che hanno contribuito o meno, secondo l'opinione dello studente, ad arrivare alla comprensione del significato di quella unità in quel contesto.

Keywords: comprensione; dizionario monolíngue; dizionario bilíngue; ricerca empirica

1 Introduzione

Il tema dell'edizione 2014 del Congresso Internazionale EURALEX, "Il focus sull'utente", è particolarmente in sintonia con la ricerca[1] da me realizzata e parzialmente presentata in lingua portoghese nell'edizione del XIV EURALEX Conference, in Olanda. In quell'occasione sono stati esposti la metodologia, con esempi, e i risultati statistici (Zucchi, 2010). Adesso invece saranno presentati i risultati legati all'aspetto qualitativo della ricerca: dati che riguardano il modo in cui le informazioni nella microstruttura dei dizionari possono o meno facilitare la comprensione, dati emersi dal contributo degli informanti, e le relative implicazioni per la Lessicografia Pedagogica, teorica e pratica (Welker, 2008; Xatara et al. 2011).

[1] Ricerca per la Tesi di Dottorato presentata al Programma di Post-Laurea del Dipartimento di Linguistica/FFLCH/USP.

La ragione che ha motivato la ricerca consiste nel fatto che, nonostante il ruolo dei dizionari nell'ambiente scolastico sia riconosciuto, si vedono tuttora pubblicate e commentate opinioni diverse quanto alla efficacia del suo uso nell'apprendimento di una lingua straniera (LS) o lingua seconda (L2); inoltre viene spesso espressa e discussa l'idea che il dizionario monolingue sia più adeguato degli altri all'acquisizione delle lingue (Nunes e Finatto, 2007:40; Corda e Marello, 2004:98). Pertanto, si è voluto controllare, attraverso test basati su lettura e verifica, la reale comprensione di determinate unità lessicali (UL) contestualizzate, con l'uso di un dizionario monolingue (DM – De Mauro, 2000), di un dizionario bilingue (DB – Michaelis, 2003) - ambedue in formato elettronico e allora disponibili *on-line* gratuitamente - o senza l'uso di dizionario (SD).

Da questa ricerca, oltre al risultato quantitativo, si sono potuti raccogliere dati precisi riguardo alle informazioni presenti nella microstruttura dei dizionari considerate utili dagli studenti, per la comprensione delle unità lessicali oggetto dello studio ed evidenziate nei testi letti. Inoltre, si è potuta verificare la maggiore o minore efficacia di alcune delle risorse dei dizionari elettronici. Per chiarire meglio, si riassumeranno qui brevemente la metodologia e i risultati statistici; in seguito, verrà fornita parte dei risultati qualitativi oggetto di questo intervento, derivati dalle osservazioni dei partecipanti sulle informazioni lessicografiche.

2 Metodologia[2], organizzazione dei dati e risultati statistici

I soggetti della ricerca empirica, realizzata per quasi quattro mesi nel primo semestre del 2008, erano ventiquattro studenti brasiliani del corso di laurea in Lettere/Lingue della Facoltà di Filosofia, Lettere e Scienze Umane (FFLCH), della Università di San Paolo (USP), indirizzo di lingua, letteratura e cultura italiana, livelli corrispondenti a A2 e B1 (QCER, 2001), della stessa Facoltà a cui appartiene come docente chi scrive. Alla FFLCH, gli studenti cominciano le lezioni di italiano nel secondo anno della laurea in Lettere/Lingue e, in generale, senza nessuna conoscenza previa di questa lingua straniera. Al primo anno frequentano sei ore di lezioni di lingua italiana obbligatorie e possono frequentare discipline di cultura italiana come complementari. Quelli che hanno collaborato con la ricerca volontariamente erano studenti del primo e del quinto semestre di lingua italiana e nessuno di loro era stato prima in Italia o aveva genitori o dei parenti con cui parlare in italiano.

Prima della realizzazione del test si è spiegato ai partecipanti lo scopo della ricerca e quali sarebbero state le condizioni di lettura per ogni testo presentato. Per conoscere gli informanti, a ciascuno è stato chiesto di compilare un questionario con domande che riguardavano il suo profilo di studente, il suo uso di dizionari e la sua competenza nella consultazione di un dizionario.

Il test consisteva nella lettura di quattro testi autentici (si presenta il testo 4 in appendice) o adattati in lingua italiana, di differente argomento, in cui venivano evidenziate complessivamente quaranta

[2] Un riassunto in inglese della metodologia e dei risultatati di questa ricerca è presentato in Welker (2010)

unità lessicali (qui elencate in appendice), di cui cinque nei due primi testi e quindici negli ultimi due. È importante dire che tutte le UL scelte fanno parte del vocabolario di base dell'italiano, presentato nelle liste del *DAIC – Dizionario Avanzato dell'Italiano Corrente*, di Tullio De Mauro (1997). La loro scelta è stata fatta a seconda della loro disponibilità nei testi e della possibilità della loro rappresentazione in immagine, perché questo è stato il criterio per la verifica della comprensione come si vedrà in seguito. Ci sono voluti due mesi di preparazione per l'elaborazione del test finale e prima di applicarlo definitivamente, è stata realizzata una prova con un piccolo gruppo di studenti di altri corsi di italiano e poi sono stati fatti i dovuti cambiamenti nel test.

Nella ricerca, queste UL sono state denominate "parole-stimolo" e non sono state presentate in un elenco nella loro forma lemmatizzata, ma solo direttamente nel testo, flesse o no, in modo evidenziato in neretto e numerate. Dopo aver letto il testo, il collaboratore riceveva il foglio del test, su cui doveva segnalare l'alternativa corretta (fra A, B, C o D), corrispondente al significato relativo al testo di ogni UL evidenziata, seguendo i passi più avanti descritti. Queste alternative erano costituite da immagini (disegni o fotografie) che alludevano al significato della UL in quel dato testo, secondo il modello metodologico di "stimolo verbale/ reazione non-verbale" (Morales, 1994), si è dunque cercato di selezionare prima dei sostantivi concreti, ma si è verificato possibile creare anche l'immagine di alcuni verbi e sostantivi astratti e in inoltre due unità fraseologiche (*servizio di posate* e *capelli legati*). Fra le alternative, una indicava la giusta corrispondenza e le altre tre erano distrattori, alternative non corrette riguardo al dato enunciato. Si è considerato tuttavia che i distrattori dovevano essere plausibili.

Al collaboratore veniva richiesto di seguire i seguenti passi: 1. leggere prima il testo e osservare le UL lì evidenziate; 2. dichiarare se la UL gli era già nota (nel plico per le risposte del test c'era un foglio per ogni UL); 3. scrivere una possibile traduzione della UL in portoghese; 4. consultare il dizionario (DM o DB a seconda del gruppo; passo non incluso nel test per il gruppo senza dizionari - SD); 5. osservare le figure e segnalare l'alternativa corretta; 6. scrivere nell'apposito spazio le informazioni lette sul dizionario che gli erano state di aiuto e quelle che, al contrario, gli avevano reso più difficile la comprensione della UL. La ricercatrice seguiva i passi degli studenti e chiedeva loro di rispondere nel modo richiesto spiegando che la ricerca aveva obiettivi statistici, che i parametri dovevano essere rispettati e che si voleva verificare se le informazioni fornite dai dizionari erano utili. È stato spiegato che il test non aveva lo scopo di testare la loro competenza linguistica, bensì di verificare i risultati dell'uso dei due dizionari (o la sua mancanza) per la comprensione di quelle determinate unità lessicali contestualizzate.

Il test era uguale per tutti e tre gruppi, con eccezione della consultazione del dizionario per il gruppo senza dizionario, al quale non venivano richiesti i passi 5 e 6. Gli informanti erano ventiquattro, di sesso femminile e maschile, distribuiti ugualmente nei tre gruppi (DM, DB e SD), con età media di ventiquattro anni, tutti di madrelingua portoghese, tutti studenti di Lettere/Lingue, indirizzo Lingua e Letteratura Italiana, livelli A2 e B1 (QCER).

La raccolta dei dati è avvenuta nell'università, durante incontri individuali o in piccoli gruppi. I dizionari utilizzati erano allora disponibili in rete gratuitamente: il monolingue De Mauro, edizioni Paravia (2000), ora fuori consultazione, e il bilingue Michaelis (2003) *on-line*, ancora disponibile. La scelta di dizionari gratuiti in rete è dovuta al frequente uso di questo tipo di dizionari da parte degli studenti e allora specialmente di questi due.

I dati raccolti sono stati organizzati in fogli Excel con le risposte degli studenti alle alternative, la loro traduzione per ogni UL, il modo in cui digitavano le UL nel campo di ricerca, la voce completa data dal dizionario per ogni UL e le informazioni contenenti in essa separate in colonne, poi, nelle righe riferenti a queste informazioni, le osservazioni degli studenti su quello che serviva loro di ausilio o meno per la comprensione. Si è creata a questo fine una scheda denominata FLICULAD (*Ficha lexicográfica informacional da compreensão de unidade lexical com auxílio de dicionário* – Scheda lessicografica informazionale della comprensione di unità lessicale con l'ausilio di dizionario) che ha facilitato l'analisi statistica e poi l'analisi qualitativa dei dati.

I fogli sono stati inviati al *Centro de Estatística Aplicada* (CEA), Centro di Statistica Applicata, dell'Istituto di Matematica della stessa università (IME-USP), e sottoposti ad analisi statistica (tecniche utilizzate: analisi descrittiva unidimensionale, multidimensionale e test di ipotesi non parametriche). Da tale analisi si sono ricavate due tipologie di risultati: 1. relazione tra il profilo degli informanti (dettagliato attraverso il questionario) e la loro performance in termini di percentuale di risposte adeguate; 2. relazione tra tipo di dizionario utilizzato, o suo mancato uso, e percentuale di risposte adeguate (per ogni testo separatamente).

L'informazione più importante emersa da questi risultati è che tutti e due i gruppi che hanno fatto uso del dizionario per la comprensione delle unità lessicali segnalate nel testo hanno avuto più successo del gruppo senza dizionario, per tutti e quattro i testi, come si può verificare nelle tabelle in appendice. L'analisi inferenziale non ha mostrato evidenze che indicassero che l'uso del dizionario monolingue fosse superiore a quello bilingue o viceversa. Il gruppo che ha usato il dizionario monolingue ha avuto una performance migliore in due testi (n.1, n.4), mentre il gruppo con il dizionario bilingue ha avuto una performance migliore nei confronti di altri due (n.2, n.3). In futuro sarà realizzata un'analisi sui vari aspetti dei testi e dei co-testi (tipologia testuale, campo semantico, argomento, ordine di apparizione delle parole ecc) e sulle UL che hanno portato a questo risultato. Da questi risultati statistici si può concludere che in ambedue i dizionari esistono caratteristiche che facilitano la comprensione e altre che la ostacolano. Alcune di queste caratteristiche sono state menzionate dagli studenti e a partire di tali dichiarazioni messe in relazione con i risultati corretti si è fatta l'analisi qualitativa descritta qui di seguito.

3 Dati ricavati dalle dichiarazioni degli studenti e dalle loro azioni

In questa e nella sezione seguente si presentano i risultati dell'analisi fatta a partire dai dati raccolti negli esercizi e posteriormente organizzati nella scheda FLICULAD citata prima, che non viene riprodotta in questo articolo per mancanza di spazio. Ogni UL aveva la sua scheda in foglio Excel, con le informazioni riguardanti gruppi e dizionari DM e DB, dove sono state riprodotte le voci lessicografiche di ogni UL e le corrispondenti informazioni ricavate dalle azioni – digitazione nel campo di ricerca e scelta dell'alternativa - e dichiarazioni (passo n.6 nella metodologia) dei sedici studenti dei due gruppi riguardo alla comprensione di tale UL con l'uso dei dizionari monolingui e bilingui.

Il foglio è stato diviso in due parti orizzontalmente, per il DM e per il DB. Per ogni dizionario, si è riprodotta la voce completa dell'UL in un'unica cella del foglio e in seguito, nella stessa riga, si sono scomposte le stesse informazioni in celle separate: divisione in sillabe; marca d'uso; categoria grammaticale; esempi; polirematiche (denominazione del DM). Anche la definizione è stata scomposta in celle per forma equivalente e per semi identificanti ogni caratteristica, questi ultimi divisi ancora fra *semi descrittivi* e *semi applicativi* (denominazione di Pottier, 1970).

La prima colonna del foglio conteneva l'identificazione degli studenti come **DMLI1** (gruppo dizionario monolingue, studente 1 di Lingua Italiana I) successivamente per gli otto studenti del gruppo e sotto la parte riguardante il dizionario bilingue DBLI**1** (gruppo dizionario bilingue, studente 1 di Lingua Italiana I). La seconda colonna indicava l'alternativa scelta dallo studente: in neretto quelle indicanti l'alternativa giusta. Nella terza colonna venivano riprodotte le digitazioni[3] degli studenti inserite nel campo di ricerca lemma del dizionario elettronico. Sulla riga corrispondente ad ogni studente seguendo le informazioni di ogni colonna (i dati lessicografici), nelle celle è stata segnalata la lettera **F** quando quell'informazione del dizionario (i semi, l'esempio ecc.) ha reso facile la comprensione, secondo quello che ha scritto lo studente sull'apposito spazio nel compilare l'esercizio, oppure **D** quando il dato ha reso la comprensione difficile e sono rimaste vuote le celle corrispondenti alle informazioni che non sono state citate. Le 'marche d'uso', per esempio, non sono state notate da nessuno studente. Nel compilare le schede, oltre a creare spazi uniformi e comparabili, si è lasciato uno spazio perché si annotassero le informazioni che i discenti ritenevano rilevanti nella loro ricerca della UL.

Le alternative scelte e le dichiarazioni dei discenti riguardanti ognuna delle quaranta UL e i due gruppi di studenti (DM e DB) organizzati in un unico foglio hanno permesso una visualizzazione che ha favorito l'analisi dei dati per confrontarli secondo i risultati positivi, le azioni dello studente e le informazioni lessicografiche dei dizionari usati.

Per quanto riguarda il gruppo degli studenti che hanno usato il dizionario bilingue, l'uso del DB è risultato efficiente e efficace tutte le volte in cui esso offriva come primo equivalente quello che corrispondeva a una delle figure presentate e che lo studente riusciva subito ad identificare come pertinente al contesto letto. Invece, non è stato efficace per la comprensione di alcune unità lessicali che presentavano caratteristiche particolari, come quelle riguardanti gli utensili di cucina *mestolo* e *tegame* (testo n.1). Nel primo caso, invece di *mestolo* il DB conteneva la parola *mestola* e il suo equivalente in portoghese *escumadeira*, che fa parte dello stesso campo semantico, ma è un utensile con funzioni diverse. Dal contesto si deduceva facilmente che la mestola non era l'utensile adeguato perché non può

[3] È stato richiesto agli studenti di scrivere in un apposito spazio del foglio di risposta il modo in cui digitavano la UL. Per esempio, la ricerca del lemma corrispondente al verbo coniugato *marciscono* ha rivelato molte difficoltà.

contenere l'acqua; nonostante ciò, alcuni studenti nella risposta hanno selezionato la sua immagine, fatto che può dimostrare maggiore fiducia nel dizionario che nel contesto, o semplicemente un caso di disattenzione. Altri invece hanno scritto che il dizionario ha reso la comprensione più difficile. Nel caso di *tegame*, il DB forniva equivalenti generici come *panela, caçarola* [IT: *pentola, casseruola*], ma questi equivalenti non aiutavano a scegliere fra le quattro figure di pentole diverse presentate, fra le quali appunto il *tegame,* caratterizzato dai bordi bassi. In questo caso, gli studenti hanno annotato che sarebbe stata utile una descrizione dei tipi di pentole, perché non ne conoscevano le differenze.

Si è verificato che la presenza di dettagli descrittivi nella definizione del DM è stata di gran aiuto agli utenti. Per esempio, gli studenti hanno dichiarato che il tratto distintivo *di bordi bassi* nella voce *tegame* li ha condotti alla figura di una pentola con i bordi bassi e che nella definizione di *mestolo* le informazioni *di metallo di forma piuttosto incavata* hanno facilitato la comprensione di quale utensile si trattasse.

Da un'altra parte, l'eccesso di informazioni tecnico-scientifiche è stato indicato dai consultatori del DM come un ostacolo alla comprensione, mentre la presenza di dati culturali li ha favoriti. Per esempio, la voce *alloro* (testo n.1) presenta il primo significato relativo all'albero *alto fino a dieci metri, con foglie coriacee aromatiche* [...], e nella seconda accezione presenta l'uso culturale relativo al *simbolo di sapienza, di gloria [...].* L'italiano e il portoghese, ambedue lingue di origine latina, condividono l'immagine metaforica delle foglie di alloro e ciò facilita la comprensione meglio delle caratteristiche dell'albero. Tuttavia, alcuni studenti hanno indicato che sarebbe stata utile una descrizione più precisa della forma delle foglie, giacché davanti alle fotografie del test (A. *ruta*; B. *prezzemolo;* C. *rosmarino;* D. *alloro*) chi non era un po' esperto di cucina non era in grado di distinguerle. In modo analogo, i semi descrittivi *con* [...] *iridescenze sulla testa e sul collo* nella voce di *piccione* sono stati indicati come causa di difficoltà, mentre l'esempio dato come unità polirematica (nomenclatura di De Mauro per le unità fraseologiche) *piccione viaggiatore* è stata indicata come informazione che facilitava la comprensione.

Dalle azioni degli studenti, si è osservato che, riguardo a tutti e due i dizionari elettronici utilizzati, il fatto che accanto allo spazio di digitazione non fosse presentato il lemmario è stato un fattore sfavorevole per l'apprendente di italiano LS, perché a volte questo utente scrive la parola in modo sbagliato o non ne conosce la forma lemmatizzata presente nei dizionari (singolare, maschile, verbi all'infinito). Per esempio, la UL *piccioni* (testo n.1) è stata digitata spesso in iversi modi, come *picciono, picciona, piccio*, prima di arrivare alla forma base *piccione.* Se ci fosse stato il lemmario sullo schermo, sarebbe stata più facilmente identificata. Nello stesso modo, si verificata molta difficoltà con il verbo coniugato *marciscono*, prima di arrivare a *marcire.* Con lo sviluppo della tecnologia, si spera che i dizionari elettronici possano condurre l'utente alla pagina del lemma anche se viene digitato un verbo coniugato.

D'altro canto, alcuni comportamenti dei discenti hanno ostacolato l'efficace consultazione, come per esempio, tanto nel gruppo DM quanto nel DB, il non proseguire nella lettura completa della voce quando il significato adeguato al testo era uno degli ultimi (es. *busta* e *carrello*, testo n.3). Questo fatto è stato osservato anche in uno studio condotto da Cote Gonzales e Tejedor Martinez (2007) nell'ambito dell'insegnamento dell'inglese in Spagna.

Un'altra difficoltà si è avuta quando la forma flessa di un verbo corrispondeva a un lemma indipendente, come nel caso di *inginocchiata* (testo n.4), sostantivo femminile (un tipo di finestra con inferriata curva e sporgente nella parte inferiore) omografo del participio passato femminile del verbo *inginocchiare* (significato corretto nel testo). In questo caso, alcuni studenti hanno letto soltanto la voce non adeguata e hanno segnalato la definizione come difficoltosa. Anche in questo caso, la presenza del lemmario sarebbe stata utile. Proficua sarebbe stata anche una nota con l'informazione sulla forma omografa del verbo.

4 Contenuto e ordine di informazioni nella microstruttura

Non si intende discorrere su ogni informazione contenuta nella microstruttura dei dizionari consultati, ma soltanto su quelle che hanno contribuito o meno alla comprensione delle unità lessicali evidenziate nel testo.

Nei riguardi del DB, anche se la microstruttura è più semplice di quella del DM, è doveroso notare che si è avuto il cento per cento di risposte corrette quando il DB ha presentato un equivalente pertinente al significato del testo e alla rappresentazione della UL e lo studente è riuscito a trovarlo digitando la forma giusta e scegliendo l'accezione adatta.

Una caratteristica che riguarda la micro e la macrostruttura del dizionario bilingue utilizzato ha influenzato la performance degli studenti: l'organizzazione delle parole omonime e polisemiche. Anziché presentare le parole omonime in differenti lemmi con i loro equivalenti, il DB presenta un unico lemma con tutti i suoi equivalenti, riguardanti le diverse accezioni in portoghese, distinti solo da numeri in neretto (**1. 2. 3.**). Questa caratteristica è stata segnalata dagli studenti come un fattore di difficoltà. La scelta del lessicografo è giustificabile in un dizionario cartaceo, dove esiste il problema dello spazio fisico e raggruppare i significati è una soluzione efficace. Tuttavia in un dizionario elettronico, dove il problema dello spazio non esiste, questa scelta non è la più adeguata, come si è comprovato in questa ricerca.

Dai risultati del gruppo con il dizionario monolingue (DM), si è notato che in tutte le definizioni delle UL in cui c'erano *semi applicativi* indicanti cioè 'qual è l'uso', 'a cosa serve', la presenza di questi semi ha contribuito alla comprensione, secondo l'opinione degli informanti. Gli utenti che hanno notato questi semi nella definizione hanno avuto successo nella comprensione, come nei casi delle UL *piccioni* - sema applicativo: *allevato per le carni gustose* - e *alloro* – sema applicativo: [*le foglie di tale pianta*] *usate in cucina per il tipico gradevole aroma* (testo numero 1).

È stato accennato prima l'insuccesso discente dovuto al non proseguire fino alla fine nella lettura della voce, tuttavia è da segnalare che è auspicabile l'ordine dal più alto al più basso uso nella sequenza delle accezioni, in modo che, *alloro* sia descritto prima come foglie e dopo come albero.

La relazione (osservata nella scheda FLICULAD prima citata) fra il numero di risposte corrette e le dichiarazioni degli studenti riguardo alla microstruttura delle UL mostra quanto l'organizzazione di un'opera lessicografica, specie se in formato elettronico, possa influire sulla comprensione del lemma. Un altro dato importante da considerare è anche l'atteggiamento dell'utente, cioè le sue azioni, che possono o no contribuire al conseguimento del successo nella ricerca di un vocabolo. Su questo punto spetta al professore di italiano LS/L2 il compito di portare il dizionario, "il tesoro della lingua", in aula, far vedere agli studenti le possibilità di consultazioni e sviluppare insieme la loro competenza di utenti.

5 Considerazioni finali

Queste sono solo alcune delle osservazioni ricavate dai risultati ottenuti, che non hanno la pretesa di essere una critica lessicografica ai dizionari utilizzati, ma che possono dare un ausilio a lessicografi, docenti e studenti di italiano LS/L2 nell'elaborazione di nuove opere lessicografiche, nell'affermazione del ruolo dei dizionari in aula, bilingui o monolingui, giacché si è verificato che i risultati della comprensione quando si parte dal contesto (gruppo senza dizionario) sono stati meno efficaci di quelli raggiunti con l'uso dei dizionari, e grazie all'azione del docente nell'insegnare a sfruttare bene questo importante strumento.

Le ricerche empiriche sull'uso dei dizionari offrono nuovi orizzonti per l'elaborazione di opere lessicografiche, principalmente quelle con obiettivi pedagogici, perché, una volta individuate e sistematizzate le attitudini e le opinioni degli utenti, è possibile perfezionare la stesura delle voci e riflettere sull'inclusione o meno di certe informazioni nella micro- e macrostruttura del dizionario.

6 Riferimenti bibliografici

De Mauro, T. (1997) DAIC – Dizionario Avanzato dell'Italiano Corrente. Torino: Paravia, 1997.

De Mauro, T. (2000), *Il Dizionario Italiano On-Line*, prima disponibile http://old.demauroparavia.it/ [10/2009]

Michaelis Dicionário Escolar Italiano-Português (2003) http://michaelis.uol.com.br/escolar/italiano/index.php [10/2009].

Corda, A., Marello, C. (2004) *Lessico insegnarlo e impararlo*. 1°. Ed. 1999, Torino, Paravia e 2°Ed. Perugia, Guerra.

Cote Gonzalez, M., Tejedor Martinez, C. (2007) Dictionary use and translation activies in the classroom. In: Welker, H.A. (Org.) *Horizontes De Linguistica Aplicada*, Ano 6, N.2. Brasilia, UnB.

Morales, H. L. (1994) *Métodos De Investigación Linguistica*. Salamanca, Ediciones Colegio De España.

Nunes, P. A., Finatto, M.J.B. (2007) Dicionários monolíngues para aprendizes de inglês como língua estrangeira: alguns elementos para o Professor. In: Welker, H.A. (Org.) *Horizontes De Linguistica Aplicada,* Ano 6, No.2. Brasília, UnB.

Welker, H. A.(2008) *Panorama Geral da Lexicografia Pedagógica*. Brasília, Thesaurus Editora.

Welker, H. A. (2010) *Dictionary Use. A general survey of empirical studies*. Brasilia, Author's Edition. http://pgla.unb.br/hawelker/images/stories/professores/documentos/2010_Dictionary_use.pdf [04/2014]

Xatara, C. et al [Org.] (2011) Dicionários na teoria e na prática como e para quem são feitos. São Paulo, Parábola.

Pottier, B. (1978) *Linguistica Geral - Teoria e Descrição*. Rio de Janeiro, Presença.

Zucchi, A. M. T. (2010a) O dicionário nos estudos de língua estrangeiras: os efeitos de seu uso na compreensão escrita em italiano. Tese de Doutorado. PPG Semiótica e Linguística Geral, FFLCH, USP. Orientadora: Profa. Dra. Maria Aparecida Barbosa. São Paulo.
http://dedalus.usp.r/F/5862IAKNNJG93L2EM1K6XSADLDTS9SR5UD7XBEG214VAPINLP7-55951?func=full-set-set&set_number=061963&set_entry=000001&format=999

Zucchi, A.M.T. (2010b) O uso de dicionários na compreensão escrita em italiano LE. In Dykstra, A. e Schoonheim, T. *Proceedings of the 14th EURALEX International Congress*. Leeuwarden, Friske Akademy. http://www.euralex.org/proceedings-toc/euralex_2010/ [04/2014]

Appendix 1 – Le unità lessicali

		Le unità lessicali (UL) evidenziate, qui presentate solo nel co-testo
TESTO 1	1	mia madre rosola i **piccioni** in olio
	2	i piccioni in olio in un **tegame**, li sgocciola
	3	soffrigge un **trito** abbondante di cipolla
	4	una foglia di **alloro** e un po' di vino
	5	aggiunge due **mestoli** di acqua calda
TESTO 2	1	si siede a **capotavola** e si accorge
	2	si accorge di aver dimenticato la **dentiera**. Girandosi
	3	infilato una mano in **tasca** ed averne tolto uma dentiera
	4	fa uso anche dello **stuzzicadenti**. Alla fine
	5	Non sono un dentista. Sono un **becchino**."
TESTO 3	1	rifiuti solidi che non **marciscono** per i solidi
	2	di attirare **topi** e scarafaggi.
	3	di attirare topi e **scarafaggi**. Se questo solido
	4	io sono una che non **spreca**. Non sono una mangiona
	5	il mio vitto è pasta o **riso**, pane, formaggio
	6	un po' di **affettati**, e poi gli avanzi
	7	e poi gli avanzi li do alla **cagna** della vicina.
	8	le **scatole** dei regali, invece, le tengo io per organizzare le mie cose
	9	uso le **bottiglie** come vasi.
	10	avevo una **busta** per l'umido
	11	fino a quando **scendevo**,
	12	mi portavo con il **carrello** le mie buste
	13	attaccate dei **manifesti** informativi
	14	con un **timbro** ufficiale del Comune
	15	invece di sprecare **soldi** in corsi professionali di mestieri

TESTO 4	1	quando vede sua madre, Anna, **inginocchiata** davanti a um mucchio di foto
	2	inginocchiata davanti a um **mucchio** di foto
	3	mi ero messa a fare ordine nei **cassetti** e ho trovato delle foto
	4	era così buffo, con quel **pizzo** da diavoletto
	5	e il suo inseparabile **berretto** rosso...
	6	quando noleggiammo il **motoscafo** e ce ne andammo
	7	com papá che non sapeva manovrare il **timone** e ci sballottava
	8	mah, sembro una suora con le **ballerine**, gli occhiali, i capelli legati
	9	mah, sembro una suora con le ballerine, gli **occhiali**, i capelli legati
	10	con le ballerine, gli occhiali, i **capelli legati**, senza un filo di trucco
	11	dai, vieni qua, **bacia** la tua vecchia mamma...
	12	devo scappare, sennò il direttore **si arrabbia** e chi lo sente dopo!
	13	gli ho detto che la prendevi oggi la **valigia**, eh!
	14	voglio chiedergli in prestito il **servizio di posate** francese per la cena
	15	chiedigli di renderti l'**impermeabile** che hai lasciato nella sua macchina.

Appendix 2: Alcuni dei risultati statistici (CEA/USP)

Testo	**proporzione di risposte corrette**	
	media	ds
1	0,64	0,20
2	0,83	0,25
3	0,83	0,13
4	0,76	0,15

Tabella 1: Proporzioni di risposte corrette d'accordo con il testo analizzato.

Testo 1	**proporzione di risposte corrette**	
	media	ds
DM	0,75	0,14
DB	0,70	0,15
SD	0,48	0,18

Tabella 2: Proporzioni di risposte corrette nel testo 1 d'accordo con l'uso del dizionario.

Testo 2		
Gruppi e dizionario	**proporzione di risposte corrette**	
	media	ds
DM	0,90	0,19
DB	0,93	0,10
SD	0,68	0,34

Tabella 3: Proporzioni di risposte corrette nel testo 2 d'accordo con l'uso del dizionario.

Testo 3		
Gruppi e dizionario	**proporzione di risposte corrette**	
	media	ds
DM	0,84	0,14
DB	0,88	0,09
SD	0,76	0,15

Tabella 4: Proporzioni di risposte corrette nel testo 3 d'accordo con l'uso del dizionario.

Testo 4		
Gruppi e dizionario	**proporzione di risposte corrette**	
	media	ds
DM	0,83	0,09
DB	0,79	0,15
SD	0,67	0,16

Tabella 5: Proporzioni di risposte corrette nel testo 4 d'accordo con l'uso del dizionario.

Appendix 3 – Esempio di uno dei quattro testi

Testo 4

Vecchie foto e ricordi

Sono le otto di mattina, Lidia si affretta alla porta perché è leggermente in ritardo per il lavoro, quando vede sua madre, Anna, ***inginocchiata (1)*** *davanti a un* ***mucchio (2)*** *di foto, così assorta che non si accorge della figlia. Lidia, curiosa, le si avvicina:*

Lidia: Mamma, che stai facendo? Buongiorno...

Anna: Buongiorno tesoro... niente, mi ero messa a fare ordine nei **cassetti (3)** e ho trovato delle foto che non vedevo da anni...

Lidia: Ah... fai vedere anche a me.

Curiosa, Lidia si avvicina e comincia a guardare le foto che Anna ha in mano.

Lidia: Guarda guarda! Ma questo non è Franco?

Anna: Proprio lui! Te lo ricordi? Eri così piccola...

Lidia: Piccola! Avrò avuto 14 anni... ma già, per te sono piccola anche adesso... Certo che mi ricordo di Franco! E come potrei dimenticarlo? Era così buffo, con quel **pizzo (4)** da diavoletto e il suo inseparabile **berretto (5)** rosso...

Anna (mostrando una foto) - Questa è quella volta che facemmo la gita in Corsica con papà, ti ricordi?

Lidia: Certo! Quando noleggiammo il **motoscafo (6)** e ce ne andammo tutti a Saint Florent, con papà che non sapeva manovrare il **timone (7)** e ci sballottava tutti da una parte all'altra! Ti ricordi quando ci rovesciammo addosso il caffè appena fatto perché lui fece una manovra brusca e nessuno riuscì a restare in piedi? Come si arrabbiò Franco quando vide che si era macchiato tutto il suo candido completo da marinaio! Me lo ricordo ancora!

Anna: E guardati in questa foto di scuola! Certo che eri proprio carina!

Lidia: Carina! Mah, sembro una suora, con le **ballerine (8)**, gli **occhiali (9)**, i **capelli legati (10)**, senza un filo di trucco...

Anna: La bellezza dell'asino... dai, vieni qua, **bacia (11)** la tua vecchia mamma...

Lidia: Va bene, va bene (la bacia; poi guarda l'orologio); uh, come è tardi! Devo scappare, sennò il direttore **si arrabbia (12)** e chi lo sente dopo! Ciao mamma.

Anna: Ah! E non dimenticarti di passare da Gianni! Gli ho detto che la prendevi oggi la **valigia (13)** eh!

Lidia: Sì, sì, stai tranquilla. Voglio anche chiedergli in prestito il **servizio di posate (14)** francese per la cena di sabato...

Anna: Allora, già che ci sei, chiedigli di renderti l'**impermeabile (15)** che hai lasciato nella sua macchina.

Lidia: È vero... l'impermeabile!! E chi ci pensava più!

Anna: L'impermeabile ti ci vuole proprio ora, che comincia a piovere parecchio.

Lidia: Sì, mamma, non ti preoccupare, lo prendo lo prendo. Ciao, e buoni ricordi!

Anna: Buona giornata cara, a più tardi

Lexicography and Language Technologies

ALIQUOT – Atlante della Lingua Italiana QUOTidiana

Michele Castellarin, Fabio Tosques
Humboldt-Universität zu Berlin
michele_castellarin@yahoo.it, ftosques@gmail.com

Abstract

Der folgende Beitrag erklärt und stellt die Methoden und Resultate dar, die im Projekt ALIQUOT (Atlante della Lingua Italiana QUOTidiana – Atlas der italienischen Alltagssprache) verwendet bzw. gewonnen wurden, um die italienische Alltagssprache zu untersuchen. Das Ziel des Projekts besteht darin, geolinguistische Karten zu erzeugen, die die enorme Vielfalt der regionalen und lokalen Varianten der italienischen Sprache im täglichen Gebrauch sichtbar machen. Die Daten für ALIQUOT werden auf der Basis der indirekten Methode mit einem Online-Fragebogen anonym erhoben. Nur wenige persönliche Daten (Alter, Geschlecht, Wohnort, Beruf, Bildungsgrad usw.) werden abgefragt. Mit Hilfe dieser Daten sollen spätere soziolinguistische Analysen ermöglicht werden. Sämtliche geolinguistische Karten werden online publiziert und sind für jeden frei zugänglich. Wir möchten der Forschung und Lehre mit ALIQUOT ein nützliches Tool in die Hand geben, welches sprachliche Phänomene und Variationen des Italienischen präzise darstellt. ALIQUOT ist damit der erste Sprachatlas, der exakte und umfangreiche Karten der italienischen Alltagssprache anschaulich präsentiert.

Keywords: Geolinguistik; Geosynonym; Digitaler Sprachatlas; Alltagssprache; Regionalismen; Dialekte; sprachliche Variation

1 Das Projekt

Abbiocco, *papagna*, *papazze* oder *cicagna*? Und wie sagt man bei Euch, wenn einen nach dem Essen die typische Müdigkeit überkommt? Das fragte am 10. September 2013 die Moderatorin Isabella Eleodori vom Radiosender *R101* ihre Hörer in einer Sendung nach 15:00 Uhr. Die Antworten der Hörer kamen aus ganz Italien, von Bozen bis Palermo. Nicht ganz klar wurde in der kurzen Umfrage, ob es sich bei den Antworten um Dialektismen oder um Regionalismen bzw. um Alltagssprache handelt. In der Tat ist es häufig nicht ganz einfach zu entscheiden, ob eine bestimmte Bezeichnung nun dialektal, regional oder alltagssprachlich gefärbt ist.

Der enorme Reichtum im Bereich Wortschatz, der häufig von Region zu Region variiert sowie die Variationen im Bereich der morphosyntaktischen und phonologischen Strukturen werden unter der Bezeichnung *italiano regionale* zusammengefasst. Mit diesem Konzept werden alle diatopischen Variationen subsumiert, die sich vom Standarditalienischen unterscheiden und für die Variationen in Italien grundlegend sind.

Die Verknüpfung von sprachlicher Variation und die Darstellung im geographischen Raum wird schon seit vielen Jahrzehnten – besonders in diversen Sprachatlanten – durchgeführt. Dabei ist zu beachten, dass es zwar eine große Zahl von Untersuchungen zu den Dialekten und Minderheitensprachen in Italien gibt, hingegen die regionalen Varianten eher stiefkindlich behandelt werden.

Erstmals wurde das Konzept des *italiano regionale* 1939 von Devoto eingeführt (vgl. Devoto 1939). Dabei handelt es sich um das Ergebnis, welches durch die (sprachliche) Einigung Italiens entstanden ist, als die Nationalsprache auf die vielen dialektalen Varianten traf: „quante sono le regioni italiane, altrettanti sono i tipi di italiano regionale che si vanno costituendo“ (Devoto 1939: 60 cit. in Cerruti 2009: 18-19). Soziale Klassen, die bis dahin ausschließlich Dialekt gesprochen haben wurden zu „creatrici di lingua“ (Cerruti 2009: 18).

Zu den ersten wissenschaftlich-systematischen Untersuchungen zu diatopsichen Variationen des italienischen Wortschatzes zählt die 1956 von Robert Rüegg verfasste Dissertationsschrift „Zur Wortgeographie der italienischen Umgangssprache“ (vgl. Rüegg 1956).

Mit dem Konzept der Regionalismen in Italien beschäftigten sich beispielsweise De Mauro (1963), De Felice (1977), Sobrero (1996) sowie Untersuchungen wie jene von Antonini und Moretti (2000), Sobrero und Miglietta (2006), Cerruti (2007) oder Poggi Salani (2010). Wie schon Devoto in seiner Untersuchung, kommen auch diese Autoren zu dem Ergebnis, dass es sich beim *italiano regionale* um eine Art „Zwischensprache“ (Interlingua) handelt, die zwischen Dialekt und Standardsprache anzusiedeln ist. Sobrero und Miglietta definieren diesen Typ als eine Sprache, die von phonetischen, morphologischen, syntaktischen und lexikalischen Erscheinungen der lokalen Dialekte beeinflusst wird.

Für Telmon hingegen, der von einer „nuova dialettizzazione” (Telmon 1990: 14) spricht, handelt es sich bei den Regionalismen um Wörter „che provengono dal fondo lessicale del dialetto (o dei dialetti) e, trovandosi in un contesto italiano, sono adattate al sistema morfo(no)lessicale [sic!] dell’italiano, quale risulta da analoghe trasferenze ai diversi livelli” (Telmon 1990: 14-15).

Bervor das *italilano regionale* die sprachwissenschaftliche Forschung erreichte, verzeichneten einzelne Wörterbücher diatopische Varianten des täglichen Gebrauchs. Dazu zählen beispielsweise das in den 1950er Jahren erschienen *Dizionario Moderno delle parole che non si trovano nei dizionari comuni* von Alfredo Panzini (1950), das *Grande Dizionario della Lingua Italiana* von Salvatore Battaglia (1972), das *Grande Dizionario Italiano dell’Uso* (GRADIT) (cfr. De Mauro 1999) und das *Vocabolario della Lingua Italiana* (Treccani 2009).

Was sowohl in den Forschungsarbeiten wie auch in den Wörterbüchern fehlt, sind Karten, auf denen die verschiedenen Varietäten im Raum visualisiert werden. Ein erster Versuch, diese Lücke zu schließen und Geosynonyme auf einer Karte zu präsentieren, wurde in *Il Vocabolario della Lingua Italiana 2009* (Treccani 2009) realisiert. Die 100 Karten, von *accendere* bis *vigile urbano*, die im Anhang des Treccani-Wörterbuchs zu finden sind, zeigen einen ersten Ansatz, wie Geosynonyme verzeichnet werden können. Problematisch ist hier jedoch, dass die verschiedenen Bezeichnungen nach Regionen eingetragen wurden und Sprachgrenzen in der Realität nur selten mit den Regionengrenzen übereinstimmen.

Einen anderen Weg der Visualisierung geht das von Elspaß und Möller (vgl. Möller, Elspaß 2008) vor gut zehn Jahren initiierte Projekt *Atlas zur deutschen Alltagssprache* (AdA). Dort werden keine kompletten Regionen nach einer bestimmten Antwort eingefärbt, sondern die Antworten der einzelnen Orte werden als verschiedenfarbige Punkte auf der Karte sichtbar. Möglich wird dies auch durch die von ihnen verwendete Technik, da hier mit Hilfe von Software die geographische Information so ausgewertet wird, dass die Antworten der Informanten problemlos auf den Karten verortet werden können.

Um eine ähnliche Darstellung, wie wir sie im AdA vorfinden, auch für Italien zu ermöglichen, wurde, ausgehend von den im Treccani verzeichneten Karten, die Idee des *Atlante della Lingua Italiana QUOTidiana* (ALIQUOT) geboren. Sinn und Zweck des Projekts ALIQUOT besteht darin, die enorme lexikalische Vielfalt des italienischen Sprachraums anschaulich darzustellen und der Öffentlichkeit einen digitalen interaktiven Sprachatlas zur Verfügung zu stellen. Dafür werden seit Anfang 2013 alltagssprachliche Bezeichnungen zu den unterschiedlichsten Begriffen abgefragt.

In der ersten Fragerunde, die vom 1. Januar bis zum 31. Juli 2013 durchgeführt wurde, wurde nach regionalen Varianten zu den folgenden zehn Lexemen gefragt: *marinare la scuola, lavorare, schiaffo, anguria, topo, fidanzato/a, calzetto, appendiabito, immondizia* und *pungere*. Die Datenerhebung erfolgte bzw. erfolgt ausschließlich online, d.h. die Nutzer sind aufgerufen, einen webbasierten Fragebogen auszufüllen. Dabei haben sie die Möglichkeit, aus vorgegebenen Antworten auszuwählen oder – falls keine davon zutreffen sollte – eine Antwort in das dafür vorgesehene Textfeld („altro“) einzutragen.

Die einzelnen Fragen der Fragerunden sind thematisch kategorisiert, z.B. Obst und Gemüse (*anguria, melone, fagiolini*) oder Gegenstände des täglichen Gebrauchs (*grembiule, balcone, immondizia, appendiabito*) etc. Die Kategorisierungen sollen die spätere Veröffentlichung in thematisch gegliederte Wörterbücher oder Atlanten ermöglichen.

Bei der Auswahl der indirekten Erhebungsmethode stützen wir uns auf die Erfahrungen Eichhoffs, da besonders bei Wortschatzuntersuchungen die „Vorteile der schriftlichen Befragung [= indirekten Befragung] voll zur Geltung [kommen], während die Nachteile das Datenmaterial in seinen wesentlichen Aspekten nicht berühren (Eichhoff 1982: 550). Besonders die Möglichkeit, viele Informanten in kurzer Zeit gewinnen zu können, war ausschlaggebend für diese Form der Erhebung. Nachteile, wie Spektrum der Informanten, Datenqualität, Datenschutz usw., die im Allgemeinen bei online-Erhebungen angeführt werden, scheinen nicht zuzutreffen. Das zeigen auch die jahrelangen Erfahrungen die Elspaß und Möller im Projekt AdA gesammelt haben (vgl. Elspaß, Möller 2006).

Nach Abschluss der ersten Fragerunde konnten wir feststellen, dass die Daten konsistent sind und besonders alle Altersgruppen der Gesellschaft erfasst wurden (vgl. Abbildung 1). Auch der Vergleich unserer Karten mit jenen von *Il Vocabolario della lingua italiana* (Treccani 2009) macht deutlich, dass die Antworten der Informanten weitgehend mit den dort verzeichneten übereinstimmen.

Sowohl die Erhebung der Daten als auch die Präsentation derselben erfolgt mit Hilfe von eigenständig entwickelter Software. Dabei wurde stets darauf geachtet, dass möglichst einfache Techniken eingesetzt werden, die es nicht erfordern, dass der Nutzer zusätzliche Plug-ins o.ä. installieren muss. Als Lösung hat sich hier der sogenannte LAMP-Stack (Linux, Apache, MySQL, PHP) angeboten, da hier freie Tools zur Anwendung kommen und diese sich seit langer Zeit etabliert haben. Zusätzlich setzen wir auf modernste Techniken wie HTML5, CSS3 und verschiedene JavaScript Bibliotheken wie jQuery usw. Die Daten werden direkt nach dem Absenden des Fragebogens in einer relationalen Datenbank gespeichert.

Die Informantensuche erfolgt in erster Linie durch das Anschreiben von Freunden, Bekannten, Kollegen, Studenten und Universitäten, mit der Bitte, den Fragebogen auszufüllen und diesen an möglichst viele Freunde und Bekannte weiterzuleiten. Die Suche erfolgt somit nach dem bekannten Schneeballprinzip. Darüber hinaus nutzen wir für den Kontakt mit den Informanten das am häufigsten genutzte soziale Netzwerk *Facebook*. Das gibt uns die Möglichkeit, unsere Teilnehmer über aktuelle und neue Entwicklungen, die das Projekt betreffen, zu informieren und durch deren Wertungen und Kommentare neue Informanten zu gewinnen. Die vorhandene Technik erlaubt es uns, die eingegebenen Daten permanent zu überprüfen und zu visualisieren. Wo wir noch während der Fragerunde große Lücken in der Datenmatrix entdecken, schreiben wir gezielt Kommunen, Provinzen, kulturelle Einrichtungen und Schulen an. Immer mit der Bitte, den Fragebogen auszufüllen und unsere Umfrage weiterzuleiten.

Obschon wir bei der Erhebung der Daten nur auf das Internet setzen, hat sich gezeigt, dass alle Altersgruppen gut vertreten sind und die Verteilung zwischen männlichen und weiblichen Informanten ausgewogen ist (vgl. Abbildung 1). Trotz der im Grunde zufälligen Auswahl der Informanten konnten wir feststellen, dass die Daten konsistent und vertrauenswürdig sind.

Neben den eigentlichen Fragen zum Lexikon, müssen die Informanten auch einige sozio-demographische Fragen beantworten. Dazu zählen beispielsweise: Wohnort, Postleitzahl, Geburtsort, Alter, Ausbildung, Beruf, Geschlecht sowie der Geburtsort und Wohnort der Eltern. Dank dieser Daten ist es möglich, sozio-demographische Aussagen zu den Informanten zu machen und diese mit den Antworten zu verknüpfen, wodurch u.a. auch soziolinguistische Karten erzeugt werden können.

Die Präsentation der Daten (vgl. Castellarin, Tosques 2012) erfolgt ebenfalls ohne Installation von zusätzlicher Software seitens des Nutzers. Hier wird auf die, für unsere Zwecke völlig ausreichenden Möglichkeiten zurückgegriffen, die von Google Maps mit Hilfe der speziellen API (Application Programming Interface) zur Verfügung gestellt werden. Google Maps bietet schließlich nicht nur die Präsentation der Daten in einer großen Karte an, sondern ermöglicht auch, spezielle Regionen, Provinzen oder Städte heranzuzoomen. Da wir nach der Postleitzahl fragen, erhalten wir mit Hilfe der Zoomfunktion die Möglichkeit, in größeren Städten sehr detaillierte Karten zu erzeugen.[1]

[1] Dabei zeigte sich, dass in den Städten tatsächlich Varianten auftreten, die sich bestimmten Stadtvierteln zuordnen lassen (vgl. Abbildung 7).

2 Ergebnisse

In den folgenden Abschnitten werden exemplarisch fünf Karten aus den drei Fragerunden vorgestellt und zum Teil mit jenen des Treccani-Wörterbuchs verglichen. Auf den Webseiten des Projekts (http://www.atlante-aliquot.de) stehen mit dem Ende der dritten Fragerunde am 30. Juni 2014 Interessierten 41 Karten zur Verfügung. Bisher (Stand: 05/2014) haben insgesamt 3.110 Informanten teilgenommen (1. Runde: 867; 2. Runde: 1406; 3. und aktuelle Runde 837)[2]. Dabei zeigt sich, wie schon erwähnt, dass einerseits alle Altersgruppen hinreichend repräsentiert sind (vgl. Abbildung 1 links) und auch die Verteilung zwischen weiblichen und männlichen Teilnehmern ausgewogen ist (vgl. Abbildung 1 rechts).

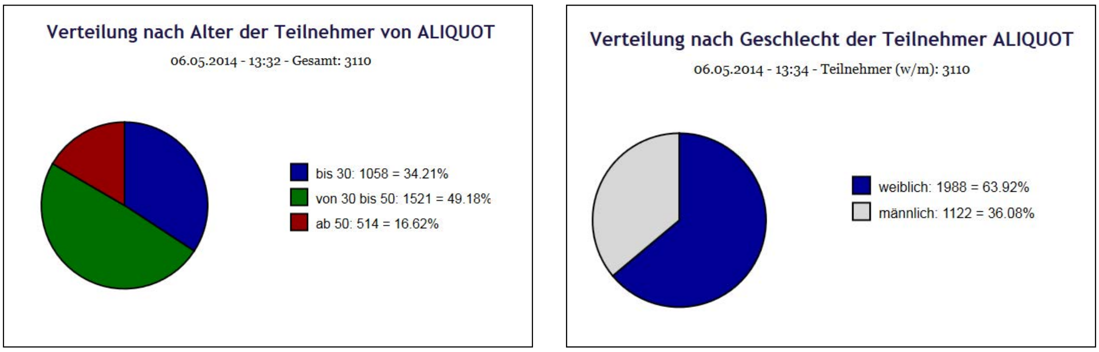


Abbildung 1: Verteilung nach Alter (links) und nach Geschlecht (rechts) der Teilnehmer von ALIQUOT.

Beim Vergleich mit den Treccani-Karten geht es weniger um einen direkten Vergleich der Daten als um einen ersten Eindruck. Es ist uns durchaus bewusst, dass sich unsere und die Treccani-Karten nur mit Einschränkungen vergleichen lassen. Haben die Treccani-Karten (vgl. Abbildung 2 und 3 rechts) modellhaften Charakter und bilden die Realität eher synoptisch ab, kommen unsere Karten (vgl. Abbildung 2 und 3 links, 4, 5 und 6), denen eine völlig andere Datenbasis zu Grunde liegt, der sprachlichen Realität weitaus näher auch wenn diese natürlich immer noch komplexer ist als die, die sich – sei es in gedruckter, sei es in elektronischer Form – darstellen lässt. So zeigen die Treccani-Karten die Verwendung der Geosynonyme pauschal in den Grenzen der Regionen Italiens. Von einer regionen-basierten Darstellung haben wir von Anfang an Abstand genommen, da die administrativen Grenzen nur selten deckungsgleich sind mit den Sprachgrenzen. Nichtsdestotrotz sind die Treccani-Karten für uns eine große Hilfe: erstens bei der Auswahl der Geosynonyme für vergangene, laufende und zukünftige Fragerunden und zweitens für einen ersten Eindruck von der Qualität der im Projekt ALIQUOT erhobenen Daten.

[2] Vgl. http://www.atlante-aliquot.de/aliquot/showall_locations.php.

Die von uns generierten Karten stehen somit nicht in direkter Konkurrenz zu den Treccani-Karten. Letztere dienen dem Projekt als Ausgangsbasis für die Präsentation der in Italien verwendeten Alltagssprache. Da es sich bei der Alltagssprache um die am häufigsten verwendete und – besonders im Vergleich zu den Dialekten – weit weniger untersuchten Sprachform handelt, legt das Projekt ALIQUOT ebenfalls den Focus auf deren Verwendung, Formenreichtum und detaillierten Darstellung[3].

2.1 Fragerunde 1: Karte zu *lavorare* („arbeiten")

Die für ALIQUOT erhobenen Geosynonyme für *lavorare* bestätigen die allseits bekannte Dreiteilung der Halbinsel, wie sie in der Literatur, z.B. in De Felice (1984), Coveri, Benucci, Diadoro (1998) oder Grassi, Sobrero, Telmon 2003, in Sprachatlanten wie beispielsweise dem *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS, vgl. Jaberg, Jud 1928-1940) oder dem *VIVaio Acustico delle Lingue e dei Dialetti d'Italia* (Kattenbusch 1999 ff.) sowie im *Vocabolario della lingua italiana* Treccani (2009) beschrieben wurde. Durchgehend wird dort der Sprachraum in einen nördlichen (*lavorare*) und einen südlichen (*faticare*) unterteilt. In der dritten und kleinsten Zone, die Sizilien, Sardinien, den äußersten Westen Piemonts und Ligurien betreffen, benutzen die Sprecher Formen von *travagliare*.

Die im Treccani Wörterbuch publizierte Karte zu *lavorare* zeigt klar und deutlich, wo die drei Zonen der Verteilung der Geosynonyme von *lavorare* liegen (vgl. Abbildung 2 rechts).

Die im Projekt ALIQUOT erhobenen Daten zu *lavorare* (vgl. Abbildung 2 links) unterscheiden sich im Ergebnis nur leicht von jenen, die im Treccani Wörterbuch dargestellt werden.

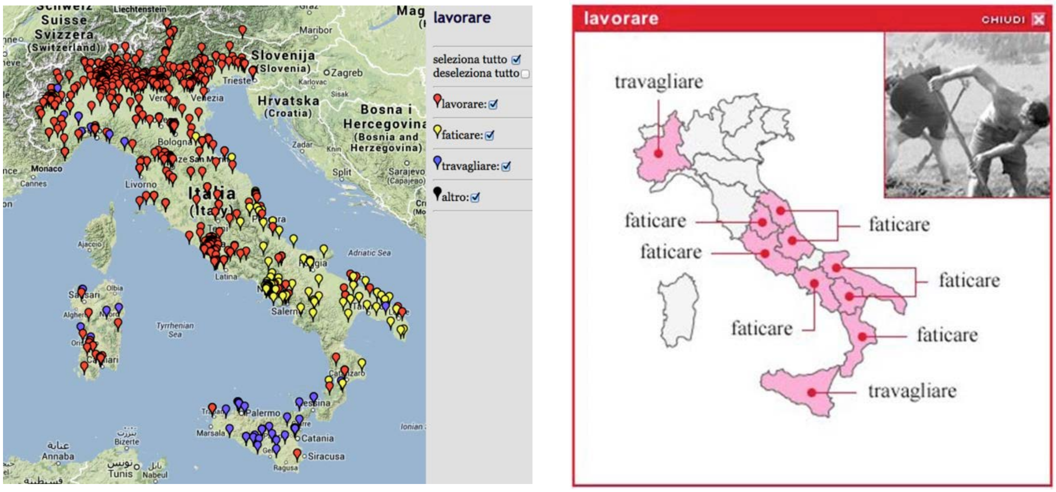


Abbildung 2: ALIQUOT-Karte (links) und Treccani-Karte (rechts) der Geosynonyme für das Lexem *lavorare*.

[3] Wir möchten darauf hinweisen, dass die hier vorgestellten Karten teilweise etwas vereinfacht wurden. Dies hängt damit zusammen, dass das Punktenetz von ALIQUOT sehr engmaschig ist und die Karten aufgrund der zahlreichen Varianten und der hohen Teilnehmerzahl ohne graphische Aufbereitung ansonsten in gedruckter Form in diesem Größenformat nur schwer lesbar wären.

Die Karte scheint in jedem Fall die Annahme zu bestätigen, dass der Dialekt auf die Formen der Alltagssprache einwirkt. Dazu bedarf es jedoch weiterer Annahmen. Eine erste betrifft die offensichtliche Ausdehnung der Standardsprache, d.h. des Lexems *lavorare* bis in den Süden der Halbinsel. Während beispielsweise die wenigen Punkte von *faticare* und *travagliare*, die außerhalb der typischen Sprachgrenzen liegen, vermutlich auf den familiären Ursprung des Informanten zurückzuführen sind, trifft dies für *lavorare* nicht zu. Schon ein flüchtiger Blick zeigt, dass *lavorare* die dialektale Form *travagliare* in Piemont und Sardinien so gut wie ersetzt hat und auf dem besten Wege zu sein scheint, die regionale Form in Zentren wie Neapel, Bari, Lecce usw. zu verdrängen.

Der Gebrauch eines Geosynonyms anstelle eines anderen kann – was der hier beschriebene Fall zeigt – sehr wahrscheinlich auf eine systematische Italianisierung zurückgeführt werden. Dieser Prozess hat mit der politischen Einigung Italiens begonnen. So scheinen die südlichen Varianten nach und nach vom Standarditalienischen verdrängt zu werden.

Voraussagen, die das „große Rätsel der Sprachwissenschaft“ (Nützel 2009: 85), den Sprachwandel, betreffen sind häufig schwer zu treffen und können postwendend unseriös werden. Dennoch möchten wir anhand der dargestellten Situation kurz über die Zukunft der beiden Lexeme *faticare* und *travagliare* nachdenken. Die gegenwärtige Konstellation zeigt, dass die hochsprachliche Variante *lavorare* sich auch im Süden offensichtlich weiter verbreiten wird und damit eine immer größere Bedeutung im Wortschatz des täglichen Gebrauchs der Sprecher einnehmen wird.

2.2 Fragerunde 1: Karte zu *fidanzato/a* („Verlobter/Verlobte“)

Bezüglich der Geosynonyme von *fidanzato/a* werden im Treccani-Wörterbuch drei zusätzliche Varianten aufgezeigt (vgl. Abbildung 3 rechts). Im Norden der Halbinsel Formen von *moroso/a*, in Umbrien *frego/a* und in Kalabrien und Sizilien *zito/a*. Alle weiteren Regionen bleiben weiß, was vom Leser so verstanden werden könnte, dass hier die standardsprachliche Form *fidanzato/a* verwendet wird.[4] Die ALIQUOT-Karte zeigt eine sprachliche Situation, die komplexer ist als jene, die im Treccani-Wörterbuch verzeichnet ist. Ein erster Blick auf unsere Karte verdeutlicht, dass *zito/a* eine größere Verbreitung erfährt als auf der Treccani-Karte, d.h. neben Kalabrien und Sizilien auch die Basilikata und Teile Apuliens.

Auch bei der Verbreitung des Lexems *moroso/a* werden Unterschiede deutlich. Während im Treccani-Wörterbuch ganz Norditalien – ohne Trentino-Südtirol – genannt wird, zeigt die ALIQUOT-Karte hier und da Unterschiede: wird im Veneto ausschließlich *moroso/a* für die „geliebte Person“ verwendet, ist in Piemont und in der Lombardei *fidanzato/a* die gebräuchlichste Form. Auch bei der Verbreitung in Mittelitalien werden Unterschiede deutlich: so wird in der Emilia-Romagna großteils *moro-*

[4] Dabei ist zu beachten, dass die Karten im Treccani-Wörterbuch keinen Anspruch auf Vollständigkeit haben. Das zeigt ein Blick auf alle 100 Karten. Häufig bleiben Regionen weiß, obschon dort eigentlich Varianten des Standards zu erwarten wären. Sehr deutlich wird dies z.B. bei der Region Molise, wo laut Treccani durchgehend die hochsprachliche Variante verwendet wird.

so/a verwendet. Das Lexem *frego/a*, welches im Wörterbuch von Treccani für Umbrien angegeben wird, ist auf der ALIQUOT-Karte nur sporadisch in dieser Region vorhanden.[5]

Ein anderes Geosynonym, welches bei Treccani nicht erwähnt wird, laut der Informanten von ALIQUOT aber vor allem in den großen Zentren wie Rom, Neapel, Cagliari, Florenz, Mailand und Turin zu finden ist, ist *ragazzo/a*. Ebenfalls nur bei ALIQUOT verzeichnet ist das Geosynonym *sposo/a*, das auch in anderen Teilen Süditaliens zu erwarten wäre.

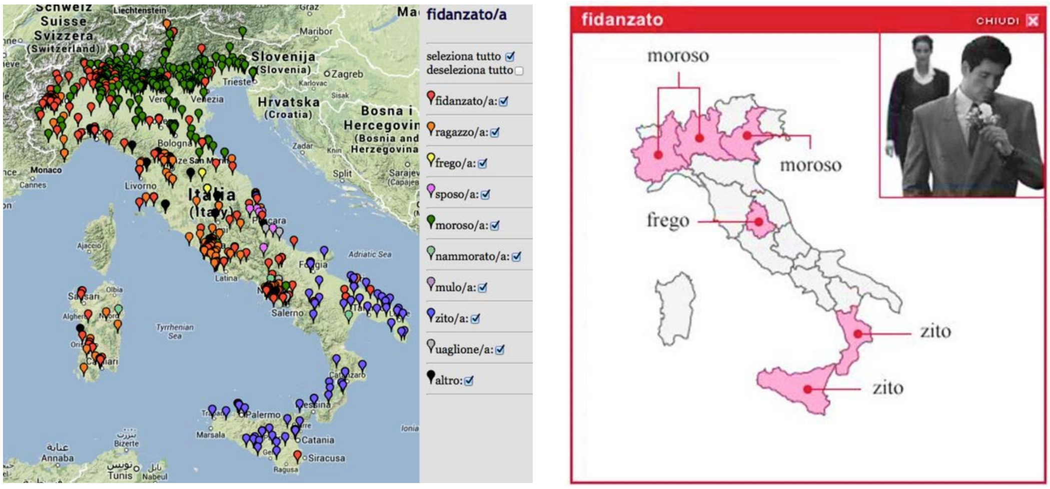


Abbildung 3: ALIQUOT-Karte (links) und Treccani-Karte (rechts) der Geosynonyme für das Lexem *fidanzato/a*.

Die im Projekt ALIQUOT erhobenen Daten zeigen somit für *fidanzato/a* und *ragazzo/a* ein anderes Bild als das Treccani-Wörterbuch. *Fidanzato/a* wird – wie zu erwarten – in der Toskana gebraucht, zeigt jedoch auch eine beachtliche Verbreitung auf der gesamten Halbinsel. Besonders in Rom, Neapel, Pescara, Sardinien, im westlichen Piemont, in Ligurien, Venetien, Friaul, im Trentino und etwas weniger in Apulien und Sizilien.

Ähnlich sieht es auch für den Terminus *ragazzo/a* aus. Ursprünglich vor allem in Rom verwendet, hat der Begriff eine beachtliche Ausdehnung erfahren, wenn auch nicht in dem Maße wie *fidanzato/a*.

Formen von *ragazzo/a* werden, wie auf der Karte zu erkennen ist, auch in den großen Zentren Italiens verwendet. In Turin haben beispielsweise *ragazzo/a* und *fidanzato/a* die zu erwartende Form *moroso/a* vollständig verdrängt.

[5] Dabei darf nicht unerwähnt bleiben, dass die Region Umbrien in der ersten Fragerunde nicht flächendeckend beantwortet wurde.

2.3 Fragerunde 2: Karte zu *padre/papà* („Vater/Papa“)

Eine Karte zu den Geosynonymen für *padre* ist im Treccani-Wörterbuch nicht vorhanden. Daher können unserer Ergebnisse nicht verglichen werden. Grossomodo existieren für *padre* fünf Geosynonyme in Italien (vgl. Abbildung 4). In der Toskana und in den Marken hören wir hauptsächlich Formen von *babbo*. In der Basilikata und in Apulien (ohne Salento) ist die häufigste Form *attane*. Im Salento selbst wird *sire* verwendet und seltener *tata*. Die Informanten ALIQUOTs antworteten im Norden Italiens hauptsächlich mit dem Terminus *papà*. Bei genauerer Betrachtung der Karte fällt auf, dass das Lexem *padre* in der Alltagssprache so gut wie unbekannt ist. Nur sehr wenige Informanten haben von den fast 1500 Teilnehmern die hochsprachliche Form *padre* als gebräuchlichste angegeben.
Sprachlich lassen sich auf der Karte für *padre* deutliche Zonen erkennen. Sofort ins Auge fällt das am häufigsten gebrauchte Lexem *papà*. In der Toskana und im Norden Sardiniens wird durchgehend *babbo* verwendet.
In Süditalien konkurrieren neben der häufigsten Form *papà* verschiedene Geosynonyme: *attane*, *tata* und *sire*.

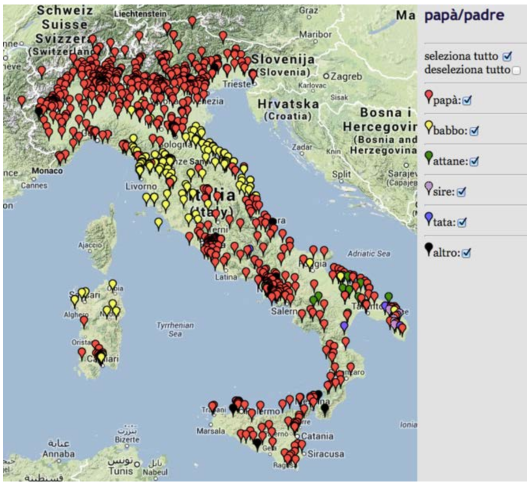


Abbildung 4: ALIQUOT-Karte für die Geosynonyme des Lexems *padre.*

2.4 Fragerunde 2: Karte zu *gomma da masticare* („Kaugummi“)

Als letzte lexikalische Karte möchten wir im folgenden die Ergebnisse für *gomma da masticare* vorstellen. Dabei handelt es sich um einen relativ jungen Begriff, der erst nach dem Zweiten Weltkrieg in den italienischen Wortschatz Eingang gefunden hat.
Die Karte (vgl. Abbildung 5) stellt eine Besonderheit dar, da hier eine Fremdsprache (englisch) auf die einheimische Sprache bzw. den einheimischen Dialekt trifft. Deutlich wird, dass die englische Be-

zeichnung und die englische Aussprache (*chewing-gum*) nur wenig vertreten sind. Die meisten Sprecher ziehen eine wörtliche Übersetzung des englischen Terminus dem Original vor und benutzen Formen von *gomma da masticare* oder einfach die Kurzform *gomma*. Diese beiden Formen verteilen sich über den ganzen Sprachraum. Eine beachtliche Teilnehmerzahl, die über ganz Italien verteilt ist, antwortet mit phonetischen und/oder morphologischen Varianten der englischen Form. So wird in manchen Gegenden aus *chewing-gum ciunga* oder *cevingum* [ˈtʃɛːviŋgum]. Daneben finden sich auch Lehnübersetzungen wie *cingomma*, *gingomma* oder *cincingomma*. Sehr weit verbreitet sind die beiden Formen *cicca* (von Norditalien entlang der Adriaküste bis nach Apulien) und *cicles* (besonders in Turin und im Großraum Bologna).

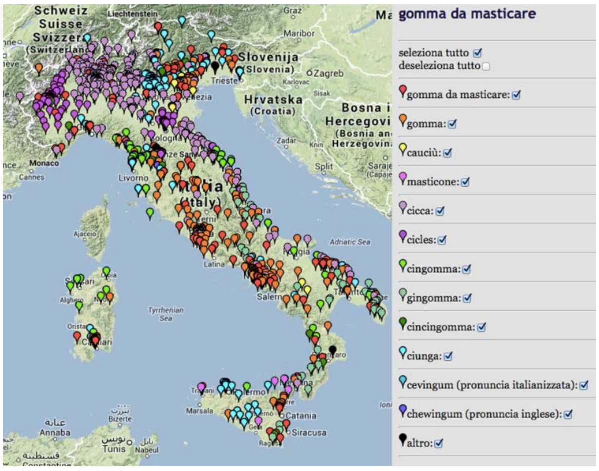


Abbildung 5: ALIQUOT-Karte für die Geosynonyme des Lexems *gomma da masticare*.

2.5 Fragerunde 3: Karte zum transitiven Gebrauch intransitiver Verben

Wir möchten den vorliegenden Beitrag nutzen, um vorausschauend die Ergebnisse zu betrachten, die den transitiven Gebrauch von intransitiven Verben untersucht.[6] Ein flüchtiger Blick auf die Karte (vgl. Abbildung 6) verdeutlicht eine Zweiteilung des Untersuchungsgebiets. Im Norden und in Mittelitalien (Toskana, Marken, Umbrien) werden Aussagen wie *scendimi le chiavi* („bring mir den Schlüssel nach unten") als grammatisch falsch interpretiert. Im täglichen Gebrauch werden sie nicht verwendet. Südlich der Linie Neapel-Pescara werden die Aussagen als korrekt empfunden und in der Alltagssprache gebraucht.

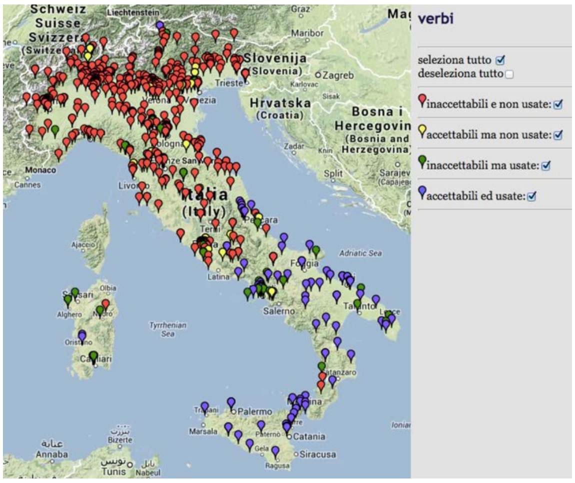
[6] Vorausschauend deshalb, da es eine Frage betrifft, die Teil der aktuellen dritten Fragerunde ist, die vom 01. Januar 2014 bis zum 30. Juni 2014 läuft. Zum Zeitpunkt der Erstellung der Karte haben ca. 850 Personen geantwortet. Wie auf der Karte zu sehen ist, sind die Punkte so gut verteilt, dass eine erste Aussage getroffen werden kann.

Offen bleiben noch die Antworten *accettabili ma non usate* („akzeptabel aber nicht verwendet“) und *inaccettabili ma usate* („inakzeptabel aber verwendet“). Der erste Fall trifft nur auf Nord- und Mittelitalien zu, wo – wie gesehen – der Großteil der Sprecher Aussagen dieses Typs als ungrammatisch komplett ablehnt. Sprecher, die die Meinung vertreten, solche Ausdrücke seien *inaccettablili ma usate* sollten sich konsequenterweise nur südlich der oben genannten Linie finden.

Überraschend ist vielleicht die Wahl der Antwort *inaccettablili ma usate* in den norditalienischen Zentren wie Mailand, Turin, Venedig und Bologna. Eine Einordnung der dort gegebenen Antworten erfordert jedoch weitere Untersuchungen, die besonders die soziolinguistischen Aspekte sowie die sozio-kulturellen Umstände der Informanten genauer unter die Lupe nehmen. Da sich diese Antwort auf die großen Städte beschränkt, sollte hier die interne Migration vom Süden nach Norden genauer untersucht werden.

Eine allgemein gültige Erklärung für dieses Phänomen wurde bisher noch nicht gefunden. Die Antworten lassen sich vielleicht auf das Sprachbewusstsein der Italiener zurückführen. Während für Norditaliener Aussagen wie *esci il cane* („führ den Hund Gassi“) zwar unüblich sind und nicht gebraucht werden, sind für Süditalien solche Aussagen typisch.

Aussagen wie diese werden von einer großen Sprechergruppe im Süden als akzeptabel empfunden und sie benutzen sie auch im täglichen Gebrauch. Interessanterweise findet sich in unseren Antworten im Süden auch eine Gruppe, der durchaus bewusst ist, dass es sich um eine inkorrekte Form handelt, diese aber dennoch Verwendung findet.



verbi

seleziona tutto ☑
deseleziona tutto □

🟥 inaccettabili e non usate: ☑
🟨 accettabili ma non usate: ☑
🟩 inaccettabili ma usate: ☑
🟪 accettabili ed usate: ☑

Abbildung 6: ALIQUOT-Karte für den transitiven Gebrauch intransitiver Verben.

3 Ausblick und Zukunftsperspektiven

Die Entscheidung, die Karten von ALIQUOT ausschließlich online zu publizieren, wurde aus zwei Gründen gefällt: Erstens können wir so schnell und unkompliziert die Karten schon kurz nach dem Ende eines Fragezyklus publizieren. Zweitens haben wir damit die Möglichkeit, Probleme der Visualisierung, die mit der Komplexität der Karten einhergehen, relativ einfach zu lösen.

Nicht unwesentlich ist, dass es sich bei den ALIQUOT-Karten um interaktive Karten handelt und nicht um statische. So erlaubt uns die Technik, einzelne Geosynonyme an- und abzuwählen, wodurch die Lesbarkeit und der Erkenntnisgewinn bezüglich der Verteilung im Raum enorm verbessert wird.

Der Technik, namentlich von Google Maps, verdanken wir auch die schon erwähnte Zoomfunktion, da dabei Variationen innerhalb einer Stadt sichtbar gemacht werden können. Hilfreich ist, dass wir nach den Postleitzahlen fragen und diese als Referenzpunkt für die Darstellung dient (vgl. Abbildung 7). Auf der Karte *anguria* zeigt sich, dass im Großraum Rom die Informanten am häufigsten mit dem für den mittelitalienischen Raum typischen Lexem *cocomero* (blau) antworten. In den nördlichen Stadtvierteln (Montemario, Vittorio, Nomentano) hingegen wurde erstaunlicherweise mit *anguria* (gelb) geantwortet, ein Lexem, das v.a. in Norditalien zu erwarten wäre. Weshalb auch in Rom mit der nördlichen Variante (*anguria)* geantwortet wird, obschon wir in Mittelitalien ausschließlich *cocomero* vorfinden und damit keine Transitionszone vorhanden ist, erfordert eine genauere Untersuchung. In jedem Fall zeigte sich, dass in den genannten Stadtvierteln im Norden Roms auch bei anderen unserer Fragen eher mit einem standarditaliensichen Lexem geantwortet wurde als mit dem für Rom und Umgebung typischen.

Ziel von ALIQUOT ist, in der Zukunft detaillierte Studien zur diatopischen und diastratischen Variation der italienischen Alltagssprache auf der Halbinsel und in der italienischen Schweiz vorzunehmen. Dafür erheben wir bei den Nutzern einige sozio-demographische Daten wie das Alter, Geschlecht, die Schulbildung usw. Eine Verbindung der sozio-demographischen Daten mit den linguistischen führt somit zu neuen Ergebnissen, die im Bereich der Alltagssprache noch nicht genauer untersucht wurden.

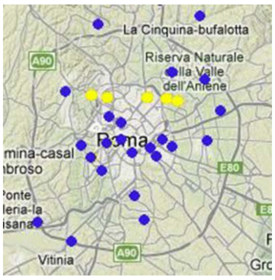


Abbildung 7: Varianten zur Frage *anguria* innerhalb Roms (*cocomero* blau, *anguria* gelb).

Schließlich erlaubt uns die Technik auch den Ursachen für den Gebrauch bestimmter Geosynonyme auf den Grund zu gehen. Häufig liegen den alltagssprachlichen Bezeichnungen dialektale Formen zu Grunde, die glücklicherweise in den Sprachatlanten vorbildlich dokumentiert sind. So können die Daten von Sprachatlanten wie dem AIS oder dem ALI mit den Daten von ALIQUOT korreliert werden, wodurch die Interferenzen zwischen Dialekt und Alltagssprache verdeutlicht werden.

Zuletzt soll noch die Möglichkeit erwähnt werden, die Karten von ALIQUOT in einem gedruckten Atlas zusammenzufassen oder diese als Basis für zukünftige *joint-ventures* mit Lexikographen oder Lexikologen zur Verfügung zu stellen und damit die verschiedenen Kompetenzen optimal zu vereinen. So könnten erstmalig detaillierte Karten zur italienischen Alltagssprache entstehen, die dem Formenreichtum derselben würdig wären.

4 Literatur

AdA: Elspaß, S, Möller, R. (2001). *Atlas zur deutschen Alltagssprache.* [http://www.atlas-alltagssprache.de].

AIS: Jaberg, K., Jud, J. (1928–40). *Sprach- und Sachatlas Italiens und der Südschweiz.* 8 vol. Zofingen: Ringier.

ALI: Bartoli, M. G. (1995). *Atlante linguistico italiano.* Roma: Istituto Poligrafo e Zecca dello Stato.

ALIQUOT: Castellarin, M, Tosques, F. (2013): *Atlante della Lingua Italiana QUOTidiana.* [http://www.atlante-aliquot.de].

Antonini, F., Moretti, B. (2000). *Le immagini dell'italiano regionale. La variazione linguistica nelle valutazioni dei giovani ticinesi.* Locarno: Dadò.

Battaglia, S. (1972). *Grande Dizionario della Lingua Italiana.* Vol. 21. Torino: UTET.

Castellarin, M., Tosques, F. (2012). ALIQUOT – L'Atlante della Lingua Italiana QUOTidiana. In *Rivista Italiana di Dialettologia. Lingue dialetti società.* XXXVI, pp. 245-262.

Cerruti, M. (2009). *Strutture dell'italiano regionale. Morfosintassi di una varietà diatopica in prospettiva sociolinguistica.* Frankfurt a. M. u.a.: Peter Lang.

Cerruti, M. (2007). Sulla caratterizzazione aspettuale e la variabilità sociale d'uso di alcune perifrasi verbali diatopicamente marcate. In *Archivio Glottologico Italiano* a. 92, n. 2, S. 203-247.

Coveri, L., Benucci, A. & Diadori, P. (1988). *Le varietà dell'italiano: manuale di sociolinguistica italiana; con documenti e verifiche.* Roma: Bonacci.

De Felice, E. (1977). Definizione del rango, nazionale o regionale, dei geosinonimi italiani, in Italiano d'oggi. Lingua nazionale e varietà regionali. In *Atti del Convegno internazionale di studio* (Trieste, 27-29 maggio 1975). Trieste: Lint, S. 109-117.

De Felice, E. (1984). *Le parole d'oggi: il lessico quotidiano, religioso, intellettuale, politico, economico, scientifico, dell'arte e dei media.* Milano: Mondadori.

De Mauro, T. (1963). *Storia linguistica dell'Italia unita.* Bari: Laterza.

Devoto, G. (1939). La norma linguistica nei libri scolastici. In *Lingua nostra*, 1. Frankfurt a. M.: Vitorio Klostermann, S. 57-61.

Eichhoff, J. (1982). Erhebung von Sprachdaten durch schriftliche Befragung. In W. Besch u.a. (Hrsg.) *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung.* 1. Halbband. Berlin u.a.: De Gruyter, S. 549-553.

Elspaß, S., Möller, R. (2006). Internet-Exploration: Zu den Chancen, die eine Online-Erhebung regional gefärbter Alltagssprache bietet. In *Osnabrücker Beiträge zur Sprachtheorie* 71: S. 141-156.

GRADIT: De Mauro, T. (a cura di) (1999). *Grande Dizionario Italiano dell'uso.* Torino: UTET.

Grassi, C., Sobrero, A. A. & Telmon, T. (2003), *Introduzione alla dialettologia italiana.* Bari: Laterza.

Möller, R., Elspaß, S. (2008). Erhebung dialektographischer Daten per Internet: Ein Atlasprojekt zur deutschen Alltagssprache. In S. Elspaß, W. König (a cura di). *Sprachgeographie digital. Die neue Generation der Sprachatlanten (mit 80 Karten),* Hildesheim u.a.: Olms. S. 115-132.

Nützel, N. (2009). *Sprache oder Was den Mensch zum Menschen macht.* München: cbt Verlag.

Panzini, A. (1950). *Dizionario moderno delle parole che non si trovano nei dizionari comuni.* Milano: Hoepli.

Poggi Salani, T. (2010). Italiano regionale. In *Enciclopedia dell'italiano 2010.* Treccani, [http://www.treccani.it/enciclopedia/italianoregionale_(Enciclopedia_dell'Italiano)/#] [04/03/2014].

Rüegg, Robert (1956). *Zur Wortgeographie der italienischen Umgangssprache.* Köln: Kölner Romanistische Arbeiten.

Sobrero, A. (1996). *Introduzione all'italiano contemporaneo : la variazione e gli usi.* Roma: Laterza.

Sobrero, A., Miglietta, A. (2006). *Introduzione alla linguistica italiana.* Bari: Laterza.

Telmon, T. (1990). *Guida allo studio degli italiani regionali.* Alessandria: Edizioni dell'Orso.

TRECCANI = Treccani (2009). *Il vocabolario della lingua italiana.* Roma: Treccani.

VIVALDI: Kattenbusch, D. (1999): *VIVaio Acustico delle Lingue e dei Dialetti d'Italia.* [http://www2.hu-berlin.de/vivaldi].

Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples

Paul Cook[1], Michael Rundell[2], Jey Han Lau[3], and Timothy Baldwin[1]
[1]Department of Computing and Information Systems, The University of Melbourne
[2]Lexicography Masterclass and Macmillan Dictionaries
[3]Department of Philosophy, King's College London
paulcook@unimelb.edu.au, michael.rundell@lexmasterclass.com, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

There have been many recent efforts to automate or semi-automate parts of the process of compiling a dictionary, including building headword lists and identifying collocations. The result of these efforts has been both to make lexicographers' work more efficient, and to improve dictionaries by introducing more systematicity into the process of their construction. One task that has already been semi-automated is that of finding good dictionary examples, and a system for this, GDEX, is readily available in the Sketch Engine. An ideal system, however, would be able to automatically retrieve candidate examples of a particular sense of a word, which is beyond the current scope of GDEX. In this paper, as a step towards this ambitious goal, we propose and evaluate a method for applying a 'word-sense induction' system to automatically extract examples that exhibit a greater diversity of usages of a target word than can currently be obtained through GDEX. We then discuss the future prospects for systems that are able to automatically select candidate dictionary examples for a particular word sense.

Keywords: dictionary examples; word-sense induction; computational lexicography

1 Automating Lexicography

A major challenge facing contemporary lexicography – in the commercial sector, at least – is the goal of maximizing the potential of digital media and abundant language data, against a background of limited financial resources. As a response to this challenging climate, there has been significant progress in the automation of some of the tasks involved in compiling dictionaries and lexical databases – an approach which has the potential to deliver reduced development costs together with improved coverage of lexicographically-relevant facts (e.g., Rundell & Kilgarriff 2011, Cook et al. 2013, Kosem, Gantar & Krek 2013).

One component of this project is the automatic retrieval from corpus data of 'good' dictionary examples, whether for presenting editors with a shortlist of appropriate candidates, populating a dictionary database with examples, or providing the end-user with a range of instances of words in context (Kilgarriff et al. 2008; Kosem, Husák & McCarthy 2011). Notwithstanding these advances, there is scope for improvement in two areas. First, example-finding software does not yet routinely achieve the contextual diversity that characterizes example-sets selected by skilled lexicographers. Secondly, it does not attempt the difficult but critical task of mapping corpus instances onto dictionary senses. Some current dictionaries provide a range of (automatically-retrieved) examples to complement the manually-selected ones in the dictionary. This approach can be found, for example, in the 5th edition of the *Longman Dictionary of Contemporary English* (*LDOCE*) (http://ldoce.longmandictionariesonline.com/dict/SearchEntry.html), where users can opt to see up to ten 'Examples from the corpus', and in *Wordnik* (www.wordnik.com), where numerous authentic examples are shown on the right-hand side of the screen. Google Translate (http://translate.google.com) has a somewhat similar feature. But in none of these cases are the examples attached to word senses: in *LDOCE*, the entry for *pool* (noun) includes a random assortment of examples, including references to swimming pools, pools of investment funds, and even football pools. Ideally, we need a system which automatically selects optimally-diverse examples for a polysemous word (so that users are offered examples exhibiting the full contextual range of the word's behaviour) and matches the examples to the individual dictionary senses whose meaning they instantiate.

In this paper we report an experiment in which a 'word sense induction' methodology is applied to extracting corpus examples in a way that fulfills the first of these goals – identifying examples showing the diversity of contexts in which a word is used. We conclude by discussing the prospects for using the output of the word-sense induction system to map the 'induced senses' the system discovers in the corpus to dictionary senses.

2 Diversifying Example Sentences with Word-sense Induction

In this section we describe the GDEX method for identifying good dictionary examples and a recently-presented word-sense induction (WSI) system, and then propose a method to combine these two technologies to automatically select more-diverse dictionary examples.

2.1 GDEX

GDEX (Kilgarriff et al. 2008) is a system for automatically selecting good dictionary examples from a corpus. Sentences containing a given target word are scored based on a number of heuristics about what makes a sentence a good dictionary example, such as sentence length, the position of the target word in the sentence, and the other words occurring in the sentence. Given a query for a particular word, GDEX then returns the top-scoring sentences in a corpus, which can be manually examined for selection as examples. GDEX has become a standard lexicographical tool, and is available for use with many corpora in the Sketch Engine (SkE, http://www.sketchengine.co.uk/, Kilgarriff & Tugwell 2002). Adaptations of GDEX (Kosem, Gantar & Krek 2011) have incorporated simple notions of diversity to avoid selecting duplicate or very similar sentences. We propose a more sophisticated notion of diversity targeted at selecting a set of candidate example sentences exhibiting a wider range of senses of the target word.

2.2 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' (Cook et al. 2013). Crucially this grouping is done without reference to a pre-existing sense inventory. WSI is the automatic counterpart to the manual lexicographic process of word sense disambiguation (WSD, Atkins & Rundell 2008: 269).[1] Although the notion of word sense is of course controversial, it is nevertheless standard for dictionaries to carve up the meanings of a word into senses, even though dictionaries will vary in terms of the sense distinctions made for a given polysemous word.

Topic modeling (Blei, Ng & Jordan 2003) is a computational technique for automatically discovering latent structure in a corpus that has recently been successfully applied to a wide range of NLP tasks. A typical topic model automatically 'learns' the topics in a corpus, and the mixture of topics in each document in the corpus. Each topic is represented as a probability distribution over words; each document is represented as a probability distribution over topics.

Lau et al. (2012) present a WSI system based on topic modeling. Rather than building a topic model for an entire corpus, they build a separate model for each target word. In this model the "documents" are short contexts – typically 3 sentences – containing a usage of the target word. There is not necessarily a correspondence between the topics in a topic model and topics in the sense of the subject of a text (although a topic model for a corpus will often learn topics that do indeed correspond to this more common usage of *topic*). In Lau et al.'s WSI methodology, the topics in the topic model are interpreted as word senses.

In traditional topic models (e.g., latent Dirichlet allocation, Blei, Ng & Jordan 2003) the number of topics to be learned must be specified manually in advance. For WSI this would mean that the number of senses for each word would need to be set by hand. Words of course differ with respect to their polysemy, and an appropriate number of senses could only be determined based on corpus analysis. Lau et al. therefore use hierarchical Dirichlet process (Teh et al. 2006), a type of topic model that also automatically learns the appropriate number of topics for a given document collection.

[1] It is worth clarifying a terminological difference between the lexicographic and natural language processing (NLP) communities. In NLP, WSD refers specifically to the task of selecting the most appropriate sense, from a given sense inventory, for a given instance of a word in context. In lexicography WSD typically refers to identifying the various senses of a word, i.e., constructing a sense inventory, based on corpus evidence.

In Lau et al.'s model, each 'document' – typically consisting of a sentence including an instance of the target word, and one sentence of context before and after – is represented as a bag-of-words, i.e., word order is ignored, but the frequency of words in the context is maintained. Because the immediate context surrounding a word can be highly informative of that word's sense, positional word features that encode the specific three words occurring to the left and right of the target word are also included. In this document representation stopwords are ignored, and all other words are lemmatized. An example of the representation used is given in Table 1. For reasons of brevity the 'document' in this example is a single sentence (whereas it would typically be three sentences). An example of the senses induced by the model for the lemma *box* (n) is given in Table 2. Recall that each sense is a probability distribution over words; here the top-10 most likely terms for each sense are shown. Examining these terms allows us to roughly interpret the senses the system has induced. For example, senses 1 and 2 seem to correspond to usages of *box* in the context of sports and elections, respectively. Random samples of 5 corpus instances corresponding to induced senses 1 and 4 for *box* (n) are given in Table 3. For sense 1, all of the usages seem to correspond to an area of a sports ground, although the first four relate to soccer, but the last relates to tennis. For sense 4, the model has identified usages related to the entertainment industry, but includes a mixture of usages of the expressions *box office* and *box set*, as well as other usages such as the final example.

Target Lemma	**box (n)**
'Document'	The Flames had a two-man advantage near the end of the second period when Lacroix and McSorley were in the penalty **box** for kneeing and unsportsmanlike conduct, respectively.
Bag-of-words Features	flame, two-man, advantage, near, end, second, period, lacroix, mcsorley, penalty, knee, unsportsmanlike, conduct, respectively
Positional Word Features	lacroix_#-3, mcsorley_#-2, penalty_#-1, knee_#+1, unsportsmanlike_#+2, conduct_#+3

Table 1: An example of the topic model features.

Sense Number	Top-10 Terms
1	box minute @card@ game ball goal score shot penalty play
2	box @card@ ballot ballot_#-1 police official election vote find party
3	box @card@ company computer digital cable converter black_#-1 black converter_#-1
4	office box office_#+1 million @card@ film movie weekend dollar ticket
5	box @card@ n't look find small think put room store
6	box cereal recipe cut outlet post fat gram vip wire
7	shanker teller lionel penn crush ferry jesus julie maine marshal

Table 2: The top-10 terms for each of the senses induced for the lemma *box* (n).

Sense Number	Usage
1	Bolivia added an extra insurance goal in the 80th when Ronaldo Garcia sent a long blast from outside the **box** into the upper corner. Arsenal were left nursing a justifiable grievance over the referee's failure to award a second-half penalty when Kuyt unbalanced Alexander Hleb with a tug from behind as the midfielder wriggled into space deep in the **box**. Thiago made it 3-0 in the 79th minute with a powerful left foot drive from the edge of the **box**. Masami Ihari headed away a corner from Juninho but only as far as Zinho, who let fly with a spectacular volley from the edge of the **box** that gave no chance at all. He's placing it in the **box** beautifully, hitting closer to the line with lots of aces.
4	And while Joel is paying himself a backhanded compliment, especially in the context of B-sides, live tracks and rarities **box** set, it got me thinking. It's the harsh command of the **box** office, demanding a big seller and the heck with all else. "For the most part, it'll help us," says Bobbie Welch, **box** office manager for New Mexico State University, which has hosted ZZ Top, Guns N' Roses and Paul McCartney. For those seeking more kid-friendly fare, the "Spotlight Collection" discs – including a sixth released Tuesday ($27) – offer more than 30 family-appropriate installments from the Golden **box** sets. Only 17 percent of reviews were positive, according to RottenTomatoes.com, but 82 percent of audience survey respondents checked off the "excellent" or "very good" **boxes**, according to Sony.

Table 3: Usages corresponding to induced senses 1 and 4 of the lemma *box* (n).

Lau, Cook & Baldwin (2013a,b) recently showed this WSI methodology to be the overall best performing system on two recent SemEval WSI shared tasks (Jurgens & Klapaftis 2013; Navigli & Vannella 2013). Cook et al. (2013) demonstrated that this system can be applied as a lexicographical tool for finding new word-senses. We therefore adopt this WSI system here for diversifying automatically-selected dictionary examples.

2.3 Diversification

For a given target lemma, we obtain the top-100 GDEX examples for a corpus from SkE. We further obtain a random sample of up to 50k usages of the target from the same corpus. In each case we extract the sentence containing the usage of the target, and one sentence of context on either side. Following Lau et al. (2012) we remove stopwords and lemmatize the tokens in the context. We then run the WSI system on these usages of the target lemma. The WSI system outputs a label indicating the induced sense number of each target instance. These induced senses correspond to groups of usages that exhibit the same sense, according to the WSI system, not dictionary senses.

We then use these induced sense labels to diversify the top-100 GDEX examples. To do so, we repeatedly iterate through the top-100 GDEX sentences (i.e., we consider each sentence in turn, one by one). For each pass over the sentences, we select the best GDEX sentence (according to GDEX's ranking) for each induced sense, which has not been selected in a previous pass. We repeat this until all sentences have been selected. In the subsequent analysis we compare the top-5 GDEX examples to the top-5 examples produced by this diversification procedure.

3 Analysis

For this preliminary analysis we selected 98 target lemmas to analyse: 54 from a recent SemEval WSI task (Jurgens & Klapaftis 2013) and 44 additional medium-polysemy lemmas. We extracted GDEX sentences, and the additional randomly-selected usages, from the ukWaC (Ferraresi et al. 2008).

We ran the GDEX diversification procedure described in the previous section for each target lemma. The top-5 GDEX sentences for the target lemma exhibited varying numbers of induced senses (as determined by the WSI software). However, if the top-5 GDEX sentences already exhibit many (e.g., 4 or 5) induced senses, then our diversification procedure has little or no impact. We therefore focused our analysis on lemmas where the top-5 GDEX usages exhibited less diversity. Crucially such cases can easily be automatically identified. Here we discuss the findings for twelve lemmas whose top-5 GDEX sentences were the least diverse, exhibiting just two induced senses of the target lemma in each case. (In no case did the top-5 GDEX usages exhibit just one induced sense.)

For each lemma, two sets of five example sentences were prepared: (1) the top-5 GDEX sentences, and (2) the top-5 sentences from our diversification procedure. These sets of sentences were presented to a professional lexicographer (the second author of this paper) who was asked to judge which set of examples was better. Crucially the lexicographer did not know which method the sets of examples corresponded to. For eight lemmas the examples produced through our new diversification procedure were selected as better; in the remaining four cases the default GDEX examples were chosen. To give an idea of the potential of our method, we discuss the output of the two systems for two of the target lemmas which were analysed: *exploitation* and *bitter*.

Top-5 sentences from GDEX:

(1) It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.

(2) A new minister replaces the old one, the daily grind of **exploitation** resumes.

(3) And Engage, wittingly or not, is aiding it in this **exploitation**.

(4) It's not trade we're against, it's **exploitation** and unchecked power.

(5) Others have seen Napster as little more than payback for decades of record company **exploitation** of artists.

Top-5 sentences from our new diversification approach:

(6) It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.

(7) Others have seen Napster as little more than payback for decades of record company **exploitation** of artists.

(8) Delivery will focus upon the development and **exploitation** of repertoire through the workplace.

(9) Requirement 24 - Identify all auditable events that may be used in exploitation of known covert storage channels.

(10) Workers can refer young people they think are vulnerable or at risk of homelessness and sexual **exploitation**.

The examples that are shared by the two sets are both good. However, the second and third sentences from default GDEX are weak because they offer little context for interpretation and include anaphora. Although the final sentence for default GDEX (sentence 4) is acceptable, the diversified sentences provide a better snapshot of the word overall, in that they cover the more neutral sense of exploitation (sentence 4) and include an example of *sexual exploitation* (sentence 5).

Top-5 sentences from GDEX:

(1) If cows eat too many carrots, their milk tastes **bitter**.

(2) It's so easy for us to be **bitter**.

(3) Men obeyed their base immediate motives until the world grew unendurably **bitter**.

(4) In his destroyed Urim, its lament is **bitter**.

(5) No; and henceforth I can never trust his word. **Bitter**, bitter confession!

Top-5 sentences from our new diversification approach:

(1) If cows eat too many carrots, their milk tastes **bitter**.

(2) It's so easy for us to be **bitter**.

(3) In his destroyed Urim, its lament is **bitter**.

(4) The rivalry between Soka Gakkei and Aum Shinrikyo was **bitter**.

(5) Joe Frazier never felt more **bitter** about defeat, and continues even today to hate his great rival.

In this case, three of the examples appear in both sets, and their quality varies: the first in each set is good; the second acceptable, if a little short on context; the third (*In his destroyed Urim…*) is weak, and would certainly not make an appropriate example sentence for a pedagogical dictionary (or any other dictionary, for that matter). But if we compare the two examples unique to each set, those in the second set (sentences 4 and 5) are clearly more suitable as dictionary examples than those in the first (3 and 5). *Rivalry* has a high saliency score as a collocate of *bitter* (though *disappointment* would have been even better, assuming the goal is maximum typicality). And the addition of a collocating preposition in the last example (with *about*, the most frequent preposition appearing with *bitter*) provides additional diversity.

4 Discussion

Software for selecting dictionary examples could be improved if it were optimised to select diverse examples for a polysemous word, such that the examples show the full range of usage for that word. In this paper we have proposed a novel method for automatically selecting a more diverse set of dictionary examples from a corpus than can currently be obtained using GDEX. We carried out a small-scale preliminary evaluation of this method, and found that – in terms of diversity – our approach outperformed GDEX for eight out of twelve lemmas analyzed. The results are encouraging rather than conclusive. But with further improvements based on what we have learned through this experiment, this new method could be applied to real lexicographic tasks – either for providing editors with candidate lists from which to select examples for a dictionary, or for automatically providing dictionary users with additional examples. In either case, the outcome should be a set of examples exhibiting a more diverse range of usages than current software tools usually supply.

Systems for selecting examples would be further improved if they were able to match automatically-identified examples to corresponding dictionary senses. Lau et al. (2014) recently proposed a method to link the senses induced by the same WSI system used here to senses in a dictionary. They evaluated this method in the context of identifying the relative frequencies of the senses of a given word in a corpus, and showed it to perform comparably to previously-proposed approaches for this task (McCarthy et al. 2007).[2] In future work, we intend to combine this method for linking induced senses to dictionary senses with the approach described in this paper, in order to identify good examples at the level of *word senses*, as opposed to *lemmas*. WSI remains a very difficult task for current natural language processing technologies. Crucially, in the context of identifying sense-specific dictionary examples, it might not be necessary to correctly identify the dictionary sense of every corpus instance of a target word. Instead, it might suffice to identify the dictionary sense corresponding to

[2] Very interestingly, but of less relevance to the present paper, they also showed that this method has the potential to identify dictionary senses that are unattested in a corpus, and senses that are induced by the WSI system but not listed in a dictionary.

those instances where the system is highly confident of its prediction, and to then apply GDEX to select good dictionary examples amongst those instances. We are therefore optimistic about the future possibility of automatically adding sense specific examples to dictionaries, although much work remains to be done.

The WSI system of Lau et al. (2012) lies at the core of the method presented in this paper. To encourage further research on WSI and its applications, Lau, Cook & Baldwin (2013a,b) made this system publicly available under a license which permits its use for commercial purposes (https://github.com/jhlau/hdp-wsi). We hope that others will make use of this software to consider further applications of topic modeling and WSI in computational lexicography.

5 References

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

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cook, P., Lau, J.H., Rundell, M, McCarthy, D. & Baldwin, T. (2013). 'A lexicographic appraisal of an automatic approach for detecting new word-senses', in Kosem et al. 2013: 49-65.

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.

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. 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.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008) 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus', in Bernal, E. and DeCesaris, J. (Eds) Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra: 425-433.

Kosem, I., Gantar, P., & Krek, S. (2013). 'Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing', in Kosem et al. 2013: 32-48.

Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) (2013). Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Kosem, I., Husák, M., & McCarthy, D. (2011). 'GDEX for Slovene'. In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011.* Ljubljana: Trojina, Institute for Applied Slovene Studies: 151-159.

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 European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA.

McCarthy, D., Koeling, R., Weeds, J., and Carroll J., (2007). Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics,* 33(4):553–590.

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.

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

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.

Acknowledgements

This work was supported in part by the Australian Research Council.

Cross-linking Austrian dialectal Dictionaries through formalized Meanings

Thierry Declerck, Eveline Wandl-Vogt
DFKI GmbH, Language Technology Lab, Germany;
Austrian Academy of Sciences, ICLTT, Austria
declerck@dfki.de, eveline.wandl-vogt@oeaw.ac.at

Abstract

This paper deals with the formalization of aspects of definitions used in dialectal dictionaries. We focus on the way "meanings" are encoded in such dictionaries, an essential position for many users and lexicographers, and describe how such an encoding can be re-used for cross-linking entries of on-line (dialectal) dictionaries. In this contribution we describe in some details experiments we made in this respect with two Austrian dialectal dictionaries: The dictionary of Bavarian dialects of Austria ("Wörterbuch der bairischen Mundarten in Österreich", WBÖ) and the dictionary of the Viennese dialect ("Wörterbuch der Wiener Mundart", WWM), which we ported into the SKOS and *lemon* models in order to publish them in the Linguistic Linked Open Data cloud. We show how this approach is not only appropriate for supporting the automation of the cross-linking of dialectal dictionaries, but also for linking entries of the dialectal dictionaries to other types of lexical and encyclopaedic resources in the web

Keywords: Dialectal lexicography; Semantic Web; Linguistic Linked Open Data; Austrian dialects

1 Introduction

In the context of recent work dedicated to porting the dictionary of Bavarian dialects of Austria ("Wörterbuch der bairischen Mundarten in Österreich", WBÖ)[1] and the dictionary of the Viennese dialect ("Wörterbuch der Wiener Mundart", WWM)[2] onto representation formats supporting their publication in the Linked Open Data (LOD) framework[3], and more specifically in the Linguistic Linked Open Data cloud[4], we got our attention directed to the investigation on how this approach could support an automation of the cross-linking of such dialectal language resources. For this, we focused on the way "meanings" are encoded in the selected dictionaries, an essential position for many users and

[1] See http://www.oeaw.ac.at/dinamlex/WBOE.html and (Wandl-Vogt 2005; Wandl-Vogt 2008). See also (Declerck & Wandl-Vogt 2013) for a description of the approach adopted for porting WBÖ to the SKOS representation language.

[2] See (Hornung & Grüner 2002).

[3] See http://linkeddata.org/

[4] See http://linguistics.okfn.org/resources/llod/

lexicographers. We take advantage here of a property of dialectal dictionaries concerning the expression of meanings of entries: Although conceived as monolingual reference work, dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is Standard German and Austrian German, as can be seen for example in the WBÖ entry "Puss": one of its meanings is expressed by the Standard German word "Kuß" (*kiss*) and by the Austrian German word "Busserl" (WBÖ: 1,1515)[5], as can be seen in Figure 1 below. WWM uses also the Standard German word "Kuss" for expressing a meaning of the entry "Bussal", as can be seen in Figure 2. Our assumption is thus that linking entries in distinct dialectal dictionaries can be implemented on the base of meanings that are expressed by similar means across the dictionaries.

In this paper we first briefly describe the two dialectal dictionaries we have been considering for our experiments. We then depict the processes we applied to the entries of the dictionaries for extracting and analyzing the expressions that express their meanings, and their encoding in the representation languages RDF[6], SKOS-XL[7] and *lemon*[8] for supporting their publication in the LOD.

The ultimate goal of our work is not only to be able to cross-link the lexical resources described in this paper, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way our dialectal and historical lexical resources to the same "digital dignity" as the mainstream languages have already gained.

We show also how this encoding allows enriching our lexical data with additional lexical information, mainly senses and multilingual variants.

2 The selected dialectal Dictionaries WBÖ and WWM

We describe in this section briefly the main characteristics of the two dialectal dictionaries we selected for conduction our experiments on cross-linking.

2.1 The Architecture of the selected dialectal Dictionaries in a Nutshell

The chosen dialectal dictionaries, WBÖ as well as WWM, are scientific dictionaries. Each dictionary offers references and example sentences for illustrating contexts of use for the entries. Whereas the documentation and interpretation in WBÖ is exhaustive, WWM is much shorter and comprehensive. WBÖ is a dialectal dictionary for all Bavarian dialects in the former Austrian Hungarian Monarchy (status: about 1915), whereas WWM is a dictionary for the dialect of the city and county of Vienna.

[5] See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language

[6] See http://www.w3.org/RDF/

[7] See http://www.w3.org/TR/skos-reference/skos-xl.html

[8] See (McCrae & al., 2012).

- grammar: Every entry informs about grammatical properties of the word
- etymology: Every entry contains information about the etymology of the word.
- definition(s): Definitions are a central position in both (onomasiological) dictionaries. Complementing the definitions, WBÖ presents a lot of examples of spoken and written dialect, phrases, songs and poems. Due to the fact, that approximately 10% of the material consists of excerpts of written texts and that the main aim in the beginning was to document the development of a word from its beginning to the actual dialect (see Arbeitsplan 1912) the emphasis on written texts is very high. Furthermore, WBÖ definitions often include a lot of encyclopaedic information about rural traditions and traditional customs.

  WWM presents this type of semantic information in a much more concise way.

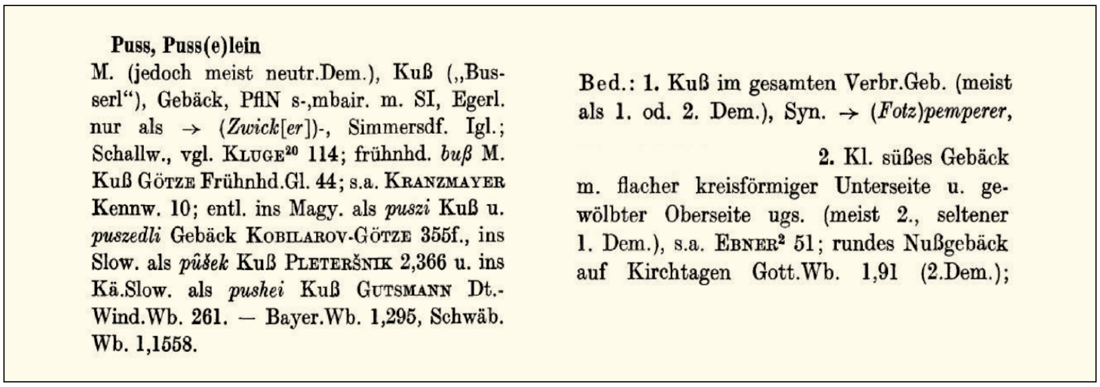
  - meanings, as a core part of definitions: Although conceived as monolingual reference work, many dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is Standard German, sometimes accompanied by Austrian German. So for example the meaning of the WBÖ entry “Puss” is expressed by the Standard German word “Kuß” (*kiss*) and by the Austrian German word “Busserl” (WBÖ 1,1515)[9], as can be seen also in Figure 1. As the reader will see later, we take advantage of this property of dialectal dictionaries concerning the expression of meanings of entries: Linking entries in distinct dialectal dictionaries can be implemented on the base of the Standard German expressions of meanings that are shared across those dictionaries.
- references to dictionaries of adjacent German dialects: WBÖ puts the information presented into a whole dialectal language area in quoting the neighbouring German dialectal dictionaries. WMM does not include this position due to not being embedded into the same / similar methodological background.
- phonetics: Phonetics played an important role in the so called “Junggrammatische Schule”[10], which is the methodological background for the WBÖ. WWM offers headwords that are transliteration based on phonetics.
- compounds: Compounds are treated within the base word entry, e.g. “Nuss)puss” (the famous creamy hazelnut truffle “Nussbusserl”). They might be dealt with in the position ‘Komp.’ (*Compounds*) within the main entry (as this is the case for the entries *Puss* (WBÖ) or *Bussal* (WWM)).
- cross references to derivations and related words: There is a position, where cross references to derivations and related words are stored, e.g. *Syn.* → *(Fotzen)pemperer* (WBÖ); *(Syn.: Schmåtss*

[9] See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language.
[10] See http://en.wikipedia.org/wiki/Neogrammarian for more details.

(WWM). Including articles of derivations or related words, especially those with "less information value" due to a rationalisation concept for WBÖ (see Straffungskonzept 1998: §§ 1.2.1-1.2.3).

- editor: Finally, each WBÖ entry closes with the initial of the author, e.g. *W.B. = Werner Bauer*; in the WWM similar signing is not existing.

In the partial entry for the word "Puss" in WBÖ shown in Figure 1, the reader can see in the right column the details for two selected meanings expressed in Standard German: 1) "Kuss" (*Kiss*) and 2) "Kl. süßes Gebäck" (*small sweet pastry*).



Puss, Puss(e)lein

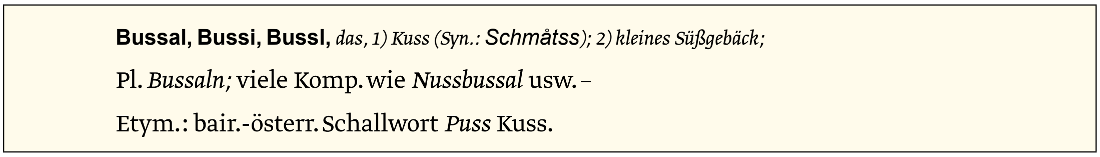
M. (jedoch meist neutr.Dem.), Kuß („Busserl"), Gebäck, PflN s-,mbair. m. SI, Egerl. nur als → (*Zwick*[*er*])-, Simmersdf. Igl.; Schallw., vgl. KLUGE^{20} 114; frühnhd. *buß* M. Kuß GÖTZE Frühnhd.Gl. 44; s.a. KRANZMAYER Kennw. 10; entl. ins Magy. als *puszi* Kuß u. *puszedli* Gebäck KOBILAROV-GÖTZE 355f., ins Slow. als *pûšek* Kuß PLETERŠNIK 2,366 u. ins Kä.Slow. als *pushei* Kuß GUTSMANN Dt.-Wind.Wb. 261. — Bayer.Wb. 1,295, Schwäb. Wb. 1,1558.

Bed.: **1.** Kuß im gesamten Verbr.Geb. (meist als 1. od. 2. Dem.), Syn. → (*Fotz*)*pemperer*,

2. Kl. süßes Gebäck m. flacher kreisförmiger Unterseite u. gewölbter Oberseite ugs. (meist 2., seltener 1. Dem.), s.a. EBNER^{2} 51; rundes Nußgebäck auf Kirchtagen Gott.Wb. 1,91 (2.Dem.);

Figure 1: WBÖ 3,1515f.: Puss, Puss(e)lein.

In the second example, taken from the WWM and displayed in Figure 2, the reader can see that very similar meanings are provided for the entry "Bussal, Bussi, Bussl". While the first meaning is expressed by using exactly the same Standard German word "Kuss" in both cases, the second meaning (*small sweet pastry*) is expressed in each dictionary by using variants: "Kl. süßes Gebäck" vs. "kleines Süßgebäck".



Bussal, Bussi, Bussl, *das, 1) Kuss (Syn.: Schmåtss); 2) kleines Süßgebäck;*

Pl. *Bussaln;* viele Komp. wie *Nussbussal* usw. –

Etym.: bair.-österr. Schallwort *Puss* Kuss.

Figure 2: WWM 199: Bussal, Bussi, Bussl.

2.2 Access Structure

The main access structure, for both WBÖ and WWM, is the macrostructure, namely the headword.[11]

- WBÖ has chosen due to etymologic-historic considerations a sophisticated, artificial headword, which is difficult to be used as access structure by scientists and in particular causes problems for laypersons. As an example, the German headword "deutsch" (*german*) is represented in WBÖ as "teütsch"; the Standard German headword "Pflaumenbaum" and the Standard Austrian word "Zwetschkenbaum" (*plum tree*) are represented in the WBÖ as "Zwëtschken)pāum". And a subentry of the main entry "Pāum" (*tree*), the WBÖ headword "Busserl", lacks a standard German representation.
- WWM chooses a transliterated headword, based on phonetics.

So that a cross-referencing and interlinking of dialectal dictionaries, even within the same language area (here: Bavarian variants), does not work without the development of mapping rules. Furthermore, such a mapping would offer just a flat and non-hierarchical interlinking.
This situation motivated the main approach of the work presented here, which consists in investigating if individual word senses, the meanings of entries expressed in Standard and Austrian German words, can serve as an access point for the cross-linking of our two selected dialectal dictionaries, as well as reference point for linking to lexical resources and other knowledge sources in the Web. The following section describes the processes we applied to the entries of the dialectal dictionaries for extracting and analyzing the expressions that express their meanings

3 Extraction and Linguistic Analysis of Expressions introducing the Meanings of Entries

Our first task consisted in detecting and extracting automatically from both dictionaries the strings expressing the core meanings for each entry. Fortunately both dictionaries have been made available to us in an electronic version: WBÖ in a proprietary XML schema and WWM in Microsoft Word. We used the TEI "OxGarage"[12] service to convert the WWM Word document into a TEI compliant XML representation. In both XML representations it was straightforward to describe in Perl scripts the patterns for extracting the meanings of the entries expressed in Standard or Austrian German.
But as mentioned at the end of section 2.1, there are discrepancies in the use of Standard or Austrian German word forms across the dictionaries, so that it is often not possible to establish a relation bet-

[11] Other important positions, navigation and access structures within (dialectal) dictionaries are – or could be – space and time. Geo-referencing with time-stamp stored within a GIS offers possibilities for spatio-temporal visualisation as well as analysis (Wandl-Vogt 2010) and exploratory visually conducted analysis (Theron & Wandl-Vogt 2014).

[12] See http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/

ween words expressing meanings across the dialectal dictionaries. Since pure string matching cannot provide accurate comparisons between those expressions, there is the need to apply basic natural language processing to the expressions and to reduce those to their lemmatized form and to mark them up with part-of-speech and morphological information. The comparison of expressions standing for meanings in both dictionaries is then made on the base of such linguistic information associated to the strings. We provided an automatic linguistic analysis of those extracted meanings, using lexical and syntactic analysis grammars written within the SCHUG tools (Declerck 2002). This included tokenization, lemmatisation, Part-of-Speech (POS) tagging and constituency as well as dependency analysis. The strings marking in both dictionaries the "small sweet pastry" meaning are enriched with the following linguistic features:

(1) WBÖ: (NP süßes (ADJ, lemma = süß, MOD) Gebäck (N, lemma = Gebäck, HEAD)) - *sweet pastry*

(2) WWM: (NP (kleines (ADJ, lemma = klein, MOD) Süßgebäck (N, compound: süß (ADJ, lemma = süß, MOD) + Gebäck (N, lemma = Gebäck, HEAD)), HEAD)) - *small sweet pastry*

In the examples (1) and (2), we can see the distinct serializations of similar concepts in German. The second example uses a compound noun ("Süßgebäck"), which has the same meaning as the simple nominal phrase in the first example ("süßes Gebäck"). In order to automatically establish this similarity, it is necessary to first perform a morphological decomposition of the head noun in the second example. And we need the lemma of the adjective in the first example, in order to be compared with the first element of the compound noun in the second example.

The fact, that both linguistically analyzed strings expressing the meanings share the same lemmas for adjectival modifiers and head nouns is the base for cross-linking the entries. As we want to formally express this relation, we need to use an appropriate representation language, opting here for Semantic Web standards (e.g. compatible to RDF), also in order to be able to publish our data in the Linked Data cloud.

4 Porting the dictionary data into the Linked Open Data framework

To mark linguistically analyzed meanings as related, it is requested to use semantic web representation languages, like those developed in the context of W3C[13] standardization activities: RDF[14], SKOS, SKOS-XL[15], *lemon*[16]. With this step we want to benefit from the inherent linking (and merging) possibilities offered by Semantic Web languages used in the Linked Data framework, and more specifically

[13] See http://www.w3.org/.

[14] See http://de.wikipedia.org/wiki/Resource_Description_Framework.

[15] See http://www.w3.org/2004/02/skos/ and http://www.w3.org/TR/skos-reference/skos-xl.html respectively.

[16] See (McCrae & al., 2012).

we aim at contributing to the emerging Linguistic Linked Open Data cloud[17], integrating dialectal language data into this framework.

4.1 Porting the dictionaries into SKOS

Based on the Resource Description Framework (RDF), SKOS (Simple Knowledge Organization System):

> provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary.[18]

Our experiment with SKOS is thus kind of novel, since we apply it to dictionaries, although one can for sure consider dictionaries as being very close to thesauri and in our approach we encode elements of entries (basically the meanings) of the dictionaries as concepts being part of a conceptual scheme. We chose this representation language, since:

- SKOS concepts can be "semantically related to each other in informal hierarchies and association networks"[19]
- "the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice"[20]
- SKOS "can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools."[21]

With the use of SKOS (and RDF), we are also in the position to make our dictionary resources compatible with other language resource available in the LOD cloud. Examples of such resources are the actual DBpedia instantiation of Wiktionary[22] or version 2.0 of BabelNet[23].

We decided in the most recent version of our model to encode the strings standing for introducing each entry of a dictionary as a skos:Concept being a member of a skos:Collection, while each associated sense is encoded as skos:Concept that is part of a concept: Scheme. In the first case we deal with a flat list of elements, while in the second case we can model (hierarchical) relations between the meanings (also called "senses"). In the following pages of this section, we present some examples of our model applied to WBÖ, using the so-called turtle serialization.

[17] See http://linguistics.okfn.org/resources/llod/
[18] http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/
[19] Ibid.
[20] Ibid.
[21] Ibid.
[22] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.
[23] http://babelnet.org/

```
(3)	icltt:Dictionary
		rdf:type owl:Class ;
		rdfs:comment “Modeling the ICLTT dictionaries”@en ;
		rdfs:label „Wörterbuch“@de , „Dictionary“@en ;
		rdfs:subClassOf owl:Thing .
```

We first introduce (ex:3) a “Dictionary” class, of which the WBÖ dictionary is an instance of (ex:4).

```
(4)	icltt:wboe
		rdf:type icltt:Dictionary , skos:Collection ;
		rdfs:comment “OEAW Dictionary for Bavarian”@en ;
		rdfs:label „Wörterbuch der bairischen Mundarten in Österreich“@de , „Bavarian dialects of Austria“@en ;
		icltt:hasLanguage icltt:bar ;
		skos:member icltt:concept_puss .
```

As the dialectal dictionaries are encoded as skos:Collection, entries of the dictionaries are modelled as being a skos:member of such collections. The WBÖ entry “Puss” (encoded in ex:4 as the icltt:concept_puss) is therefore listed as a member (for reason of space we display here only one member of the collection). The icltt:concept_puss is introduced in our model as an instance of the class icltt:Entry (ex:5):

```
(5)	icltt:concept_puss
		rdf:type icltt:Entry ;
		rdfs:label “puss”^^xsd:string .
```

```
(6)	icltt:Entry
		rdf:type owl:Class ;
		rdfs:label “Entry”^^xsd:string ;
		rdfs:subClassOf skos:Concept .
```

In doing this we have introduced entries of the WBÖ as a concept being a member of a collection. We still need to introduce in our model the concrete information about the entries, and we describe this step in the next section.

4.2 Representing the Headwords and the Meanings in SKOS-XL and lemon

Contrary to most knowledge objects described in the LOD, we do not considers strings (encoding in WBÖ lemma and word forms as part of the language) as being just literals, but as being also knowledge objects. We considered therfore the use of SKOS-XL and of the *lemon* model[24] for representing the string used for headwords and senses. SKOS-XL has proven to be adequate for encoding strings as complex knowledge objects, but not for representing the linguistically analyzed expressions used for marking the meanings. For this we opted for the *lemon* model, which is compatible to SKOS.

```
(7)  icltt:concept_puss
                 rdf:type icltt:Entry ;
                 rdfs:label “puss”^^xsd:string ;
                 skosxl:prefLabel icltt:entry_puss .
```

In the displayed code in ex:7, the reader can see now the complete representation of the “icltt:concept_puss” object (extending ex:5): the concrete headword in the dictionary is pointed to by the means of the skox-xl property “prefLabel”, which contrary to both the rdfs:label and skos:label properties is not having a literal as the possible value of the range of the property, but which is having an object (“icltt:entry_puss”, as shown in ex:7) as a value in its range:

```
(8)  icltt:entry_puss
                 rdf:type icltt:Lemma ;
                 icltt:hasPos icltt:noun ;
                 lemon:sense icltt:gebäck , icltt:kuss , icltt:süßes_gebäck ;
                 skosxl:literalForm “Puss”@bar .
```

As the reader can observe in ex:8, we can encode the fact that “puss” is a lemma and that it is a noun. More importantly, we can include the senses associated to the WBÖ entry. In fact we are adding a new sense, “icltt:gebäck” (*pastry*). This is a direct consequence of the linguistic analysis of the expression “süßes Gebäck” we described in section 3 (ex:1 and ex:2): since the word “Gebäck” is the head noun in this nominal phrase, we can assume that this head noun is also a meaning (or sense) to be associated to the entry. In doing so we introduce a hierarchical organization of the meanings associated to an entry: “süßes Gebäck” is a specialization of “Gebäck” (using the “skos:broader” relation, as shown in ex:16) . As mentioned above, this is the reason why we use the skos:ConceptScheme construct in order to encode the senses associated to the entries (ex:9-ex:11):

[24] *lemon* is also available as an ontology: http://lemon-model.net/

```
(9)   icltt:Senses_ICLTT
                    rdf:type skos:ConceptScheme ;
                    rdfs:comment “Senses that are used in ICLTT dictionnaries”@en ;
                    rdfs:label „Senses“@en , “Bedeuttungen”@de .
```

```
(10) icltt:Sense
rdf:type owl:Class ;
                    rdfs:label “Sense”@en ;
                    rdfs:subClassOf skos:Concept ;
                    owl:equivalentClass lemon:LexicalSense .
```

```
(11) icltt:kuss
                    rdf:type icltt:Sense ;
                    rdfs:label “kiss”@en , “Kuss”@de ;
                    skos:inScheme icltt:Senses_ICLTT ;
                    skosxl:prefLabel icltt:sense_kuss .
```

Lemmas of the expressions used in the original dictionaries for marking meanings are indirectly linked to the class Sense, and are directly instances of the class icltt:Lemma (ex:12-ex:14):

```
(12) icltt:sense_gebäck
                    rdf:type icltt:Lemma ;
                    rdfs:label “Gebäck”@de ;
                    icltt:hasPos icltt:noun ;
                    skosxl:literalForm “Gebäck”@de .
```

```
(13) icltt:sense_kuss
                    rdf:type icltt:Lemma ;
                    rdfs:label “Kuss”@de ;
                    icltt:hasPos icltt:noun ;
                    skosxl:literalForm “Kuss”@de .
```

```
(14) icltt:sense_süß
                    rdf:type icltt:Lemma ;
                    rdfs:label “süß”@de ;
                    icltt:hasPos icltt:adj ;
                    skosxl:literalForm “süß”@de .
```

We introduce in ex:14 the class "CompoundSense" that allows us to mark the fact that the sense(s) can be resulting from a compound term or a phrase used to express the meaning of an entry.

```
(15) icltt:CompoundSense
                rdf:type owl:Class ;
                rdfs:label "Composition of Sense"@en ;
                rdfs:subClassOf icltt:Sense .
```

An instance of such a class is displayed in ex:16, in which the reader can see how we model for the time being the hierarchical relation between the sense "gebäck" and "süßes Gebäck" (using the "skos:broader" relation). We can also encode the fact that the sense of the entry "süß_gebäck" is composed of two senses, but the model needs to be further developed, since it is clear that the sense "süß" cannot be considered only as a sub-sense of "süß_gebäck", but more as a "modifying" sense. We are currently working on representing with the help of *lemon* such cases of linguistic dependencies.

```
(16) icltt:süß_gebäck
                rdf:type icltt:CompoundSense , lemon:LexicalSense ;
                rdfs:label „sweet pastry“@en , „süßes Gebäck“@de ;
                lemon:subsense icltt:süß , icltt:gebäck ;
                skos:broader icltt:gebäck ;
                skos:inScheme icltt:Senses_ICLTT .
```

In ex:16 we can see the advantage of using a representation model that can encode linguistic properties. In this case, it is for example necessary to tokenize the string representing the meaning of the entry "Puss": the first token can then be lemmatized to "süß" (*sweet*), while for the second token the lemma is identical to the written form used. We represent the tokenization information using the *lemon* property "decomposition", as can be seen in ex:17:

```
(17)  lemon:decomposition
                rdfs:domain lemon:LexicalSense ;
                rdfs:range icltt:Sense .
```

For the time being we introduce in our model an explicit listing of components as subclasses of icltt:-CompoundSense (see ex:18 and ex:19). The way this encoding is used is shown in ex:21.

```
(18) icltt:Component1
rdf:type owl:Class ;
                rdfs:label ""^^xsd:string ;
                rdfs:subClassOf icltt:CompoundSense .
```

```
(19) icltt:Component2
                  rdf:type owl:Class ;
                  rdfs:label “”^^xsd:string ;
                  rdfs:subClassOf icltt:CompoundSense .
```

4.3 Linking to Resources available in the LOD

As the reader can see in the examples (20) and (21) further below, we decided to use the DBpedia instantiation of Wiktionary as a reference for the senses (meanings) of the entries of the dictionary, pointing thus to linguistic and knowledge objects that are already in the LOD. To be more precise, the link to DBpedia/Wiktionary is applied for each token of the expressed meanings. In the case of “süßes Gebäck”, we can thus point to two URLs in DBpedia/Wiktionary, each representing the adequate senses for the actual token. In the name of the URLs used for pointing to DBpedia/Wiktionary we have implicitly also the information about the language and the PoS of the entry. But one could point to the RDF version of ISO data categories[25] for making this information explicit in our model.

Additionally the linking to the appropriate senses in DBpedia/Wiktionary allows accessing all the corresponding multilingual lemmas associated in this resource with a sense. Looking for example at http://wiktionary.dbpedia.org/page/sweet-English-Adjective-1en (corresponing to the URL for the German word, we us in ex:20), we get more than 70 expressions in more than 60 languages.

And the URL http://wiktionary.dbpedia.org/page/pastry-English-Noun-1en refers to ca. 30 expressions in about 25 languages. Having one unique URL for the “sweet” sense of “süß” allows to link all the corresponding entries to a unique reference point, and so to improve comparability of dictionary resources, also at the semantic level.

```
(20) icltt:süß
                  rdf:type lemon:LexicalSense , icltt:Component1 ;
                  rdfs:label „sweet“@en , „süß“@de ;
                  skos:exactMatch <http://wiktionary.dbpedia.org/resource/süß-German-Adjective-1de> ;
                  skos:inScheme icltt:Senses_ICLTT .
```

```
(21) icltt:gebäck
                  rdf:type lemon:LexicalSense , icltt:Component2 ;
                  rdfs:label „Gebäck“@de , „pastry“@en ;
                  skos:exactMatch <http://wiktionary.dbpedia.org/resource/Gebäck-German-Noun-1de> ;
                  skos:inScheme icltt:Senses_ICLTT ;
                  skos:narrower icltt:süß_gebäck ;
                  skosxl:prefLabel icltt:sense_gebäck .
```

[25] See http://www.isocat.org/

In the two examples just above, the reader can see how we can link the senses of the entries to existing sources in the LOD. We use for this the skos:exactMatch property (although we could also use *lemon* properties for this). But our model also makes clear that the DBpedia/Wiktionary URL we use for each token is valid only in the context of the compound term we are dealing with. The word "süß" has in DBpedia/Wiktionary more senses, but in the context of "süßes Gebäck" only the one sense that refers to "sweet" is adequate. Using the lemon model allows us thus to disambiguate senses associated to the components of complex terms used in the dictionaries for expressing a meaning.

5 Cross-referencing of Dictionary Entries through shared Meanings

The establishment of a relation between the entry "Puss" in WBÖ and and the entry "Bussal" in WWM is made possible on the base of the successful mapping of both the adjectival modifier "süß" and the head noun "Gebäck", which are present in both the definitions in WBÖ and WWM, but used in the context of textual variants, as can be seen in the examples (1) and (2). Interesting is also the fact that a user searching the electronic version of the dictionaries could give the High German form "Gebäck" and would get from both dictionaries all the entries which have this word in their definition, also if the word is used in a compound form. The same for the High German adjectival form "süß", also irrespectively if this form is inflected or part of a compound word. Our work is thus also addressing in the longer term the semantic access to dialectal dictionaries.

6 Conclusion

We presented an approach consisting in extracting meanings associated to entries in two dialectal dictionaries. Comparison of the expressions used to mark those meaning can be done only after applying basic natural language processing to those expressions. Expressions that are judged as being similar are cross-linked. Furthermore we encode those meanings in Semantic Web representation languages and can so link to lexical and knowledge resources available in the Linked Data Framework. This step in supporting potentially the semantically base cross-linking of our lexical entries to other dialectal dictionaries published in the web.

Current work is dedicated in improving the model for an adequate representation of more complex linguistic phenomena, and also in investigating how our approach could be applied for linking our dictionaries not only to DBpedia/Wiktionary but also to other lexical resources. We think here in particular to portals that already offer a network of dictionaries, like the Trier Wöterbuchnetz[26], which

[26] http://woerterbuchnetz.de/

contains a lot of dialectical dictionaries, also offering cross-links between entries. A next step will consist in published the content of our dictionaries in the Web, so that references from other sources in the LOD to our dictionaries can be be implemented.

7 References

Arbeitsplan (1912). Accessed at http://www.oeaw.ac.at/icltt/dinamlex-archiv/Arbeitsplan.pdf [10/04/2014]

Declerck, T. (202). A set of tools for integrating linguistic and non-linguistic information. In: *Proceedings of SAAKM (ECAI Workshop).*

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., Fergerson, 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

Hornung, M., Grüner, S. (2002) *Wörterbuch der Wiener Mundart*; Neubearbeitung. öbvhpt, Wien.

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).

Österreichisches Wörterbuch, 42. Auflage. 2012.

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/icltt/dinamlex-archiv/Straffungskonzept_1998.pdf [10.04.2014].

Theron, R., Wandl-Vogt, E. (2014). The fun of exploration: how to access a non-standard language corpus visually. In: *VisLR: Visualization as added value in the development, use and evaluation of LRs. Workshop-Proceedings of LREC2014.*

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. (2008). ..wie man ein Jahrhundertprojekt zeitgemäß halt: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX). In: Ernst, P. (ed) 2008, *Bausteine zur Wissenschaftsgeschichte von Dialektologie / germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen*, Wien: 93-112.

Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis* 3 (2010): 35-53.

Wandl-Vogt, E., Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in* *the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Wiktionary RDF extraction. Accessed at: http://dbpedia.org/Wiktionary [10/04/2014]

Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-). Verlag der Österreichischen Akademie der Wissenschaften. Wien. Accessed at: http://hw.oeaw.ac.at/cl?frames=yes [10/04/2014].

Acknowledgements

The DFKI part of this work is supported by the LIDER Project. LIDER: "Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe" is a FP7 project with reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.

Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen

Jörg Didakowski, Nadja Radtke
Berlin-Brandenburgische Akademie der Wissenschaften,
Technische Universität Dortmund
didakowski@bbaw.de, nadja.radtke@tu-dortmund.de

Abstract

Im Hinblick auf eine Zuarbeit für die lexikographische Arbeit an einem lexikalischen Informationssystem zu deutschen Stützverbgefügen wird das DWDS-Wortprofil vorgeschlagen. Mithilfe dieses Werkzeugs kann eine zeitintensive und mühsame Recherchearbeit über eine Textsuchmaschine vermieden werden, indem auf Basis eines ausgewogenen Korpus potenzielle Stützverbgefüge bereitgestellt werden. Des Weiteren wird die Einbeziehung von Assoziationsmaßen vorgeschlagen, um die Menge des zu sichtenden Materials für die lexikographische Arbeit weiter reduzieren zu können.
Keywords: Stützverbgefüge; Assoziationsmaße; computerlinguistische Verfahren

1 Einleitung

Neben den traditionellen Textkorpora stehen den Lexikographen heutzutage auch digitale Textkorpora zur Verfügung, die über Abfrage- und Analysewerkzeuge die Möglichkeiten der lexikographischen Arbeit stark erweitern (Engelberg & Lemnitzer 2009). Eines dieser Werkzeuge stellt das DWDS-Wortprofil dar, welches beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen (das SVG-Wiki) hilft bzw. den Aufbau von diesem gar erst ermöglicht. Dieses Werkzeug ist Teil des Digitalen Wörterbuchs der deutschen Sprache (DWDS), eines Projekts der Berlin-Brandenburgischen Akademie der Wissenschaften. Ziel des vorliegenden Beitrags ist es, die Potenziale der Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis aufzuzeigen.

Im 2. Kapitel des Beitrags stellen wir zunächst das SVG-Wiki vor, dabei legen wir den Gegenstand fest, bestimmen die Zielgruppe, führen die Komponenten des SVG-Wikis ein und heben abschließend die korpusbasierte Erarbeitung des SVG-Wikis hervor. Im 3. Kapitel beschreiben wir das DWDS-Wortprofil, das für die Datenerhebung des SVG-Wikis grundlegend ist. Anschließend gehen wir im 4. Kapitel des Beitrags genauer auf die Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis ein.

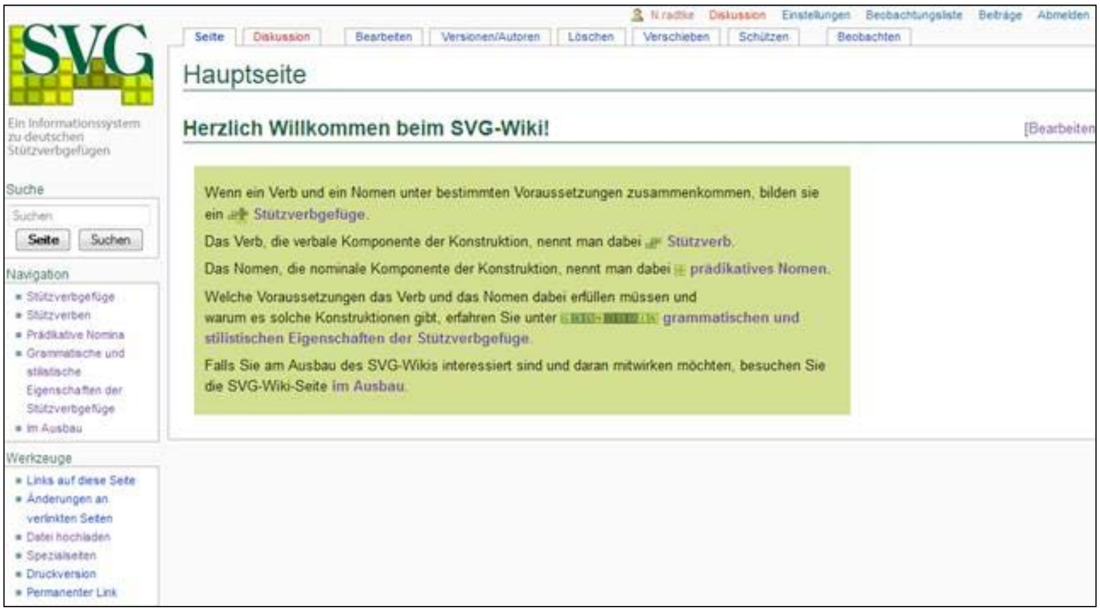
2 Das SVG-Wiki: Ein lexikalisches Informationssystem zu deutschen Stützverbgefügen

Den Gegenstand des SVG-Wikis bilden Stützverbgefüge (SVG, engl. *support verb constructions*) des Deutschen wie z.B. *Anwendung finden*, *zur Anwendung kommen* und *Kritik üben*, die aus einem prädikativen Nomen (z.B. *Anwendung*) und einem semantisch blassen Stützverb (z.B. *finden* oder *kommen*) konstruiert werden[1]. Abgegrenzt werden SVG von freien Konstruktionen (z.B. *zur Party kommen*) sowie von Idiomen (z.B. *zu Potte kommen*); beschrieben werden sie in Bezug auf ihre Systematik und hinsichtlich ihrer Leistungen (z.B. Pottelberge van 2001; Seifert 2004; Heine & Wotjak 2005; Storrer 2007).

SVG haben bereits Eingang in die Grammatiken, Wörterbücher und Lehrwerke des DaF-Unterrichts gefunden. Wie sie am besten eingeführt und behandelt werden, wird in der Forschungsliteratur intensiv diskutiert. Vermittelt werden sie bei der Arbeit mit dem Wortschatz sowie im Grammatikunterricht, indem einerseits ausgewählte SVG aufgeführt und andererseits ausgewählte grammatische und stilistische Eigenschaften der SVG beschrieben werden. Wie man dabei zu einer adäquaten Auswahl der SVG gelangt und wie Vollständigkeit und Systematik hinsichtlich der grammatischen und stilistischen Eigenschaften der SVG erreicht werden können, bleibt jedoch offen. Die zu vermittelnden SVG auszuwählen, um dann diese in Hinblick auf ihre Systematik behandeln zu können, fällt hier nach wie vor in das Aufgabenfeld der DaF-Lehrenden. Eine vor diesem Hintergrund hilfreiche Ressource zu deutschen SVG für die DaF-Lehrenden gibt es noch nicht. Das SVG-Wiki schließt diese Lücke, indem es die DaF-Lehrenden als zukünftige Nutzer vorsieht und aufbauend auf den bereits ausgereiften Vorschlägen für die Printlexikographie (Heine 2006; Heine 2008) eine Wörterbuchkomponente (Spezialwörterbuch) mit einer Grammatikkomponente verbindet. Das SVG-Wiki ist als ein digitales wikibasiertes Informationssystem realisiert[2], das den Nutzern im Internet zur freien Verfügung stehen wird und kontinuierlich erweitert werden soll (siehe Abbildung 1).

[1] In der deutschen linguistischen Fachliteratur findet man unterschiedliche Termini wie *Nominalisierungsverbgefüge, Funktionsverbgefüge* und *Streckverbgefüge,* die ebenfalls unterschiedlich begrifflich gefasst sind. Bei der Bestimmung des Gegenstandes des SVG-Wikis bedienen wir uns einem möglichst weiten Begriff der SVG: Das prädikative Nomen steht dabei im Akkustativ oder kommt als Präpositionalphrase vor, es ist abstrakt, wird deverbal oder deadjektivisch gebildet, kann ein Fremdlexem sein und idiomatisch verwendet werden.

[2] Das Wikisystem (hier: MediaWiki) ermöglicht u.a., die jeweiligen Komponenten des SVG-Wikis kollaborativ auszubauen. So können z.B. die Nutzer die noch nicht im SVG-Wiki berücksichtigten Stützverben zur Erweiterung des SVG-Wikis auf den dafür vorgesehenen Seiten des SVG-Wikis vorschlagen.



Hauptseite

Herzlich Willkommen beim SVG-Wiki!

Wenn ein Verb und ein Nomen unter bestimmten Voraussetzungen zusammenkommen, bilden sie ein Stützverbgefüge.

Das Verb, die verbale Komponente der Konstruktion, nennt man dabei Stützverb.

Das Nomen, die nominale Komponente der Konstruktion, nennt man dabei prädikatives Nomen.

Welche Voraussetzungen das Verb und das Nomen dabei erfüllen müssen und warum es solche Konstruktionen gibt, erfahren Sie unter grammatischen und stilistischen Eigenschaften der Stützverbgefüge.

Falls Sie am Ausbau des SVG-Wikis interessiert sind und daran mitwirken möchten, besuchen Sie die SVG-Wiki-Seite im Ausbau.

Abbildung 1: Hauptseite des SVG-Wikis.

Ziel des SVG-Wikis ist es, die SVG in ihrem Grundbestand festzuhalten, die DaF-Lehrenden bei der Auswahl der SVG zu unterstützen und die SVG in Bezug auf ihre grammatischen und stilistischen Eigenschaften zu beschreiben.

Das Besondere an dem SVG-Wiki ist, dass die jeweiligen Komponenten korpusbasiert erarbeitet werden. Ausgegangen wird hierbei von einer Lemmaliste der Stützverben, die aus 23 ausgewählten Grammatiken und unter Berücksichtigung der Forschungsliteratur erstellt wurde; sie enthält ca. hundert Stützverben mit den für sie charakteristischen und für die Datenerhebung relevanten Merkmalen. So wird z.B. bei dem Stützverb *finden* eingetragen, dass dieses mit einem prädikativen Nomen im Akkusativ vorkommt. Anhand dieser Lemmaliste werden zunächst die prädikativen Nomina der jeweiligen Stützverben[3] mithilfe des DWDS-Wortprofils ermittelt; von den ermittelten prädikativen Nomina ausgehend, besteht im weiteren Schritt der Erarbeitung die Möglichkeit, die im SVG-Wiki noch nicht berücksichtigten Stützverben ebenfalls mithilfe des DWDS-Wortprofils zu entdecken. Anschließend findet die lexikographische Beschreibung der Stützverben, der prädikativen Nomina und der SVG anhand des DWDS-Kernkorpus statt. Parallel dazu wird die Grammatikkomponente des SVG-Wikis erarbeitet. Das DWDS-Kernkorpus stellt somit die primäre Quelle für die Wörterbuchbasis der Wörterbuchkomponente dar und bildet gleichzeitig die Datengrundlage für die Grammatikkomponente des SVG-Wikis.

[3] Kamber (2008) geht ebenfalls in seiner korpusbasierten Untersuchung zu den nominalen Prädikaten des Deutschen von den jeweiligen Verben aus.

3 Das DWDS-Wortprofil

Das DWDS-Wortprofil (Didakowski & Geyken 2013) stellt Kookkurrenzpaare für verschiedene grammatische Relationen wie z.B. Akkusativ-/Dativobjekt, Genitivattribut, Adjektivattribut und präpositionales Komplement/Modifizierer bereit. Die Kookkurrenzpaare werden mithilfe von computerlinguistischen Verfahren automatisch extrahiert. In Kilgarriff et al. (2004) wird für die automatische Extraktion grammatischer Kookkurrenzpaare die flache *Sketch-Grammar* vorgeschlagen, mit der über reguläre Ausdrücke Kookkurrenzpaare für bestimmte grammatische Relationen extrahiert werden können. Ivanova et al. (2008) zeigen jedoch, dass es für das Deutsche sinnvoll ist, um gute Ergebnisse zu erzielen, auf eine reichhaltigere linguistische Analyse zurückzugreifen. Beim DWDS-Wortprofil wird für eine reichhaltigere Analyse die TAGH-Morphologie (Geyken & Hanneforth 2006) und der regelbasierte Parser SynCoP (Syntactic Constraint Parser, Didakowski 2008) verwendet. So kann die relativ reichhaltige Morphologie und freie Wortstellung im Deutschen angemessen behandelt und die Kookkurrenzen mit gewünschter Qualität extrahiert werden.

Im DWDS-Wortprofil sind die Kookkurrenzpaare mit Werten verschiedener Assoziationsmaße versehen. Derzeit werden drei verschiedene Assoziationsmaße berechnet: 1) die reine Frequenz, 2) das auf dem Dice-Koeffizienten basierende logDice-Maß (Rychlý 2008) und 3) das auf Mutual-Information basierende MI-log-Freq-Maß (Kilgarriff & Tugwell 2002). Mithilfe dieser Maße können Kookkurrenzpaare nach Verbindungsstärke bzw. Anziehungskraft sortiert werden. Hierbei wird das Assoziationsmaß in der Regel so gewählt, dass die entsprechende Sortierung für eine bestimmte Aufgabe am geeignetsten ist (Evert 2008).

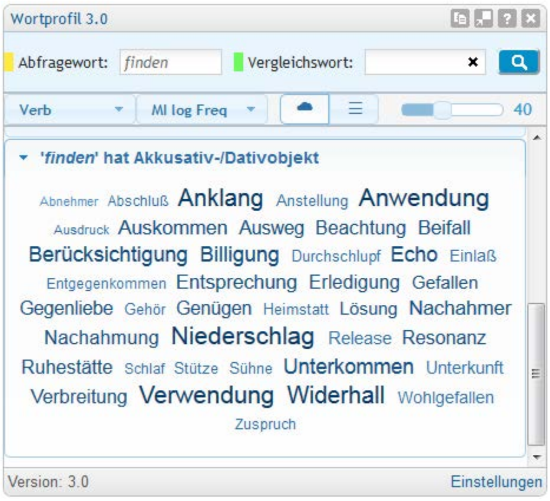
Die Korpusgrundlage für das DWDS-Wortprofil bilden das DWDS-Kernkorpus und verschiedene verbreitete Zeitungen (*Süddeutsche Zeitung*, *DIE ZEIT*, *Berliner Zeitung*, *DIE WELT*, *Der Tagesspiegel*, *Bild*). Das DWDS-Kernkorpus ist ein Referenzkorpus der *deutschen* Sprache des 20. Jahrhunderts und ist ausgeglichen bezüglich verschiedener Textsorten, die zudem gleichmäßig über das 20. Jahrhundert verteilt sind. Es umfasst über 100 Millionen laufende Wortformen (Tokens) und hat damit eine vergleichbare Größe wie das British National Corpus (Geyken 2007). Das DWDS-Kernkorpus stellt somit als ausgeglichenes Referenzkorpus das Herzstück der Korpusbasis des DWDS-Wortprofils dar und nimmt damit eine besondere Stellung ein. Die gesamte Korpusgrundlage des DWDS-Wortprofils umfasst ca. 1,7 Milliarden laufende Wortformen (Tokens) und reicht zeitlich vom Anfang des 20. Jahrhunderts bis heute. Die Kookkurrenzpaare sind hierbei für die gesamte Korpusbasis und auch für die einzelnen Subkorpora berechnet. So können Kookkurrenzpaare z.B. ausschließlich auf Basis des DWDS-Kernkorpus abgefragt werden.

Das DWDS-Wortprofil ist einerseits über die DWDS-Webseite und andererseits innerhalb der CLARIN-Infrastruktur über WebLicht (Hinrichs et al. 2010) zugänglich, wo es in Verarbeitungsketten integriert werden kann.

Ein Beispiel für eine DWDS-Wortprofil-Abfrage auf der DWDS-Webseite ist in Abbildung 2 zu sehen. Hier wurde das Verb *finden* für die Akkusativ-/Dativobjekt-Relation unter Verwendung des MI-log-

Freq-Maßes auf der Basis des DWDS-Kernkorpus abgefragt. Die relevanten Kookkurrenzpartner zu dem Verb *finden* werden als Wortwolke dargestellt. Je größer der Wert des Assoziationsmaßes eines Kookkurrenzpaares ist, desto größer wird der Kookkurrenzpartner in der Wolke dargestellt. Eine alternative Darstellungsform zu dieser Wortwolke ist die Tabellenansicht, in der die Kookkurrenzpartner nach dem Assoziationsmaß sortiert aufgelistet und genauere Informationen zu Wortkategorien und Assoziationswerten aufgeführt sind.

Hervorzuheben ist dabei, dass im DWDS-Wortprofil über die einzelnen Kookkurrenzpartner direkt auf die entsprechenden Korpusbelege zugegriffen werden kann. Erst dadurch wird eine sinnvolle lexikographische Arbeit möglich. In Abbildung 3 sind die Belege für das Kookkurrenzpaar *Anwendung finden* aufgeführt.



Wortprofil 3.0

Abfragewort: finden

Vergleichswort:

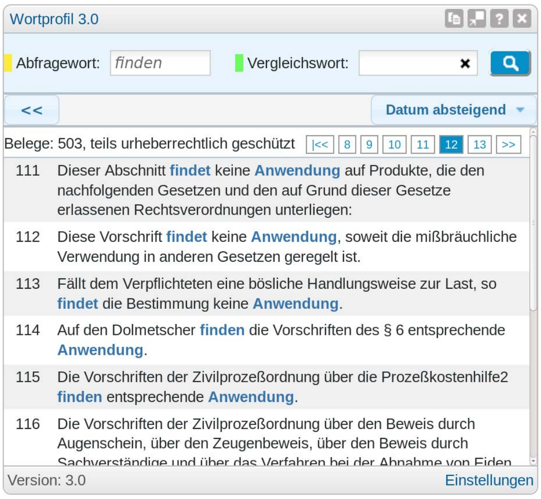
Verb

MI log Freq

40

'finden' hat Akkusativ-/Dativobjekt

Abbildung 2: DWDS-Wortprofil-Wortwolke.



Wortprofil 3.0

Abfragewort: finden

Vergleichswort:

<<

Datum absteigend

Belege: 503, teils urheberrechtlich geschützt |<< 8 9 10 11 12 13 >>

111 Dieser Abschnitt **findet** keine **Anwendung** auf Produkte, die den nachfolgenden Gesetzen und den auf Grund dieser Gesetze erlassenen Rechtsverordnungen unterliegen:

112 Diese Vorschrift **findet** keine **Anwendung**, soweit die mißbräuchliche Verwendung in anderen Gesetzen geregelt ist.

113 Fällt dem Verpflichteten eine bösliche Handlungsweise zur Last, so **findet** die Bestimmung keine **Anwendung**.

114 Auf den Dolmetscher **finden** die Vorschriften des § 6 entsprechende **Anwendung**.

115 Die Vorschriften der Zivilprozeßordnung über die Prozeßkostenhilfe2 **finden** entsprechende **Anwendung**.

116 Die Vorschriften der Zivilprozeßordnung über den Beweis durch Augenschein, über den Zeugenbeweis, über den Beweis durch Sachverständige und über das Verfahren bei der Abnahme von Eiden

Abbildung 3: DWDS-Wortprofil-Belege.

Mit dem DWDS-Wortprofil ist es also möglich, auf strukturierte Weise Kookkurrenzpaare mit den dazugehörigen Korpusbelegen zu ermitteln. Verschiedene Assoziationsmaße können dazu verwendet werden, bestimmte Kookkurrenzpaare aus der Menge der Kookkurrenzpaare hervorzuheben. Die große Korpusbasis zusammen mit dem DWDS-Kernkorpus als ausgeglichenem Bestandteil gewährleistet hierbei ein breites Spektrum an Kookkurrenzen und repräsentative Ergebnisse.

4 Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis

Im Folgenden begründen wir die Wahl der Korpusbasis für die Ermittlung der prädikativen Nomina beim Aufbau des SVG-Wikis und gehen kurz auf verschiedene Möglichkeiten zur Ermittlung der prädikativen Nomina ein. Hierbei heben wir die Potenziale des DWDS-Wortprofils hervor. Im Weiteren erläutern wir, ob die Assoziationsmaße dabei helfen können, die Menge der Kookkurrenzpaare für mögliche prädikative Nomina zu verkleinern, sodass weniger Kookkurrenzpaare gesichtet werden müssen und dabei trotzdem der Grundbestand der SVG festgehalten werden kann.

4.1 Ermittlung der prädikativen Nomina

Bei der Ermittlung der prädikativen Nomina wird die Korpusbasis auf das DWDS-Kernkorpus eingeschränkt, damit mit Blick auf die Zielsetzung des SVG-Wikis die Dekaden des 20. Jahrhunderts sowie verschiedene Textsorten gleichermaßen vertreten sind.

Zur Ermittlung der prädikativen Nomina kann einerseits die DWDS-Suchmaschine und andererseits das DWDS-Wortprofil genutzt werden. Bei der Nutzung der DWDS-Suchmaschine erhält man z.B. für die Abfrage zu dem Verb *finden* eine Liste mit 82.864 Treffern, in der nach den prädikativen Nomina manuell gesucht werden muss. Bei der Nutzung des DWDS-Wortprofils reduziert man bereits durch die Wahl einer grammatischen Relation die Menge der Treffer. So erhält man bei dem Verb *finden* durch die Wahl der Akkusativ-/Dativobjekt-Relation eine Liste mit 779 Kookkurrenzpaaren, die dann durch das Zugreifen auf einzelne Korpusbelege manuell nach prädikativen Nomina klassifiziert werden können. Hierbei beträgt die durchschnittliche Anzahl an Kookkurrenzpaaren für die 31 in den Grammatiken am häufigsten genannten Stützverben pro Verb mit dem Nomen im Akkusativ ca. 437 und pro Verb mit dem Nomen als Präpositionalphrase ca. 843. Hervorzuheben ist im Weiteren, dass dem DWDS-Wortprofil eine reichhaltigere linguistische Analyse (siehe Kapitel 3) zugrunde liegt und somit bestimmte Fälle, die die Suche mit der DWDS-Suchmaschine zusätzlich erschweren, vermieden werden. Zu solchen Fällen gehören Verben mit einem abtrennbaren Präfix (z.B. *ausüben*) oder Verben, die bezüglich einer Wortform homograph zu einem anderen Verb sind (z.B. *geraten* und *raten*). Somit ermöglicht das DWDS-Wortprofil, die Ermittlung der prädikativen Nomina überhaupt in einem realistischen Zeitrahmen bewältigen zu können.

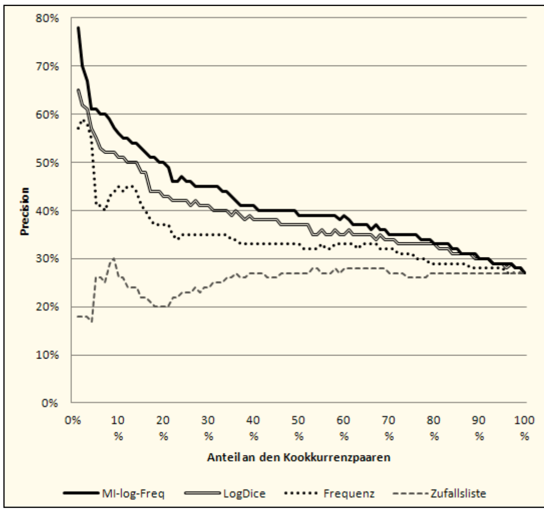
4.2 Verwendung von Assoziationsmaßen bei der Ermittlung der prädikativen Nomina

Über das DWDS-Wortprofil ist es möglich, die abgefragten Kookkurrenzpaarlisten nach verschiedenen Assoziationsmaßen zu sortieren (siehe Kapitel 3). Hierbei ist die Frage, ob die Assoziationsmaße in der Lage sind, die Kookkurrenzpaarlisten so zu sortieren, dass am Anfang der Listen die Dichte der prädikativen Nomina sehr hoch ist und am Ende nur wenige prädikative Nomina vorkommen. So könnten die Kookkurrenzpaarlisten verkleinert werden, ohne dass zu viele prädikative Nomina verloren gehen. Auf diese Weise kann Recherchearbeit eingespart werden.[4]

[4] Langer (2009) versucht, Funktionsverbgefüge vollautomatisch aus Korpora zu gewinnen. Dies will er hauptsächlich über Assoziationsmaße realisieren. Er zeigt, dass die Maße für so eine Aufgabenstellung

Um in Bezug zu der oben genannten Fragestellung eine Bewertung der Assoziationsmaße durchzuführen, folgen wir der Methodik in Evert et al. (2000) und beurteilen die Maße anhand von Precision und Recall. Hierzu wurden vier Verben herangezogen: *bringen*, *finden*, *kommen* und *üben*. Zu diesen Verben wurden mithilfe des DWDS-Wortprofils unter Berücksichtigung der jeweiligen grammatischen Relationen und durch die Wahl des DWDS-Kernkorpus als Korpusbasis Kookkurrenzpaarlisten ermittelt, wobei eine Minimalfrequenz für die Kookkurrenzpaare auf fünf festgelegt wurde. Die Listen wurden dann manuell vollständig gesichtet und nach prädikativen Nomina klassifiziert. Insgesamt wurden 859 prädikative Nomina aus 9.166 Kookkurrenzpaaren identifiziert. Die prädikativen Nomina machen demnach 27% der Gesamtmenge aus. Über die so erstellte Referenzmenge können nun Precision und Recall ermittelt werden. Zur Bewertung des Nutzens der Assoziationsmaße wird hier zusätzlich eine Zufallssortierung der Kookkurrenzlisten hinzugezogen. Die Zufallsliste deckt den Fall ab, dass kein Assoziationsmaß zur Sortierung verwendet wird.

Der Verlauf von Precision und Recall zu den einzelnen Assoziationsmaßen und der Zufallsliste ist in den Diagrammen in Abbildung 4 und 5 zu sehen. Auf den X-Achsen ist der Anteil an Kookkurrenzpaaren, der durch das Verkürzen der Kookkurrrenzpaarlisten entsteht, in Prozent aufgetragen. Precision und Recall zu den einzelnen Anteilen sind jeweils auf den Y-Achsen in Prozent ablesbar.



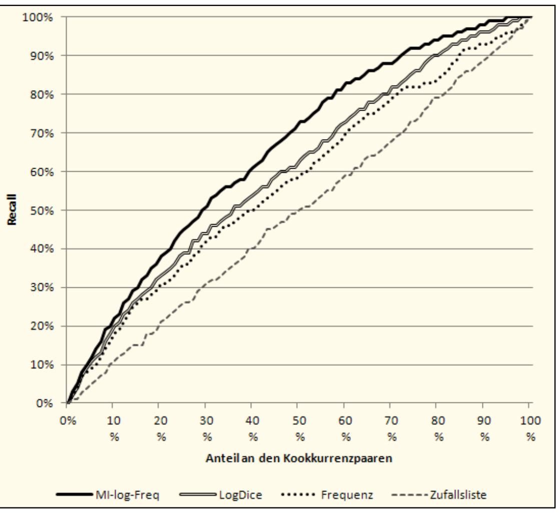


Abbildung 4: Precision-Kurven.

Abbildung 5: Recall-Kurven.

Der Verlauf der Precision-Kurven in Abbildung 4 zeigt, dass über die Sortierung nach dem MI-log-Freq-Maß die besten Precision-Werte erreicht werden. Hier liegt der Anteil der prädikativen Nomina sogar bei 78%, wenn 1% der Kookkurrenzpaare herangezogen wird. Mit anwachsendem Anteil an Kookkurrenzpaaren flachen die Precision-Kurven der Assoziationsmaße anfangs ab und fallen stetig auf das Grundniveau von 27%, welches durch den Anteil an prädikativen Nomina gesetzt ist. Bei der Zu-

nicht ausreichend sind. Die vollautomatische Extraktion, die Langer (2009) im Sinn hat, wird in unserer Vorgehensweise jedoch nicht verfolgt.

fallssortierung hingegen bewegt sich die Precision lediglich nahe am Grundniveau und sogar darunter. Der Verlauf der Recall-Kurven in Abbildung 5 zeigt ergänzend dazu, dass unter Verwendung des MI-log-Freq-Maßes nur die Hälfte der Kookkurenzpaare gesichtet werden muss, um bereits 72% der prädikativen Nomina zu ermitteln. Sind die Kookkurrenzpaare nach dem Zufall sortiert, bekommt man hingegen lediglich ca. 50% der prädikativen Nomina.

Hier wird deutlich, dass die Assoziationsmaße hilfreich sind, wenn man bei der Ermittlung der prädikativen Nomina den Umfang und den damit verbundenen zeitlichen Aufwand reduzieren möchte und gleichzeitig möglichst viele prädikative Nomina als Grundbestand ermitteln will. Hierbei hat sich das MI-log-Freq-Maß als am geeignetsten herausgestellt.

5 Zusammenfassung

Das DWDS-Wortprofil hat das Potential, den Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen entscheidend zu erleichtern. Die Aufgabe des DWDS-Wortprofils liegt bei dem Aufbau des SVG-Wikis darin, mögliche prädikative Nomina für Stützverben bereitzustellen. Dadurch wird eine zeitintensive und mühsame Recherche über eine Textsuchmaschine vermieden. Unter dem Aspekt zeitlicher Restriktionen ist das DWDS-Wortprofil sogar unabdingbar. Über eine zusätzliche Bewertung nach Assoziationsmaßen kann zudem weitere Recherchezeit eingespart werden.

6 Literaturhinweise

Das digitale Wörterbuch der deutschen Sprache (DWDS). Accessed at: www.dwds.de [11/04/2014].

Didakowski, J. (2008). Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers. In A. Storrer, A. Geyken et al. (eds.) Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing, KONVENS 2008. Berlin et al.: Mouton de Gruyter, pp. 65-78.

Didakowski, J., Geyken, A. (2013). From DWDS corpora to a German Word Profile – methodological problems and solutions. In Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information, 2nd Work Report of the Academic Network "Internet Lexicography" (OPAL - Online publizierte Arbeiten zur Linguistik X/2012). Mannheim: Institut für Deutsche Sprache, pp. 43-52.

Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Evert, S., Heid, U. et al. (2000). Methoden zum qualitativen Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In W. Zühlke, Ernst G. Schukat-Talamazzini (eds.) Sprachkommunikation, KONVENS 2000. Berlin et al.: VDE, pp. 215-220.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, M. Kytö (eds.) Corpus Linguistics. An International Handbook of the Science of Language and Society. Berlin: Mouton de Gruyter, pp. 1212-1248.

Geyken, A., Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In A. Yli-Jyrä, L. Karttunen et al. (eds.) Finite State Methods and Natural Language Processing. Berlin et al.: Springer, pp. 55-66.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In Ch. Fellbaum (eds.) Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London et al.: Continuum, pp. 23-41.

Heine, A., Wotjak, B. (2005). Zur Abgrenzung und Beschreibung verbonominaler Wortverbindungen (Wortidiome, Funktionsverbgefüge, Kollokationen). In *Deutsch als Fremdsprache. Zeitschrift für Theorie und Praxis des Deutschunterrichts für Ausländer*, 42(3), pp. 143-153.

Heine, A. (2006). Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-) Lexikographie. Frankfurt a. M. et al.: Peter Lang.

Heine, A. (2008). Funktionsverbgefüge richtig verstehen und verwenden. Ein korpusbasierter Leitfaden mit finnischen Äquivalenten. Frankfurt a. M. et al.: Peter Lang.

Hinrichs, E., Hinrichs, M. et al. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010, System Demonstrations (ACLDemos '10), Association for Computational Linguistics*. Stroudsburg, PA (USA), pp. 25-29.

Ivanova, K., Heid, U. et al. (2008). Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008. Marrakech (Morocco), pp. 2101-2107.

Kilgarriff, A., Tugwell, D. (2002). Sketching Words. In M.-H.Corréard (eds.) Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S.Atkins, EURALEX, pp. 125-137.

Kilgarriff, A., Rychlý, P. et al. (2004). The Sketch Engine. In *Proceedings of EURALEX* 2004. Lorient (France), pp. 105-116.

Kamber, A. (2008). Funktionsverbgefüge – empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen. Tübingen: Max Niemeyer.

Langer, S. (2009). Funktionsverbgefüge und automatische Sprachverarbeitung. München: LINCOM.

MediaWiki. Accessed at: www.wikipedia.org [11/04/2014].

Pottelberge van, J. (2001). Verbonominale Konstruktionen, Funktionsverbgefüge. Vom Sinn und Unsinn eines Untersuchungsgegenstandes. Heidelberg: Universitätsverlag C. Winter.

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka, A. Horák (eds.): *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*. Brno: Masaryk University, pp. 6-9.

Seifert, J. (2004). Funktionsverbgefüge in der deutschen Gesetzessprache (18. – 20. Jahrhundert). Hildesheim et al.: Georg Olms.

Storrer, A. (2007). Corpus-based investigations on German support verb constructions. In Ch. Fellbaum (eds.) Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London et al.: Continuum, pp. 164-187.

WebLicht. Accessed at: https://weblicht.sfs.uni-tuebingen.de [11/04/2014].

Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies

Iztok Kosem[1], Polona Gantar[2], Nataša Logar[3], Simon Krek[4]
[1]Trojina, Institute for Applied Slovene Studies
[2]Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia
[3]Faculty of Social Sciences, University of Ljubljana
[4]Jožef Stefan Institute, Ljubljana
iztok.kosem@trojina.si, apolonija.gantar@guest.arnes.si,
natasa.logar@fdv.uni-lj.si, simon.krek@guest.arnes.si

Abstract

Due to increasingly large amounts of authentic data to analyse, lexicographers are nowadays looking to language technologies to provide them with not only the tools to analyse the data, but also with tools and methods that ease and speed up the data analysis. One of the most promising avenues of research has been the automation of early stages of the corpus data analysis, with the aim to summarize, and consequently reduce, the amount of corpus data that the lexicographers need to examine. However, most of this research deals with general lexicography; terminology is yet to extensively test these methods. This paper attempts to address this gap by presenting two separate Slovene research projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that used the same method of automatic extraction of corpus data (presented in Kosem et al. 2013). After describing the projects and the corpora use, similarities and differences in the parameter settings and the quality of extracted data in the two projects are presented. We conclude with discussing the further potential of automation in both general and specialised lexicography.

Keywords: data extraction; terminology; general language; collocations; dictionary; GDEX

1 Introduction

In recent years, lexicography has witnessed several projects where automation of different aspects of lexicographer's work has been successfully implemented, such as detection of new words or meanings (Cook et al. 2013) or initial data extraction (Kosem et al. 2013). This trend of increasing the role of a computer in the dictionary-making process follows Rundell and Kilgarriff's (2011) vision of focussing lexicographer's tasks towards validating and completing the data extracted by a computer.

The calls for automation originate mainly from general lexicography where lexicographers are faced with increasingly larger corpora that they need to analyze. But what about using automation in the making of dictionaries, such as terminological dictionaries, where much smaller and more specialized corpora are used? To what extent can automation methods used in general lexicography be transferred to specialized lexicography or terminology? This paper attempts to provide some answers to these questions by describing and discussing two Slovenian projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that tested the use of automation in database compilation.

We first briefly describe both projects, and the corpora used for automatic extraction of data. This is followed by the description of the automatic process, and an overview of the settings used in the two projects. Then, the findings are presented, focusing on the differences as well as similarities identified between the results of automatic data extraction in the two projects. We conclude with some thoughts on the further potential of automation in both general and specialised lexicography, and outline our plans for the future.

2 Slovene Lexical Database

The Slovene Lexical Database (SLD; Gantar & Krek 2011) is one of the results of the Communication in Slovene[1] project that has developed language data resources, natural language processing tools and resources, and language description resources for Slovene. The Slovene Lexical Database 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. 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:

- lemma or the headword,
- senses and subsenses (labelled with semantic indicators and in many cases described with semantic frames),
- multi-word expressions,
- syntactic structures (representing a formalization of typical patterns on the clause and phrasal level),
- collocations, and
- corpus examples.

[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/.

3 Terminological database Termis

Applied research project Termis[2] took place between 2011 and 2013. The aim of the project was the compilation of an online dictionary-like terminological database for the discipline of public relations. The basis of the project was KoRP,[3] a corpus of public relations texts (Logar 2013). It has been envisaged from the beginning that the entries in the terminological database would contain English translations of headwords, explanations, syntactic and collocational information, and corpus examples. The database in now completed and is freely accessible online at http://www.termania.net. It comprises 2000 entries that also offer links to the KoRP corpus and the Gigafida corpus, a reference corpus of Slovene.

4 Using automatic data extraction in the two projects

The decision to use automatic extraction of lexical information from the corpus in both projects comes from the need to reduce time and cost connected with the production of dictionaries, by utilizing new possibilities offered by state-of-the-art tools for corpus analysis. The main idea behind using automatic extraction of corpus data is to reduce the amount of time spent by lexicographers on examining corpus data, especially on browsing through plethora of corpus examples. Lexicographic analysis remains corpus-based (or driven); however, the initial selection of corpus data to be analysed is left to the computer. The lexicographer then examines, validates, and completes the information and shapes it into the final dictionary entry.

The automatic method used in the two projects relies heavily on Word Sketch (Kilgarriff & Tugwell 2002) and GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), two functions that are part of the Sketch Engine corpus tool. The method requires a lemma list, sketch grammar for the building of word sketches, GDEX configuration(s), and settings that set thresholds for data extraction. An API script is then used to extract from the corpus collocates under grammatical relations, defined in the sketch grammar, and examples of their use. The method is described in more detail in Kosem et al. (2013), thus the next sections focus on the main differences in the automatic method used by the two projects.

4.1 Corpora

The basis for the extraction of lexical information for the Slovene Lexical Database was the Gigafida corpus[4] (Logar Berginc et al. 2012), containing 1.18 billion words or 39,427 texts created between 1990

[2] http://www.termis.fdv.uni-lj.si/

[3] http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl

[4] http://www.gigafida.net/

and 2011 with printed texts representing 84.35% and internet texts 15.65%. Printed part contains fiction (2%), non-fiction and textbooks (4%), and periodicals such as daily newspapers (56%) and magazines (21%). Text originating from the web were published on news portals, pages of large Slovene companies and more important governmental, educational, research, cultural and similar institutions.

Automatic extraction of lexical information for the Termis project was conducted on a much smaller, specialised corpus – the KoRP corpus – containing 1.8 million words. The texts in the KoRP corpus were selected according to carefully designed criteria (Logar 2007) that make the corpus representative of a public relations field in Slovenia. It is important to note that the two corpora were lemmatised and morphosyntactically tagged with the same statistical tagger (Grčar, Krek & Dobrovoljc 2012), enabling comparisons of extracted data.

4.2 List of lemmas

The two projects used completely different approaches to devising a list of lemmas for automatic extraction. For the Slovene Lexical Database, a more homogenous group of lemmas was used, mainly comprising of not too frequent lemmas that were either monosemous of less polysemous according to sloWNet, a Slovene version of Wordnet (Fišer, 2009). Less polysemous nature of lemmas also enabled a better comparison of data extraction with the Termis project, given that the terms in Termis were mainly monosemous. An additional criterion for selection, which was preferred but not mandatory, was the absence of the lemma in the Dictionary of Standard Slovenian (SSKJ). 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).

In the Termis project, the lemma list was in fact a headword list and was built using a term extraction tool (Vintar 2010)[5]. The list contained 2127 items: 941 nouns, 199 verbs and 987 multi-word terms. Single- and multi-word term candidates have been extracted using morphosyntactic patterns and term weights, calculated by comparing their frequencies in the KoRP corpus and in a reference corpus of Slovene called FidaPLUS (Arhar Holdt & Gorjanc 2007), as well as phraseological stability of the extracted terminological unit. Each term candidate was carefully examined in its natural environment – the texts in the KoRP corpus – by a terminologist and experts in the field of public relations.

4.3 GDEX configurations

The GDEX tool (Kilgarriff et al. 2008) ranks corpus examples according to their quality, using measurable parameters such as example length, whole sentence form, syntax, and presence/absence of rare words, etc. The majority of work associated with devising GDEX configurations for automatic extraction was done during the SLD project; drawing on the experience in developing the first version of

[5] http://lojze.lugos.si/cgitest/extract.cgi

GDEX for Slovene (Kosem et al., 2011), four different configurations were designed, one for each word class in the SLD (noun, verb, adjective, adverb), the process involving several iterations of evaluation and comparison of results produced by the last two versions of configuration (Kosem et al. 2013). A good indication of the difference between the first version of GDEX for Slovene and the version for automatic extraction is that the former was designed to provide at least three good examples among the ten offered, while the latter aimed to have the top three examples meet the criteria of a good example. The Termis project's point of departure was using the final GDEX configurations used by the SLD project, evaluating them on a sample of lemmas and making adjustments to the heuristics, which proved to be minor, until the results were satisfactory. In the end, two different GDEX configurations were used, one for nouns and multi-word units, and one for verbs.

4.4 Settings for extraction

This part of the automatic extraction introduced the greatest number of differences between the two projects. Preparation of settings for extraction included providing values for the following six parameters:

- 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.

For the SLD project, three examples per collocate were extracted, and for the Termis project two examples per collocate. Both projects used a limit of maximum 25 collocates per grammatical relation.
The values of the remaining four parameters had to be obtained with statistical and manual analysis of the word sketches of a sample of lemmas used in automatic extraction. Namely, initial tests during the SLD project showed that the same values could not be used for all grammatical relations and collocates; for example, more salient and frequent relations of word classes (e.g. *adjective + noun* for adjectives) required higher thresholds due to a large number of collocates. Also, corpus frequency of the lemma played a vital role in setting the values; more frequent lemmas had more extensive word sketches and required higher thresholds, whereas rarer lemmas required lower thresholds or no thresholds at all. Consequently, both projects divided lemma lists into different frequency groups, with different settings used for each group. The SLD project used three frequency groups for each word class, with different frequency ranges for different word classes. On the other hand, the Termis project used three frequency groups for verbs, four frequency groups for nouns, and three frequency groups for multi-word units. Each category in the Termis project contained one group, the so-called 0 group, that included low frequency lemmas for which all the data available in the word sketches was extracted.

The only values that were shared by the two projects were values for minimum collocation salience for nouns and values for minimum gramrel salience for verbs; all other values were (much) lower for the Termis project than for the SLD project. This was a direct result of the difference in the sizes of the corpora used for automatic extraction of data.

4.5 Extracted lexical information: general language vs. specialized language

It is worth noting that a term as a name for a concept in a certain discipline is more difficult to specify than it is presented and argumented in the general theory on terminology (Wüster 1931; Felber 1984) – at least if terms are observed and identified in the context (Pearson 1998, as well as other perspectives, e.g. Cabré Castelví 2003). Such complexity of terms has been adequately summarized by Sager (1998/99) who argued that terms are merely words with a specific function, or in other words, terms are formally not very different from other words. This fact causes great difficulties to terminographers during preparatory stages, i.e. while preparing the headword list; on the other hand, this similarity between terms and other words is an advantage during the extraction of lexical context, as terminography can utilize lexicographic knowledge and tools for the analysis and description of a general language.

So far, we have compared grammatical relations/syntactic structures found in both Slovene Lexical Database and Termis, using a smaller number of noun entries that have a higher frequency per million words in the KoRP corpus than in the Gigafida corpus. The analysis showed that a large percentage of words acquire the specialised meaning only at a context level, especially with compounds or when we are dealing with polysemous words that have one of their meanings used also in a specialised domain or have developed their own specialised meaning.

The comparative analysis also focused on identifying syntactic structures common to both the general corpus (Gigafida) and the specialised corpus (KoRP), more specific to one of the corpora, or exclusive to one of the corpora. Similar comparison was made for collocations in both vocabularies. The sketch grammar contains 258 grammatical relations functioning as syntactic structures, and the automatically extracted data for noun entries showed that there were 69 (27%) attested syntactic structures, i.e. structures with identified collocates, in both corpora, 188 (73%) syntactic structures were found only in the Gigafida corpus, while one syntactic structure was found only in the KoRP corpus. These findings confirm that terminology does not differ from general language on a syntactic level, i.e. does not form terminology-specific syntactic structures. There are exceptions, however they are specific to particular lexical items; thus, a syntactic structure can be found in the language, but is not typical for a specific verb, noun, adjective etc. as used in the general language. An example of this is the structure VERB + NOUN4 for the collocation *communicate message*, which is typical for the field of public relations, but not for general Slovene where the pattern *communicate* + *about* + NOUN5 is more commonly used.

5 Discussion

The automatic extraction approach proved successful in both projects, in terms of providing good enough data for devising database entries and saving a great deal of lexicographer's/terminologist's time spent on more routine tasks. One of the important findings is that the steps used in the general language project (SLD) could be replicated in the terminological project (Termis), with some elements requiring little change (e.g. GDEX configurations) or no change at all (e.g. sketch grammar). Also, the evaluation of extracted corpus sentences in both projects reported good quality of the examples. The comparison clearly shows that the main work on any future project adopting this methodology would be dedicated to determining the settings for data extraction. Namely, this step exhibited the greatest differences between the projects, mainly on account of a significant difference in the size of the corpora used for automatic extraction.

Different nature of projects also enabled us to evaluate and test the approach on different lemmas in terms of corpus frequency and consequently in the amount of corpus data available. In SLD, the minimum frequency of a lemma was 600 occurrences (0.5 times per million words)[6] in the Gigafida corpus, whereas the threshold in Termis was determined by terminological potential of the word rather than its frequency (for example, some terms had only two or three occurrences in the KoRP corpus[7]). For high frequency lemmas, more work on settings for extraction was required in order to find the right balance between exporting enough data and excluding irrelevant grammatical relations and/or collocates. For very rare lemmas, i.e. for those in groups 0 in the Termis project[8], it was established that the value of the automatic approach is mainly in saving lexicographer's time by directly exporting all the data for each lemma and importing it into the dictionary-writing system, thus changing the lexicographer's task from analysis-selection-copying to validation-deletion.

The automatic extraction of data for multi-word units was conducted only for Termis, as the project was conducted after the conclusion of the SLD project when a new feature called Multi-word links had already been implemented in the Sketch Engine. The automatic extraction of lexical information was only possible for two-word patterns such as *adjective* + *noun* and *noun* + *noun*, and not for others (e.g. *noun* + *preposition* + *noun*). It is therefore not possible to make comparisons of the projects as far as automatic extraction of data for multi-word units is concerned. Nonetheless, we can report that the data obtained in the Termis project was found to be of similar quality as the data for single-word terms, with the main difference being in the GDEX configuration and settings used.

What is left for lexicographers to do are tasks such as sense division, definition writing, distributing and cleaning the automatically extracted information etc.; and as shown by studies such as Kosem et al. (2013), some of those tasks can be left to non-lexicographers, e.g. by using crowd-sourcing. Further-

[6] Majority of lemmas had frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words) in the Gigafida corpus.

[7] This still meant that these terms had a higher frequency per million words (1.1) than the least frequent lemmas in the SLD project.

[8] Groups 0 contained 889 terms in total.

more, the feedback from the terminologists devising entries in the Termis project showed that many extracted examples already contained definitions of terms or at least the information needed to devise them, indicating further avenues for the implementation of automation. It is noteworthy that the entries in the Termis database contain (encyclopaedic) definitions that are short (one sentence), medium-length (multi-sentence paragraph) or even longer (several paragraphs); in general they are longer than definitions (or semantic frames) in the Slovene Lexical Database.

6 Conclusion

Technological advances gave rise to corpora, enabling lexicographers to describe language more accurately and in greater detail than ever before, but ironically, corpora have now become a problem for lexicographers due to the increasingly larger amounts of data they contain. Consequently, it seems inevitable that more and more lexicographic tasks will become automated. There is simply too much data to analyse and not enough time to do it in – in addition, users want quick(er) access to up-to-date information. Initial experience on a Slovene lexicographic project has showed promising results, but it is even more encouraging that the automatic approach appears to be suitable also for terminological purposes.

The automatic method by Kosem et al. (2013) has the most potential for projects where a dictionary or a database is devised from scratch,[9] but it can also be useful for existing dictionaries. For example, periodical automatic extraction of regularly updated corpus data could facilitate quicker detection and description of new meanings and usages of the words. This remains one of the avenues of future research; namely, how to automatically extract and include in the database only the new information on the use of a particular word or phrase. By this we do not mean only new words and meanings, but also new uses of existing meanings.

Future plans as far as the Slovene Lexical Database is concerned include a more in-depth evaluation of entries devised with automatically extracted data, as well as their comparison with manually devised entries. We also aim to test automatic extraction on more frequent lemmas, where we expect much more work with setting parameters for extraction. Further use of the automatic approach is planned on the terminological side, possibly by testing its usefulness in a few other domains. Finally, we aim to explore automatic extraction of information not covered by the existing automatic method. One of such areas is definition extraction; for example, future plans with the Termis database include conducting an experiment on automatic definition extraction from the KoRP corpus, using the recently-developed methodology, specially adapted for Slovene (Pollak 2014).

[9] For example, the automatic data extraction method is an integral part of a proposal for a new dictionary of contemporary Slovene (Krek et al., 2013).

7 References

Arhar Holdt, Š., Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52(2), pp. 95-110.

Cabré Castellví, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology* 9(2), pp. 163–199.

Felber, H. (1984). *Terminology Manual.* Paris: Infoterm.

Fišer, D. (2009). SloWNet – slovenski semantični leksikon. In M. Stabej (ed.) *Infrastruktura slovenščine in slovenistike (Obdobja 28).* Ljubljana: Znanstvena založba Filozofske fakultete, pp. 145–149.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., Baldwin, T. (2013) A lexicographic appraisal of an automatic approach for detecting new word senses In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 49-65.

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 October 2011, 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.

Logar, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: doktorska disertacija. Ljubljana: Filozofska fakulteta.

Logar, N. (2013). *Korpusna terminografija: primer odnosov z javnostmi.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

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.

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 Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425-432.

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

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., 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

Pollak, S. (2014). *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov/Semi-automatic domain modeling from multilingual corpora* (Semi-automatic Domain Modeling from Multilingual Corpora). PhD thesis. Ljubljana: University of Ljubljana, Faculty of Arts, Department of Translation. Accessed at: http://kt.ijs.si/theses/phd_senja_pollak.pdf. [25/03/2014]

Pearson, J. (1998). *Terms in context.* Amsterdam, Philadelphia: John Benjamins.

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.

Sager, J. C. (1998/99). In Search of a Foundation: Towards the Theory of the Term. *Terminology*, 5(1), pp. 41-57.

Vintar, Š. (2010). Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141-158.

Wüster, E. (1931). *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*. Berlin: VDJ.

A Corpus-based Dictionary of Polish Sign Language (PJM)

Jadwiga Linde-Usiekniewicz, Małgorzata Czajkowska-Kisil, Joanna Łacheta, Paweł Rutkowski
University of Warsaw
jlinde@uw.edu.pl, maczajkis@wp.pl,
j.lacheta@uw.edu.pl, p.rutkowski@uw.edu.pl

Abstract

The aim of this paper is to give a general overview of an on-going lexicographic project devoted to Polish Sign Language (PJM), a natural language used by the Deaf community in Poland. The project in question will result in the first-ever PJM dictionary based on extensive corpus data (encompassing more than 300 hours of video material recently collected by the Section for Sign Linguistics of the University of Warsaw). The present article discusses the most important assumptions and methodological foundations of the current PJM dictionary project and confronts them with previous work on PJM and Signed Polish glossaries, as well as with international standards in contemporary sign language lexicography. The design of the new PJM dictionary is discussed in detail, including the most problematic issues, such as lemmatization, sense division and sense description principles. Sample entries are given as illustration. It is important to note that apart from filling an important gap in the availability of sign language teaching and learning materials in Poland, the PJM dictionary outlined in this paper is also likely to further the recognition of PJM as a full-fledged natural language.

Keywords: Polish Sign Language (PJM); corpus-based dictionary; sign language lexicography

1 Introduction

Polish Sign Language (*polski język migowy*, usually abbreviated as PJM) is a natural visual-spatial language used by the Deaf community in Poland. It emerged around 1817, with the foundation of the first school for the deaf in Warsaw, and has been continually in use since then. The current number of PJM users is estimated to exceed 50,000. Despite being one of the largest minority languages in Poland, PJM has not – until recently – attracted much attention from the hearing linguistic community. The first-ever academic unit specializing in research on the grammatical and lexical properties of PJM was created at the University of Warsaw in 2010 (the Section for Sign Linguistics, SSL). The present paper is devoted to a large-scale research project that is currently being developed by the SSL team with the aim of creating a corpus-based dictionary of PJM.

2 Previous PJM lexicography

As happened with most sign languages worldwide (Zwitserlood 2010 and the references therein), early dictionaries of PJM were relatively simple glossaries or wordlists. The first one, published in 1879 and reprinted as Hollak et al. 2011, featured spoken Polish words as headwords, while the body of the entry featured descriptions of how the sign was actually produced or, if the sign language equivalent of a Polish word consisted of a combination of signs, the component signs (described elsewhere in the dictionary) were listed. Thus the entries read as follows (translation ours): "ALLOW – move the hands down in front of oneself and nod seriously"; "WIT – the sign *quick thinking* followed by the sign *wisdom*" (Hollak *et al.* 2011). The dictionary was designed for families of deaf people and for educators. Despite its simplicity, it is a very important source of knowledge on the history of the PJM lexicon.

Again, as in most countries, sign language was banned from Deaf education in Poland for most of the 20th century. The departure from strict oralism in Deaf education and the reintroduction of signs was followed by the creation and establishment of a hybrid artificial code, called the Language-Sign System (*system językowo-migowy*, usually abbreviated as SJM), to be used in Deaf schools as the official language of instruction. SJM was a combination of a set of PJM signs, other artificially created signs and Polish grammar, and is, therefore, often referred to as Signed Polish. Importantly, natural sign language signs were supplemented with signs created as word-formation calques from Polish; the syntax was that of spoken Polish, and in its most elaborate form, it involved finger-spelled Polish morphological affixes.

At that time another dictionary, or rather glossary, was published (Hendzel 1986/2006), in which signs were represented by black-and-white photos of people producing them, with arrows representing movement. The Polish wordlist was based on Bartnicka & Sinielnikoff's (1978) learner's dictionary of Polish. For each sign, spoken Polish equivalents were added, e.g. for a sign referring to friendship the equivalents are Polish lexemes meaning 'friendship', 'friendly', 'friendlily' 'to be friends', 'male friend', 'female friend'. The dictionary presents both true PJM signs and signs that are used in SJM exclusively (cf. Ruta & Wrześniewska-Pietrzak 2013).

The latest general lexicographical publication (Kosiba & Grenda 2011) follows roughly the same principle, though the pictures are in color. Both Hendzel (1986) and Kosiba & Grenda (2011) contain over 2000 signs. Both publications arrange their entries by alphabetical order of Polish words. In the latter publication the headword list was compiled on the basis of various sources, including specific purpose glossaries and phrasebooks, based on both SJM (e.g. Szczepankowski 2000) and PJM (Grzesiak 2008, 2010a,b,c).

In general, for all the lexicographic work mentioned above (Grzesiak's phrasebooks being a notable exception), the lexicographic procedure for eliciting appropriate signs was based on Polish language lexemes, with all the possible dangers and actual errors this procedure may entail (Zwitserlood 2010: 455).

3 Sign language lexicography world-wide

In the meantime, sign language lexicography world-wide has taken a radical turn from glossaries to descriptive or explanatory dictionaries. The explanatory character of such dictionaries has consisted in the meaning of each sign being given not as a spoken language equivalent, but as a sense definition. As such, entries in sign language dictionaries started to conceptually resemble those of monolingual dictionaries of spoken language, though in contrast to true monolingual dictionaries the lexicographic description was not provided in the same language as the headword: thus in the American Sign Language Dictionary (Costello 1998) the sign referring to 'bear' (animal) is identified as a noun and provided with an English definition "[...] a large, heavy mammal with thick, rough fur [...]" (Zwitserlood 2010: 447). This model was also adopted for the Auslan dictionary (Johnston 1989), where the English-language definitions were adopted from the Collins-Cobuild dictionary (Johnston, p.c.). Thus the two explanatory dictionaries were not in fact truly monolingual; a true monolingual sign language dictionary should have the definitions signed (Johnston, p.c.). Yet there is no actual dictionary of a sign language that uses the sign language as the lexicographic metalanguage (Kristoffersen & Troelsgard 2012, Zwitserlood, Kristoffersen & Troelsgard 2013). Both ASL and Auslan dictionaries were originally published as books, with signs represented by drawings. In both publications the headsigns (or lemmas) were ordered not by alphabetical order of English-language equivalents, but by formal features of signs: handshape, hand orientation, sign location, direction of movement (for more detailed discussion of sign representation systems both in printed and on-line dictionaries see Zwitserlood 2010; Kristoffersen & Troelsgard 2010, 2012; Zwitserlood, Kristoffersen & Troelsgard 2013). Nowadays, sign language dictionaries are generally fully electronic, with videos representing headsigns and even sign language examples, as can be seen for instance in the Danish Sign Language Dictionary and Finnish Sign Language Dictionary. Moreover, signs can be accessed not only by formal features but also by topics, and the dictionaries can be used as thesauri as well.

4 Research into PJM – grammar and dictionary

The changes in sign language lexicography are part of a general advance in sign linguistics, brought about by modern video and IT technology that allowed for compilation and analysis of sign language corpora (Crasborn *et al* 2008). A sign language corpus is also being compiled in Poland. The PJM corpus project was launched in 2010 and its first phase will conclude in 2014. The underlying idea is to compile a collection of video clips showing Deaf people (native signers) using PJM in a variety of different contexts. As of early 2014, more than 80 people have already been filmed. When the project is completed, approximately 300 hours of footage will be available for research purposes. The PJM corpus is diversified geographically, covering more than 10 Polish cities with significant Deaf populations. The group of signers participating in the project is well balanced in terms of age, gender, as well as for social and educational background (respective sociological metadata is an integral part of the corpus).

Recording sessions always involve two signers and a Deaf moderator. The procedure of data collection is based on an extensive list of tasks to be performed by the two informants. Typically, the signers are asked to react to certain visual stimuli, e.g. by describing a scene, naming an object, (re-)telling a story, or explaining something to their partner.

The elicitation materials include pictures, videos, graphs, comic strips, etc., with as little reference to written Polish as possible. The participants are also requested to discuss a number of topics pertaining to the Deaf. Additionally, they are given some time for free conversation (they are aware of being filmed but no specific task is assigned to them). The latter two parts of the recording session scenario are aimed at collecting spontaneous and naturalistic data.

The raw material obtained in the recording sessions is further tokenized, lemmatized, annotated, glossed and translated using the iLex software developed at the University of Hamburg (Hanke & Storz 2008). The annotation conventions employed have been designed especially for the purposes of PJM (cf. Rutkowski, Łozińska, Filipczak, Łacheta & Mostowski 2014).

The two basic outcomes of the project are the compilation of a PJM dictionary and a grammar of PJM. The two have to be compatible in terms of underlying theoretical assumptions about sign language linguistics; moreover, it has been decided that the lexicography part of the project has to suit the grammatical description and not vice versa.

The methodology used to produce the PJM dictionary is based on methodologies established for compiling corpus-based spoken language dictionaries. The lexicographers working on the project are native signers, either hearing bilinguals (children of Deaf parents) or Deaf signers with near-native fluency in spoken/written Polish, all of them trained in monolingual lexicography. Thanks to iLex, they have access to all tokens (occurrences) of a particular sign and on the basis of this usage data they establish the meaning of each sign. It is a challenging procedure that needs to take into account various issues such as homonymy versus polysemy, division into syntactic categories, sense division, and adequate sense description, described below.

5 The design of the PJM dictionary

5.1 Target audience and general purpose

Whenever any dictionary is compiled, whether of sign or spoken language, the basic question that has to be addressed is what kind of purpose it is meant to serve and, in consequence, what kind of user group it is aimed at. In terms of purpose, dictionaries can be divided into two types: descriptive or explanatory, i.e. monolingual and bilingual. Importantly, in lexicographic tradition it is the explanatory dictionaries that have always been kept in higher regard: having an explanatory dictionary devoted to it gives a language stature and recognition. This socio-cultural and sociolinguistic aspect of lexicography has lead us to design our dictionary as primarily explanatory. However, since it is not the Polish Deaf community that needs to be made to recognize PJM as a separate language of a linguistic minority, but the general society, compiling a truly monolingual explanatory dictionary of PJM would not only be technically and methodologically feasible, but also would defeat the very purpose of giving PJM its due recognition. That is why we opted for a hybrid explanatory dictionary, i.e. one in which the lexicographic description is given in spoken Polish, as was done in the ASL and Auslan dictionaries mentioned above. In spite of Zwitserlood's (2010: 463-464) criticism of the apparently similar procedure adopted by the publishers of the Dutch Sign Language dictionary (Schermer & Koolhof 2009), we believe that it offers several important advantages.

First of all, the meanings of signs are explicitly described and not left to be inferred from a set of Polish equivalents, which would have been the case had we followed the bilingual-like model that has recently gained popularity in sign language lexicography. There is a general risk of wrong inferences being drawn about the meaning of the source language item from the translation equivalents (Linde-Usiekniewicz & Olko 2006) if semantic equivalents are interpreted as translation equivalents or vice-versa (see Piotrowski 1989 for the distinction, and Piotrowski 1994: 104-155 for more detailed discussion of equivalence in general). While in the Danish dictionary this risk is minimized by linking Danish equivalents to their corresponding entries in a general Danish explanatory dictionary (Kristoffersen, p.c.), this option was not available to us.

Explicit semantic description of sign meanings makes the dictionary a useful source for research into the PJM lexicon and PJM semantics, independently of the researchers' actual proficiency in PJM. It is also a valuable source material for hearing people (native speakers of spoken Polish) learning PJM.

Paradoxically, while the design of the dictionary is such that it gives grounds for the sociolinguistic recognition of PJM, it might appear as if it were of less practical value to the Deaf community. Yet this is not the case. The majority of Polish Deaf are in fact bilingual and use spoken Polish (through lip-reading for speech and in written form). For more technical and academic texts they use Polish language dictionaries, which, in their great majority, are meant for native Polish users. Thus the PJM dictionary in its present form is a kind of a bridge dictionary (Williams 2008), which offers training in the way word senses are being defined in Polish lexicography. Yet since the original bridge dictionaries of English have English lemmas and English Collins-Cobuild definitions translated into the native language of the learners, the PJM dictionary, seen in that light, is actually the reverse combination: sign language lemmas and Polish definitions.

5.2 Directionality and looking-up options

The dictionary is conceived in principle as unidirectional, with sign language signs as lemmas. Nevertheless, when appropriate, for each defined sense, a Polish equivalent or equivalents are provided, and all equivalents are listed alphabetically, so the reverse direction is also available. If several entries share the same Polish equivalent, it is cross-referred to all of them. In the reverse looking-up mode, the definitions provide cues to the appropriate sense of a polysemous or even homonymous Polish word.

PJM signs can be identified by the aforementioned features of handshape, orientation, location and movement, as it is done in all sign language dictionaries.

5.3 Lemmatization: homonymy vs. polysemy

There are two ways in which lemmatization can be carried out in sign lexicography (and in spoken language lexicography as well): purely on formal grounds and on the basis of a combination of both formal and semantic criteria. For reasons of convenience, we opted for purely formal criteria, with no homonymous entries, i.e. with homonyms described in a single entry. Nevertheless, at the entry level homonymy-like phenomena are differentiated from polysemy. In some cases the homonymy is obvious, as the attested meanings fall into discrete bundles of related senses. For example there is a sign that may refer to an ache, pain and related phenomena, as well as to the function of director, directorship, etc. However, since there is a strong tendency for iconicity to be the motivating factor for a sign's signifier, a less obvious example consists of a single sign referring to a crown, and by the same token to a monarch, ruling, etc., as well as to someone having a spherical object on their head (Figure 1). It could be argued that the two groups of senses share the same etymology or motivation so no homonymy is involved, yet on the other hand the semantic relation is nothing but tenuous. In the case of homonymy-like phenomena the entry is split into several sections devoted to macrosenses, i.e. bundles of related senses, as shown in Figure 1. The decision as to whether a given entry should be split or not lies with the lexicographers, who nevertheless often consult other native signers to verify their intuitions about relevant groups of senses being interconnected or not.

5.4 Syntactic issues

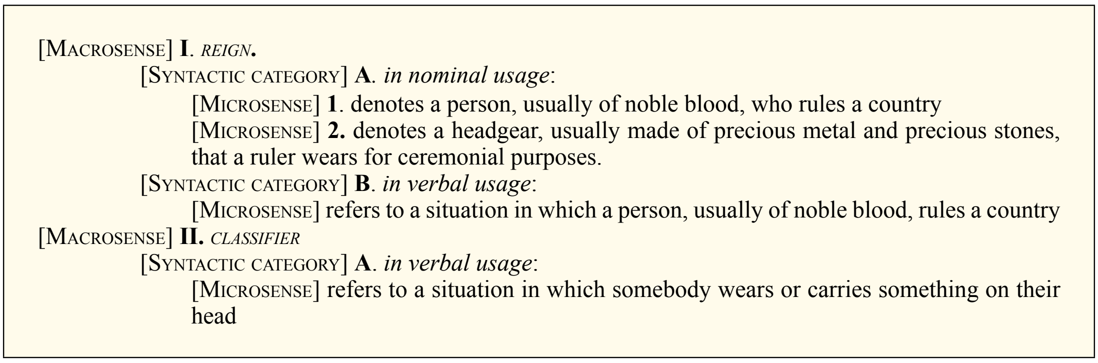
Another challenge lies in classifying senses and usage-types as belonging to a syntactic category. Sign languages, including PJM, have almost no inflection and parts of speech (POS) cannot be established on morphological grounds. Syntactic criteria are not functional enough, as research into PJM to date has tended to show that the same sequence of signs can be interpreted as corresponding to a complex nominalization or to a full clause. Moreover, in contrast to spoken languages for which there are established grammars, the grammar of PJM is being investigated in parallel, on the basis of the same corpus. However, identifying the syntactic character of a sign is necessary in order to provide the most adequate semantic definition of a given sense in Polish. At the same time, the way a given definition is formulated in Polish, i.e. using Polish verbs, nouns or adjectives as key elements, could erroneously suggest the POS of the sign being defined: using infinitives in definitions suggests a uniquely verbal character, using nouns suggests a nominal character, etc. In order to match senses with definitions we decided to eschew POS identification for signs and provide usage-type information instead. The usage-type information reflects the syntactic category identified on a semantic basis (Wierzbicka 2000; Schwager & Zeshan 2008).

Thus the syntactic properties of a sign are introduced in the dictionary as different types of usage (see Figure 1 and 2). The main usage types are named with terms based on the traditional Polish POS system. More importantly, irrespectively of their correspondence or non-correspondence to Polish POS, they are seen as falling into two largely distinct groups: one corresponding to words with predominantly conceptual meaning and the other corresponding to words with predominantly procedural meaning (Wilson 2011). Signs belonging to the first group are described as having: verb-like usage (i.e. referring to an act, an activity, a process or an event or a situation); noun-like usage (i.e. denoting a person, an animal or an object); adjective-like usage (i.e. denoting an entity's feature or property); adverb-like usage (i.e. denoting a feature of an activity or a process); and numeral-like usage, corresponding to cardinal and ordinal numerals. The second group contains pronoun-like usage, preposition-like usage and conjunction-like usage.

Not all signs and sign senses of PJM have a corresponding POS in spoken Polish. Sign languages are known for classifier signs (Zwitserlood 2012) both of nominal and verbal character. While nominal classifier signs mainly serve to identify referents, verbal classifiers usually involve incorporating both the path of movement or direction (if applicable) and the shape of the object the sign refers to. One example of such a classifier sign was the macrosense 'having a spherical object on one's head', the second macrosense of the sign also associated with king, reign and crown (see below).

5.5 Sense division and sense description

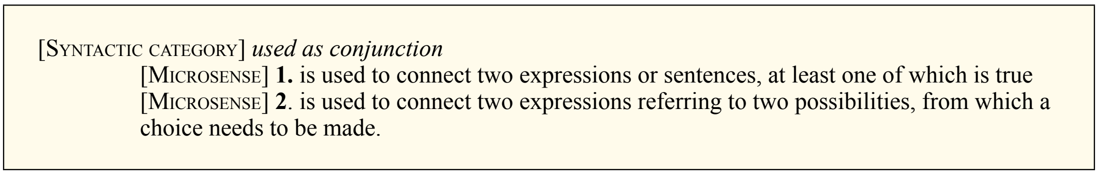
For all the senses associated with conceptual meaning the appropriate definition is not presented as semantic equivalence, as has been done in Costello (1998) and in Johnston (1989), but as a description of the sign denotation or reference. Thus, the appropriate entry fragment for the 'crown/king/rule...' sign is presented in Figure 1:



[Macrosense] **I**. *reign*.
[Syntactic category] **A**. *in nominal usage*:
[Microsense] **1**. denotes a person, usually of noble blood, who rules a country
[Microsense] **2.** denotes a headgear, usually made of precious metal and precious stones, that a ruler wears for ceremonial purposes.
[Syntactic category] **B**. *in verbal usage*:
[Microsense] refers to a situation in which a person, usually of noble blood, rules a country
[Macrosense] **II.** *classifier*
[Syntactic category] **A**. *in verbal usage*:
[Microsense] refers to a situation in which somebody wears or carries something on their head

Figure 1: The entry for king/crown/rule (translated into English).

By contrast, signs with procedural meaning (or procedural senses of a given sign) are described in terms of their function; thus the sign corresponding to the conjunction 'or' is described as shown in Figure 2:



[Syntactic category] *used as conjunction*
[Microsense] **1.** is used to connect two expressions or sentences, at least one of which is true
[Microsense] **2**. is used to connect two expressions referring to two possibilities, from which a choice needs to be made.

Figure 2: The entry for conjunction 'or' (translated into English).

The sense definitions tend to be highly detailed, and at the same time, formulated in a simple language, following the spirit, but not the letter of the Collins-Cobuild project (Sinclair 1987, and particularly Hanks 1987), and its Polish counterpart, Bańko 2000. To underline the fact that the dictionary definitions describe signs and do not provide complete semantic equivalents, the definitions are formulated in terms of specific frames, already mentioned above: thus, signs in noun-like usage *denote* entities further specified in the definition; signs in adjective-like usage *describe* entities as having some property further specified in the definition and signs in verb-like usage *refer* to a situation further described in the definition. For verb-like usages the definition contextually specifies the "subject-like" and the "object-like" syntactic arguments. Directional verbs are specified as such in the syntactic category.

Sense definitions tend to be extended and comprehensive, as we are trying not to use examples either to further specify the range of meaning or to supplement an over-general definition, but to illustrate the sign use. Specifically we will be using examples to show nouns used as modifiers, or being modified by other nouns, since this is a syntactic property of sign languages, including PJM, that is absent in Polish. Adjectives will be illustrated either by attributive or predicative use, and specifically to show if they tend to be used pre- or post-nominally (Rutkowski, Łozińska, Łacheta & Czajkowska-Kisil 2013, Rutkowski, Czajkowska-Kisil, Łacheta & Kuder 2013).

In order to arrive at the actual sense division for each sign the lexicographer consults the recorded corpus, using the iLex software (Hanke & Storz 2008). Since many tokens of the same sign are repeated in recordings of different participants performing the same task (questionnaires, recounting pictures and movies) the lexicographers are not required to check all instances of use in the same context. (This is one of the advantages of working with an elicited corpus, as opposed to spoken language corpora, where corpus search brings numerous repeated instances of the word's use). However, the lexicographers are supposed to check all the instances in which the sign appears in free discourse. Since the lexicographers are themselves native signers with fully developed vocabulary (teachers of PJM and PJM interpreters) they may apply their own knowledge of the sign, confirmed by other deaf consultants, to further develop the entry, thus providing senses not attested in the corpus. On the basis of both corpus-attested usage and not corpus-attested but nevertheless confirmed usage, they establish sense division. In so doing, they are not supposed to be guided by Polish equivalents of the sign. Thus, having a single Polish equivalent does not constitute evidence for a sign having a single sense. On the other hand, having two different translation equivalents in Polish does not constitute evidence for there being two different senses, as is also the case for two spoken languages as represented in bilingual dictionaries (Bogusławski 1995, Linde-Usiekniewicz & Olko 2006, Linde-Usiekniewicz 2011, Lew 2013). The general tendency is for "lumping" as opposed to splitting, i.e. establishing senses conceptually and not by reference or denotation.

6 Video materials

Video clips for lemmas will be recorded in order to provide non-native users with a neat example of the sign production. Variant realizations, usually differing from the basic form by one parameter alone, will also be recorded. Within the body of the entry, some examples of actual use will be provided as clips. These would be chosen from corpus recordings to complement the information explicitly provided in the entry. For example signs with nominal usage and sense defined in the entry will be illustrated in their attributive use (i.e. where they would be translated into Polish as adjectives); ordinary verbs will be exemplified by their use with standard arguments and also in patterns where they would be translated by nominalizations. Examples will also be provided for directional verbs and for classifiers. Though the examples will be initially chosen from the corpus material they will be reproduced and re-recorded in controlled conditions, for greater clarity. Examples will be glossed and accompanied by Polish translations.

7 Other features

The entries will also feature information about geographical restrictions in sign use (if applicable), since some of the signs tend to be used only in some geographically restricted areas (i.e. they are regionalisms). Another feature, meant mainly for non-native signers, is that of cross-reference and comparison. The 'compare' feature will direct the user either to a sign that is produced in a similar way, and therefore may be confused, or to a sign that differs in form but has the same Polish equivalent.

8 Size and coverage

As to the dictionary size, in order to be as comprehensive as possible we plan to include all signs which are represented by more than 5 tokens in the corpus. However, the headsign list will have to be complemented by signs taken from other sources, since the corpus frequency is influenced by the nature of the corpus: it is an elicited corpus, based largely on specific visual stimuli, with signs corresponding to these stimuli largely overrepresented.

9 Concluding remarks

Overall, the main objective of the PJM dictionary project described here is to fill an important gap in the availability of sign language teaching and learning materials in Poland, by providing a dictionary of groundbreaking functionality. Moreover, it is our expectation that the resulting dictionary is likely to further the recognition of PJM as a full-fledged natural language. As such, we have offered some justification herein for the methodological assumptions and choices made in developing this project, as being appropriate for this particular set of circumstances.

10 References

Bańko, M. (ed.) (2000). *Inny słownik języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Bartnicka, B. & Sinielnikoff, R. (1978). *Słownik podstawowy języka polskiego dla cudzoziemców. Warsaw: Wydawnictwa Uniwersytetu. Warszawskiego.*

Bogusławski, A. (1995). Bilingual general purpose dictionary. A draft instruction with commentaries. In J.Wawrzyńczyk (ed.) *Bilingual Lexicography in Poland: Theory and Practice*. Warsaw: Katedra Lingwistyki Formalnej Uniwersytetu Warszawskiego, pp. 15-55.

Costello, E. (1998). *Random House Webster's American Sign Language Dictionary*. New York: Gallaudet University Press.

Crasborn, O., Hanke, T., Efthimiou, E., Zwitserlood, I., , &. Thoutenhoofd, E. (eds.) (2008). Construction and Exploitation of Sign Language Corpora. *3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris.

Grzesiak, I. (ed,). (2008). Minirozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi przydatne w placówkach administracji publicznej. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.

Grzesiak, I. (ed,) (2010a). Minirozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi przydatne w placówkach opieki zdrowotnej. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.

Grzesiak, I. (ed,) (2010b). Minirozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi z zakresu rozwoju zawodowego. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.

Grzesiak, I. (ed,) (2010c). *Piłka nożna. Słowniczek migowo (PJM)-polski, polsko- migowy (PJM).* Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.

Hanke, T. & Storz, J., (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, E. Thoutenhoofd, (eds.) *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELDA.

Hanks, P. (1987). Definitions and explanations. In J. Sinclair (ed.) *Looking up*. London: HarperCollins Publishers Limited, pp. 116-136.

Hendzel, J. (1986/2006). *Słownik polskiego języka miganego*. Olsztyn: Rakiel.

Hollak J., Jagodziński, T., Świderski, T., Twardowska, E., Turkowska, K. & Dutkiewicz D. (2011). *Słownik mimiczny dla głuchoniemych i osób z nimi styczność mających*. Łódź: Polski Związek Głuchoniemych

Johnston, T. (1989*). Auslan Dictionary: a Dictionary of Australian Sign Language (Auslan).* Adelaide: TAFE National Centre for Research and Development.

Kosiba, O &. Grenda, P. (2011). *Leksykon języka migowego*. Bogatynia: Silentium.

Kristoffersen, J. & Troelsgard, T. (2010). The Danish Sign Language Dictionary. In A. Dykstra & T. Schoonheim, (eds.): *Proceedings of the XIV EURALEX International Congress*. Leeuwarden: Fryske Akademy

Kristoffersen, J. & Troelsgard, T. (2012). The Electronic Lexicographical Treatment of Sign Languages: The Danish Sign Language. In S.Granger & M. Paquot, (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-315.

Lew, R. (2013). Identifying, ordering and defining senses, In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing, pp. 284-302

Linde-Usiekniewicz, J. (2011). Polszczyzna w leksykografii dwujęzycznej – dylematy i postulaty. In W. Gruszczyński & L. Polkowska (eds.). *Problemy leksykografii. Historia – metodologia – praktyka*, Kraków: Wydawnictwo Lexis, pp. 105-122.

Linde-Usiekniewicz, J. & Olko, M. (2006). Multilingual dictionaries on-line: reality and perspectives. In V. Koseska-Toszewa & R. Roszko (eds.) *Semantyka a konfrontacja językowa*, Vol. 3. Warsaw: SOW, pp. 43-59

Piotrowski, T. (1989). The bilingual dictionary – a manual of translation or a description of lexical semantics. In Z. Saloni (ed.) *Studia z polskiej leksykografii współczesnej III*, Białystok: Dział Wydawnictw Filii UW, pp. 41-52.

Piotrowski T. (1994). *Problems in Bilingual Lexicography.*, Wrocław: Wydawnictwa Uniwersytetu Wrocławskiego.

Ruta K. & Wrześniewska-Pietrzak M. (2013). Rzecz o nieobecnych. O słownikach polskiego języka migowego (presentation). *IV Glosa do leksykografii polskiej,* Warsaw, September 22 -23, 2013

Rutkowski, P., Czajkowska-Kisil, M., Łacheta, J. & Kuder A. (2013). The Internal Structure of Nominals in Polish Sign Language (PJM): A Corpus-based Study (poster) *Theoretical Issues in Sign Language Research 11 – TISLR 11* London, July 11.

Rutkowski, P., Łozińska, S., Filipczak, J., Łacheta J., & Mostowski, P. (2014). Jak powstaje korpus polskiego języka migowego (PJM), *Polonica* 33.

Rutkowski, P., Łozińska, S., Łacheta, J. & Czajkowska-Kisil M., (2013). Constituent order in Polish Sign Language (PJM), *Theoretical Issues in Sign Language Research 11 – TISLR 11* London, July 11.

Schermer G.M., & Koolhof C. (eds.) (2009). *Van Dale Basiswoordenboek Nederlandse Gebarentaal.* Utrecht: Van Dale.

Schwager, W., & Zeshan, U. (2008), Word classes in sign languages. Criteria and classifications. *Studies in Language*, 32 (3), pp. 509–545.

Szczepankowski, B. (2000). *Słownik liturgiczny języka migowego.* Warsaw: Wydawnictwo św. Jacka.

Sinclair, J. (ed.). (1987). Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. London: HarperCollins Publishers Limited.

Tennant, R. (1998). *The American Sign Language Handshape Dictionary*, Gallaudet University Press

Williams, G.. (2008). A Multilingual Matter: Sinclair and the Bilingual Dictionary, *International Journal of Lexicography,* Vol. 21 , No. 3, pp. 255-266.

Wilson, D. (2011). The conceptual-procedural distinction: Past, present and future. In V. Escandell-Vidal, M. Leonetti & A. Ahern (eds.) *Procedural meaning: Problems and perspectives* Bingley: Emerald Group Publishing Limited, pp. 3-31.

Wierzbicka, A. (2000). Lexical prototypes as a universal basis for cross-linguistic identification of parts of speech. In P. M. Vogel & B. Comrie (eds.), *Approaches to the Typology of Word Classes*. Berlin: Mouton de Gruyter, pp. 285–317.

Zwitserlood, I. (2010), Sign language lexicography in the early 21st century and a recently published Dictionary of Sign Language of the Netherlands, International Journal of Lexicography, Vol. 23 No. 4, pp. 443–476

Zwitserlood, I. (2012)., Classifiers. In R. Pfau, M.Steinbach & B. Woll (eds.) Sign Language: An International Handbook (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science: 37), Amsterdam: De Gruyter Mouton, pp. 158-186.

Zwitserlood, I., Kristoffersen, J. & Troelsgard, T. (2013). Issues in Sign Language Lexicography. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing, pp. 259-283.

Acknowledgements

This research has been supported financially by Poland's National Science Center (*Narodowe Centrum Nauki*) within the project *Iconicity in the grammar and lexicon of Polish Sign Language (PJM)* (grant number: 2011/01/M/HS2/03661). The authors are also grateful to Daniel J. Sax for his help with the preparation of this paper.

Laying the Foundations for a Diachronic Dictionary of Tunis Arabic: a First Glance at an Evolving New Language Resource

Karlheinz Mörth[1], Stephan Procházka[2], Ines Dallaji[2]
[1]Institute of Corpus Linguistics and Text Technology (Austrian Academy of Sciences)
[2]Department of Oriental Studies (University of Vienna)
Karlheinz.Moerth@oeaw.ac.at, stephan.prochazka@univie.ac.at, ines.dallaji@univie.ac.at

Abstract

Arabic lexicography has a long tradition. However, at the time of writing this report, there exist only a very few digital products, let alone products documenting Arabic dialects. Our paper presents the TUNICO project (Linguistic Dynamics in the Greater Tunis Area) and digital language resources which are being produced as part of the project. The TUNICO working group is working on a digital diachronic dictionary of Tunis Arabic which is being compiled as part of a larger linguistic endeavour to document the variety of the Tunisian capital. One of the interesting features of the project is that it draws on a number of heterogeneous sources: text books, grammatical descriptions and a corpus of spoken youth language which is currently being compiled. In this project, the dictionary is used as an analytical tool, as a research instrument, by integrating the various sources into one new coherent language resource thus allowing researchers to gain unprecedented insights in material that partly has been available for quite some time.

Keywords: eLexicgraphy; diachronic lexicography; lexicography tools

1 Introduction

The compilation of the diachronic dictionary of Tunis Arabic has been started as part of a larger project investigating the linguistic dynamics caused by recent demographic and social changes in the metropolitan area of Tunis (hence Tunis Arabic and not Tunisian Arabic). The TUNICO project (funded by a three year grant of the *Austrian Science Fund*[1]) will produce two digital language resources: (a) a corpus of unmonitored speech (dialogues as well as narratives) and (b) a dictionary based on this corpus and on other historical sources published in print form.

TUNICO in turn has grown out of an ongoing cooperative project which goes by the name *Vienna Corpus of Arabic Varieties* (VICAV). VICAV has been already started several years ago and is being run with a twofold perspective in mind: proceeding from linguistic research questions VICAV has been designed

[1] Project number P 25706-G23

as a forum for collecting, producing and making available digital language resources of a wide range of spoken Arabic varieties. In addition, the project also pursues text technological interests investigating relevant standards, developing tools and workflows. At the heart of the VICAV collection there are so-called language profiles. This type of text consists in concise sketches of spoken linguistic varieties. The intention has been to proceed in a complementary manner to similar endeavours (such as the *Encyclopedia of Arabic Language and Literature* which is a standard reference work in the field). For the time being, the concept does not foresee the production of detailed grammatical descriptions. The focus is rather put on general information, research histories, relevant literature and sample texts. Another language resource represented in the collection are lists of salient grammatical features. The working group has attempted to identify particular linguistic items that are repeatedly used elsewhere in comparative Arabic investigations and make them comparable in example sentences that are the same across the various linguistic varieties. There are also digital dictionaries, texts, bibliographies, descriptions of relevant workflows and best-practises such as thorough encoding guidelines that can be reused for similar purposes in other projects. VICAV is intended as a cross-disciplinary platform for researchers in the field enabling them to exchange data, to collaborate effectively on new digital resources and to publish their findings, tools and data.

Both projects, TUNICO and VICAV, are joint initiatives of the University of Vienna (Department of Oriental Studies) and the Austrian Academy of Sciences (Institute for Corpus Linguistics and Text Technology). They are typical examples of a new brand of research that understands itself as digital humanities pursuing research with innovative methods and in accordance with new paradigms such as collaborative work, transdisciplinarity and open humanities.

2 A New Digital Dictionary

In the history of lexicography, dictionaries documenting Arabic dialects are a rather recent phenomenon. While the situation with respect to print dictionaries has improved for many areas, there are almost no digital Arabic dictionaries available so far, let alone dictionaries that come in a digitally reusable form, live up to modern standards or cover varieties other than Modern Standard Arabic.

With respect to Tunis, the situation is no different. There exists no comprehensive dictionary of the Arabic dialect of Tunis. Nicolas 1911 can be regarded as a good basis for diachronic research. However, it is – by and large – outdated. Other sources for lexicographic data are the works of Beaussier/Lentin (2006, a fusion of the 1958 edition and the 1959 supplement) which also include data on Tunis, Quéméneur (1961 and 1962) who provided useful lists of lexicographical items and Abdellatif 2010 who produced a quite useful amateur glossary. The eight volume glossary compiled by Marçais/Guîga 1958-61 covering the vocabulary of the village of Takroûna (ca.100 km south-east of Tunis) is still of unmatched value for the documentation of the lexicon of the Arabic vernacular of the Tunis area. However, it reflects mainly rural speech and is based on material from the 1920s.

Our project was designed to create up-to-date and easily accessible lexical information on Tunis Arabic, taking into account historical as well as contemporary data, by compiling a small, micro-diachronic and machine-readable dictionary of the variety. One of the many advantages of such a machine-readable dictionary is that queries in both directions (in our case Tunis Arabic – English/German/French and vice versa) are possible. All hitherto published dictionaries except the outdated work by Nicolas (1911) are unidirectional Arabic – French.

3 Heterogeneous Sources

One innovative aspect of the project lies in the fact that it is not only drawing on contemporary data taken from a digital corpus. However, it will also incorporate various sources reaching back as far as the 19th century (Stumme 1893/1896a-b). By integrating both corpus data and historical sources, we will create a new language resource, a new dictionary. Technically, the intention has been to keep each bit of information added to the dictionary traceable to its origin, thus allowing coming generations of researchers to interpret the data in accordance with their particular needs.

The basis of the dictionary was laid by data taken from didactic materials that were compiled by participants in the project for university classes. The glossaries of this course of spoken Tunisian Arabic could be easily recycled for the purpose and transformed into digital dictionary entries. In the next phases of the project, this data will be enriched from three main additional sources: the newly created corpus, interviews with first language speakers, and historical publications on the linguistic variety under investigation.

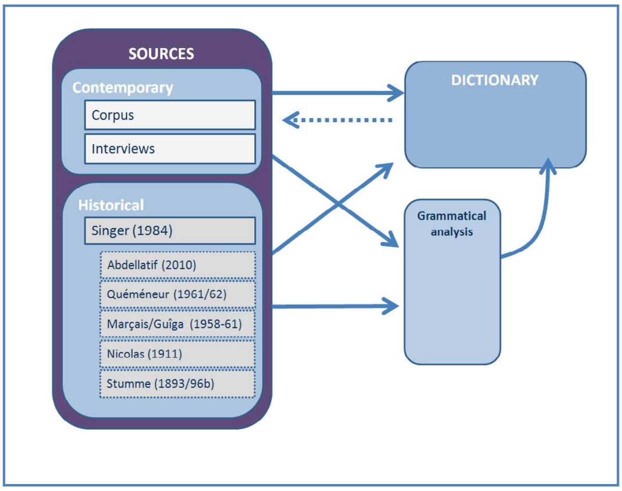


Figure 1: Basic chart of used sources.

3.1 Corpus of Spoken Youth Language

As can be seen in figure 1, the contemporary sources consist in a corpus of youth language by which we understand language produced by people under the age of 25. In Arabic dialectology, focussing on the language of the younger generation is a rather uncommon approach as traditionally linguistic interests were usually directed towards the past. Dialectological investigations often been focussed on the language of the older generation attempting to gain knowledge on older more conservative "pure" linguistic varieties. This has led to a situation (and not only in Arabic studies) where we often only know older forms of particular varieties and comparatively little about contemporary language. TUNICO (and also other language resources of the VICAV collection) has been designed to remedy this shortfall focussing on modern language, contemporary usage and lexical neologisms which, in the case of Tunis, are most often not of Arabic but French (or English) origin.

A first set of data was collected in August and September 2013 and is currently being transcribed. The field workers recorded some 33 hours of Tunis Arabic, the recordings contain data from approximately 90 different interviewees. The final version of the dictionary will contain the most frequent lemmas represented in the transcribed corpus, which will constitute the foundation of the dictionary's contemporary layer.

3.2 Additional Interviews

The second contemporary source, the interviews, will be created at a later stage of the project. As we expect numerous lacunae in the lemma list and constituents of the dictionary entries, the data gained from the corpus and also the historical sources will have to be completed with information gained from additional interviews conducted with Tunisian informants. This will be done during the third and the last field campaigns. The interviews will differ from those done in the first two campaigns as they will be semi-structured interviews that aim at the elicitation of lexical data absent in the corpus. Having collected plenty of dialogues, it is planned to also go for narratives in this phase of the project.

3.3 Historical Data

The historical aspect will be introduced by way of lexicographic items excerpted from print publications, especially the very rich lexical material contained in Singer's monumental grammar (1984; almost 800 pages) of the Medina of Tunis, which hitherto has been difficult to use which is mainly due to a lack of an index[2]. Singer's data will be evaluated systematically and integrated into the dictionary (the material will, of course, be indicated by reference to his book; however on account of the unclear

[2] It is important to note that Singer's study is based on fieldwork carried out in the early 1960s (Singer 1984: VIII); the texts and the glossary which were advertised in the foreword (p.X) have never been published.

copyright status it is not planned to create a digital version of the book itself). Additional resources (Nicolas 1911, Marçais/Guîga 1958-61, Quéméneur 1962, Abdellatif 2010) will also be consulted in order to verify and complete the collected data. The diachronic dimension will help to better understand processes in the development of the lexicon.

The rich material gathered from young people whose parents are often not natives of the city of Tunis but have migrated to the capital from rural regions or other cities of Tunisia will hopefully enable us to analyse recent developments in the lexicon, particularly semantic changes including semantic shift, semantic reduction, and semantic extension of lexical items. Our diachronic approach will also make possible to determine the influence of other dialectal varieties of Arabic which in many cases are a result of the pan-Arabic satellite channels. One focus of the interviews carried out during our fieldwork is to gain newly coined vocabulary that appeared during and after the revolution of 2011, which had an immense impact on Tunisian society and hence also on Tunisian language. This vocabulary is, however, different from real youth slang that is often only used in in-group conversation. Studies on this particular field of the lexicon of Arabic dialects are extremely rare (the best publication on this topic is Caubet 2004 dealing mostly with Morocco).

As we are not mainly interested in the "pure and real" dialect we will also pay attention to the incorporation of foreign elements into the Arabic language as spoken in Tunis both with regard to semantics and morphology. The latter is characterized by a high degree of integration into the morphological frame of Arabic. Particularly verbs of foreign origin have to be adapted to Arabic patterns for the sake of inflection. A similar development is often found with pluralisation of nouns and adjectives.

The direct connection of corpus and dictionary (see figure 1) will facilitate research on phrases, idioms and collocations. Apart from some very well-studied varieties of Arabic (especially Egyptian and Moroccan Arabic) phraseology and related subjects such as collocation have been largely neglected, mainly because of a lack of text corpora sizeable enough for these purposes. The linkage of dictionary and corpus will enable users to investigate in which way given lexical items are connected to one another.

4 Modelling the Dictionary Entries

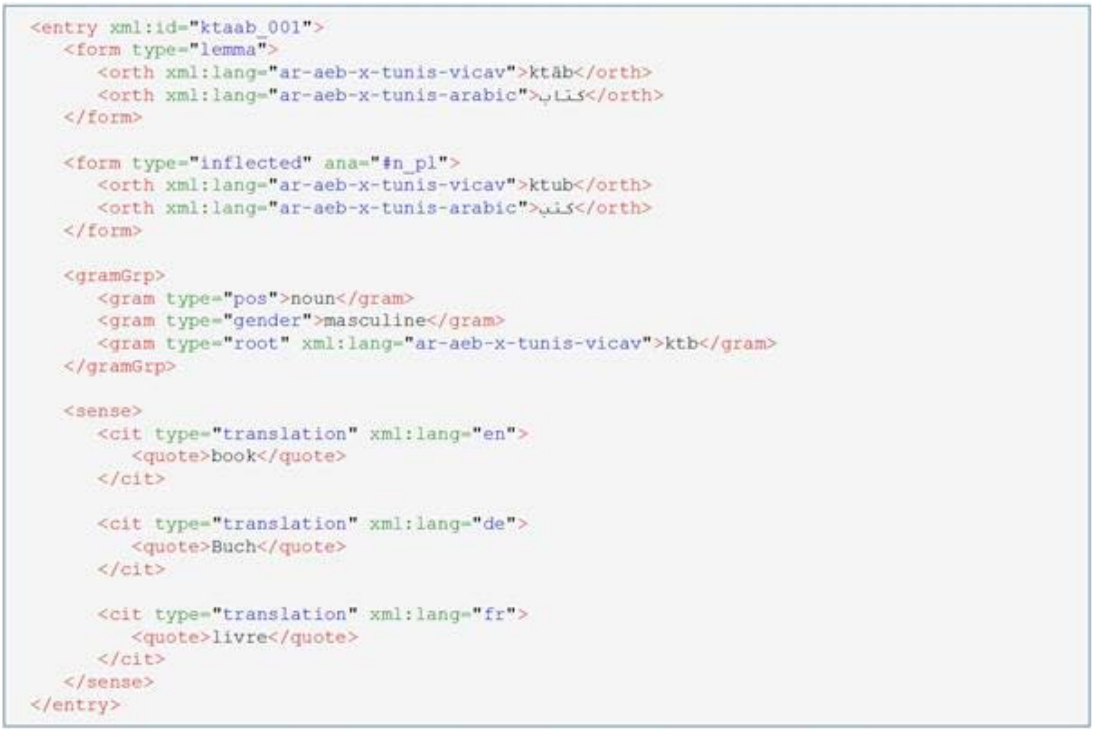
In the world of digital dictionary production, a considerable number of competing formats co-exist. We are far from any real standardisation in the field and our paper will not resume the discussion as to which format is best suited for which task (cf. Budin et al. 2012). Let it suffice to state, that using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. However, it has been shown that the TEI dictionary module is also usable for NLP purposes (Budin 2012). Data modelling for our project has been undertaken with two perspectives in mind: (a) to achieve a high degree of interoperability with comparable dictionaries of other varieties of Arabic (already existing at the same department) and (b) to stay as compliant as possible with the ISO standard LMF (Romary 2013).

A major issue in this endeavour is how to represent and how to harmonise the diverging systems of transcription and transliteration found in the historical sources. As in comparable other projects, researchers in our project try to reduce the rich inventory of combinations of diacritics and characters by applying a basically phonemic transcription. Following a widespread convention in Arabic dialectology, the data is presented in a broad phonological transcription that does not usually indicate allophones. Basically, the set of characters used in the dictionary follows widely used conventions in Arabic studies. It is by and large the system used in standard reference works such as the *Encyclopedia of Arabic Language and Linguistics* (Leiden: Brill, 2006-2009). In the future, it is planned to provide the data in IPA-transcription too.

As can be seen in figure 2 we indicate the so-called root for each lexical item in the dictionary. The root is an intrinsic feature of Arabic word formation. In all layers of Arabic the bulk of the vocabulary is built on the principle of root and pattern. To express certain semantic terms, i.e. words, a purely consonantal root carrying the basic semantic information is combined with a limited set of patterns utilizing a fixed sequence of consonants, vowels, and optional prefixes and suffixes. To make comparative cross-dialectal search possible we have decided to indicate the root in a strictly etymological way. This means, each root reflects the corresponding root of Standard Arabic wherever possible. We are convinced that this approach does not reduce the usability of the on-line dictionary because, for users familiar with Arabic morphology, it is easy to detect the dialectal root in question. The main advantage of this approach lies in the possibility to find the reflexes of a certain Standard Arabic root in all dictionaries simultaneously.

4.1 Basic Schema

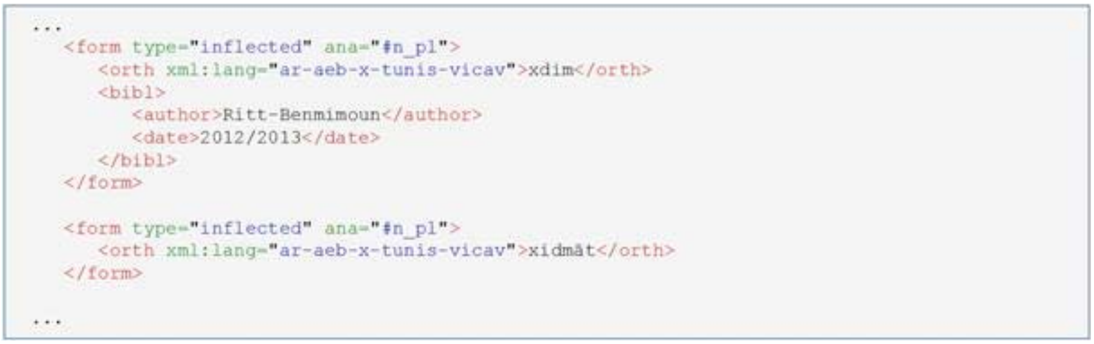
The schema applied in the compilation of the new dictionary has been used before in other projects for various languages and serves as the structural foundation of the dictionary entries. It imposes a number of very strict structural constraints on the TEI elements to ensure a high degree of interoperability with other components of the existing dictionary infrastructure at the ICLTT (Budin 2012). These constraints are defined by means of an XML Schema which only allows the use of a small subset of TEI elements and only a very few combinations thereof. The typical, slightly simplified basic structure of an entry taken from the Tunis dictionary is shown below. The entry begins with the lemma, this is followed by morphological forms and grammatical information. The system provides for translations in several languages. In the Tunis dictionary, we intend to offer German and English translations. Resources permitting, we will also add French.



```
<entry xml:id="ktaab_001">
   <form type="lemma">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktāb</orth>
      <orth xml:lang="ar-aeb-x-tunis-arabic">كتاب</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
      <orth xml:lang="ar-aeb-x-tunis-arabic">كتب</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="gender">masculine</gram>
      <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">ktb</gram>
   </gramGrp>

   <sense>
      <cit type="translation" xml:lang="en">
         <quote>book</quote>
      </cit>

      <cit type="translation" xml:lang="de">
         <quote>Buch</quote>
      </cit>

      <cit type="translation" xml:lang="fr">
         <quote>livre</quote>
      </cit>
   </sense>
</entry>
```

Figure 2: Basic encoding of a typical dictionary entry.

As can be seen in the example above, *sense* elements can have multiple translations. In a similar manner, every *entry* can contain an unspecified number of form elements. These can represent different morphological forms, variants such as for instance competing plurals or varying phonological representations. All these cases are treated similarly. Hierarchies are avoided, all *form* elements are placed directly inside the *entry* element.



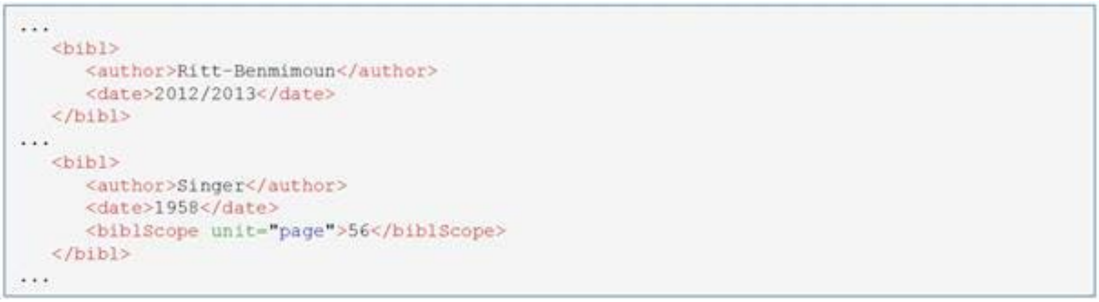
```
...
   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">xdim</orth>
      <bibl>
         <author>Ritt-Benmimoun</author>
         <date>2012/2013</date>
      </bibl>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">xidmāt</orth>
   </form>

...
```

Figure 3: Overabundance in plural forms.

The two inflected forms represent both common plurals. In cases of a clear distribution across registers, labels can be used to assign information regarding the register of the particular form. However, for the time being such forms are merely collected without adding information concerning the formality scale. Once the corpus is available, we intend to add frequency data to the lemmas and the inflected forms. As to the encoding of the frequency information, discussions concerning data modelling are still ongoing.[3]

Modelling Diachrony

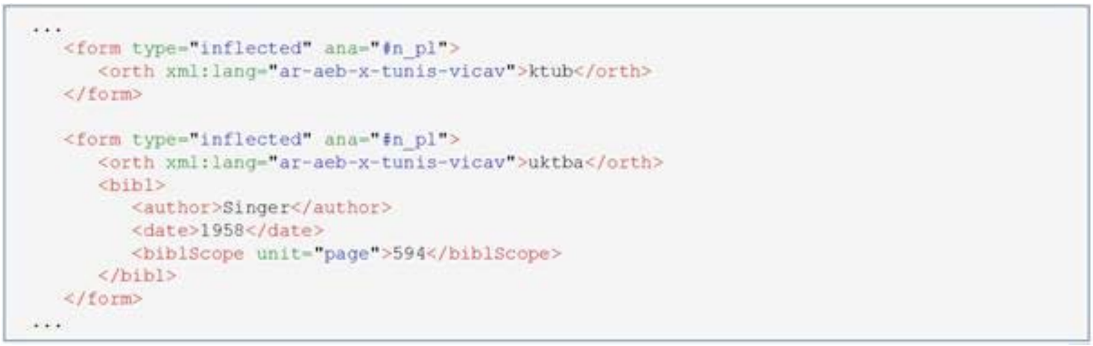
Diachrony, or as some might insist micro-diachrony (as we are only talking about a time-span of roughly a century), is represented in the dictionary by indicating the source from which the data was taken. To this end, we make use of the *bibl* (bibliographic citation) element. This is a loosely-structured element the sub-components of which may or may not be explicitly tagged (TEI Guidelines 2013).



```
...
   <bibl>
      <author>Ritt-Benmimoun</author>
      <date>2012/2013</date>
   </bibl>
...
   <bibl>
      <author>Singer</author>
      <date>1958</date>
      <biblScope unit="page">56</biblScope>
   </bibl>
...
```

Figure 4: Bibliographic citations in TEI (P5).

Diachrony is established by adding these *bibl* elements to *form* and/or *sense/cit* elements. As stated before, any entry can have multiple forms and also can have multiple instances of the same morphological form. The absence of a *bibl* element indicates that the form has been entered from contemporary sources. In this manner, each element can be historically classified. In the following example, the *entry*, a noun, has two plural forms. By contrast to the example above which displays synchronous data (*xdim* vs. *xidmāt* are both still in use) the second form here represents evidence of a historic form. An analogous contemporary form could so far not be verified.



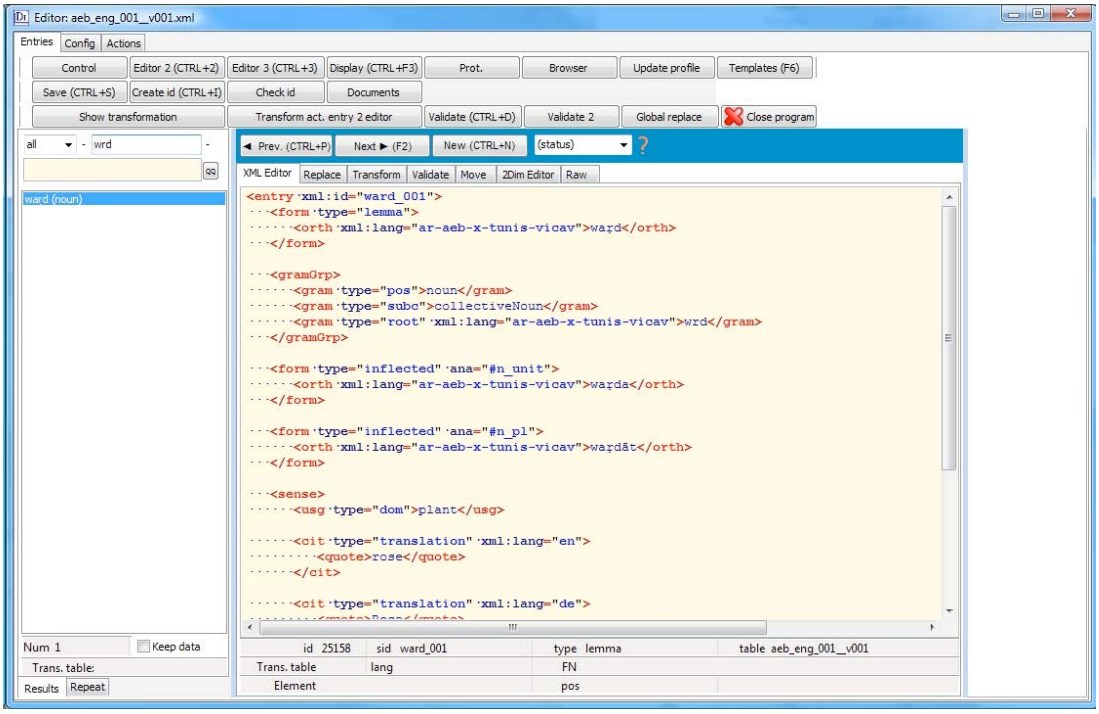
```
...
   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">uktba</orth>
      <bibl>
         <author>Singer</author>
         <date>1958</date>
         <biblScope unit="page">594</biblScope>
      </bibl>
   </form>
...
```

Figure 5: *ktub* vs. *uktba*.

[3] Details of these discussions were presented in the TEI members Meeting 2014 (Rome) and have been submitted for publication in jTEI 8 (papers from the 2013 conference).

5 Tools

The dictionary entries are compiled making use of the *Viennese Lexicographic Editor* (VLE), a general purpose XML editor providing a number of functionalities typically needed in compiling lexicographic data. It allows to collaboratively work on lexicographic data. From the very beginning of its development, it was designed to process standard-based lexicographic and terminological data such as LMF, TBX, RDF or TEI. VLE can automate many editing procedures. Most of these functions can be applied both to single and/or multiple entries. VLE has been implemented as a standalone desktop application (for Windows). VLE is the client of a server-client architecture. In order to realise a working environment, a web-server is needed. The server-side scripts (*php* + *mysql*) are also freely available and easy to setup. The program can check the structural integrity (well-formedness) of input on the fly and can validate the data against XML Schemas.
One of the particular features of VLE is a special module optimising access to external language resources such as corpora, other dictionaries, word lists etc. which makes it particularly well suited for deployment in our project with its various digital resources.[4] The fact that VLE is a product of our institute and constantly being updated will ease the implementation of necessary interfaces between the corpus and the dictionary infrastructures.



```
<entry xml:id="ward_001">
   <form type="lemma">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛd</orth>
   </form>

   <gramGrp>
      <gram type="pos">noun</gram>
      <gram type="subc">collectiveNoun</gram>
      <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">wrd</gram>
   </gramGrp>

   <form type="inflected" ana="#n_unit">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛda</orth>
   </form>

   <form type="inflected" ana="#n_pl">
      <orth xml:lang="ar-aeb-x-tunis-vicav">waṛdāt</orth>
   </form>

   <sense>
      <usg type="dom">plant</usg>

      <cit type="translation" xml:lang="en">
         <quote>rose</quote>
      </cit>

      <cit type="translation" xml:lang="de">
         <quote>Rose</quote>
```

id 25158	sid ward_001	type lemma	table aeb_eng_001__v001
Trans. table	lang	FN	
Element		pos	

Figure 6: The dictionary editor.

[4] The tools can be downloaded from the *Language Resources Portal* (*CLARIN Centre Vienna*): http://clarin.oeaw.ac.at/ccv/vle.

The online publication of both the corpus and the dictionary will be undertaken by means of *corpus_shell*, a modular framework of reusable software components which has also been developed at the ICLTT over the past couple of years. It is used to access and publish heterogeneous and distributed language resources such as language corpora, dictionaries, encyclopaedic databases, prosopographic databases, bibliographies, metadata, and schemata. Its core functionality is encapsulated in self-contained components exposing well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components.[5] This software was used in several other projects before, it is the backbone of the *Language Resources Portal*[6] which is run at the Austrian Academy of Sciences.

6 Status and Outlook

The project is being conducted in the context of CLARIN-AT, the local branch of the European infrastructure consortium CLARIN-ERIC (Common Language Resources and Technology Infrastructure). Both the corpus and the dictionary were planned as in-kind contributions to the CLARIN network for the years to come. The build-up of the corpus, the compilation of the dictionary and the development of software are being undertaken in the spirit of open-access and open-source. So far, no binding decision has been made as to the licence under which the particular language resources will be available. However, there is a strong case for a Creative Commons licence, CC-BY being the favoured option, which has been used for comparable other projects of the department. Discussions with interested researchers and other stakeholders have shown that the permission to create derivative works is usually regarded as an important prerequisite in order to ensure reuse of data. The tools, workflows and specifications created in this project can potentially also be used for other languages and many other applications.

At the time of writing this paper, the dictionary already contained roughly five thousand raw entries and several hundred edited entries. We are planning to make data available as soon as data from the corpus and historical sources have been added to the basic entries, i.e. already during the production phase which is meant to allow and to encourage other researchers in the field to contribute to this work. To our knowledge, this lexicographic undertaking is not only the first scholarly attempt to make available a digital dictionary of a spoken Arabic variety, but also the first attempt at creating a digital dictionary presenting diachronic data of a spoken Arabic variety.

[5] More details at clarin.oeaw.ac.at/ccv/corpus_shell.

[6] clarin.oeaw.ac.at/ccv/

7 Selected References

Abdellatif, K. (2010). Dictionnaire «le Karmous» du Tunisien. Accessed at: http://www.fichier-pdf.fr /2010/08/31/m14401m/ [06/04/2014].

Baccouche, T., Mejri, S. (2000). L'Atlas Linguistique de Tunisie: problématique phonologique. In *Revue Tunisienne de Sciences Sociales* 120, pp. 151-156.

Banski, P., Wójtowicz, B. (2009). FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries. In TEI-MM, Ann Arbor. Accessed at: http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz- TEIMM-presentation.pdf [06/04/2014].

Beaussier, M. (2006). *Dictionnaire pratique arabe-français: (arabe maghrébin); constitué du «Dictionnaire pratique arabe-français» de Marcelin Beaussier dans l'édition de Mohamed Ben Cheneb & de son «Supplément» par Albert Lentin,* Paris: Ibis Press.

Bel, N., Calzolari, N. & Monachini, M. (eds.) (1995). Common Specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTEXT Deliverable D1.6.1B. Pisa.

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 3 (Special issue on TEI and linguistics).*

Hass, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz. Berlin, New York: W. de Gruyter.

Ide, N., Kilgarriff, A. & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In *Euralex 2000 Proceedings*, pp. 113-126.

Marçais, W. (1925). Textes arabes de Takroûna. I. Textes, Transcription et Traduction annotée. Paris.

Marçais, W., Guîga, A. (1958-61). *Textes arabes de Takroûna. II: Glossaire*. 8 vol. Paris.

Mörth, K., Budin, G. (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 (Slovenia), pp. 52-59.

Nicolas, A. (1911). Dictionnaire français-arabe: idiome tunisien and Dictionnaire arabe-français. Tunis.

Quéméneur, J. (1962). Glossaire de dialectal. In *IBLA* (1962), pp. 325-67.

Romary, L., Salmon-Alt, S. & Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on enhancing and using electronic dictionaries.* Coling 2004, Geneva.

Romary, L., Wegstein, W. (2012). Consistent Modelling of Heterogeneous Lexical Structures. In *Journal of the Text Encoding Initiative 3 (Special issue on TEI and linguistics).*

Romary, L. (2013). TEI and LMF crosswalks. In *Stefan Gradmann and Felix Sasaki (eds.), Digital Humanities: Wissenschaft vom Verstehen.* Humboldt Universität zu Berlin. Accessed at: http://hal.inria.fr/hal-00762664 [08/03/2014].

Singer, H. (1984). Grammatik der Arabischen Mundart der Medina von Tunis. Berlin-New York.

Sperberg-McQueen, C.M., Burnard L. & Bauman S. (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlotteville, Nancy. Accessed at: http://www.tei-c.org/Guidelines/P5/ [08/03/2014].

Stumme, H. (1893). Tunisische Märchen und Gedichte. Band I: Transcribierte Texte nebst Einleitung; Band II: Übersetzung. Leipzig.

Stumme, H. (1896a). Grammatik des tunisischen Arabisch nebst Glossar. Leipzig.

Stumme, H. (1896b). Neue tunisische Sammlungen. (Kinderlieder, Strassenlieder, Auszählreime, Rätsel, 'Arôbi's, Geschichtchen u.s.w.). Berlin (ZAOS II).

Stumme, H. (1898). Märchen und Gedichte aus der Stadt Tripolis in Nordafrika. Eine Sammlung prosaischer und poetischer Stücke im arabischen Dialekt der Stadt Tripolis, nebst Übersetzung, Skizze des Dialekts und Glossar. Leipzig.

Versteegh, K., Eid, M., Elgibali, A., Woidich, M & Zaborski, A. (eds.) (2005-2009). *Encyclopedia of Arabic Language and Linguistics*. 5 vols. Leiden, Boston: Brill

BabelNet meets Lexicography: the Case of an Automatically-built Multilingual Encyclopedic Dictionary

Roberto Navigli
Sapienza University of Rome
navigli@di.uniroma1.it

Abstract

In this paper we provide a first study of the lexicographic quality of BabelNet, a very large automatically-created multilingual encyclopedic dictionary. BabelNet 2.0, available online at http://babelnet.org, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of several language resources, namely: WordNet, Open Multilingual WordNet, Wikipedia and OmegaWiki. Here we present a first analysis of the dictionary entries in terms of their coverage of English and Italian word tokens in a large corpus and in comparison to existing, well-established dictionaries, namely the Oxford Dictionary of English and the Treccani Italian dictionary. We observe that BabelNet contains most meanings of the frequent words under analysis and provides additional, often domain-specific meanings and their textual definitions unavailable in traditional dictionaries, as well as encyclopaedic coverage for those words.

Keywords: Multilinguality; Encyclopedic dictionaries; Quality evaluation of automatically-created dictionaries

1 Introduction

The textual content that is available on the Web is becoming ever increasingly multilingual, providing an additional wealth of valuable information. Most of this information, however, remains inaccessible to the majority of users because of language barriers. Consequently, both humans and automatic systems need tools which will enable them to enjoy the beauty and the usefulness of this varied multilingual world.

The wide majority of bilingual paper dictionaries, however, focus on a given language pair, which are the languages on which the lexicographers, and authors of the dictionary, are expert in. As a result, the sense inventories of dictionaries for different language pairs are different, even if the dictionaries are printed by the same publisher. Integrating these inventories, thereby enabling the creation of a multilingual dictionary, is therefore a very arduous task.

MultiJEDI (Multilingual Joint word sensE Disambiguation, http://multijedi.org) is a major project under way in the Linguistic Computing Laboratory at the Sapienza University of Rome. MultiJEDI is a 5-year Starting Independent Research Grant funded by the European Research Council (ERC) that started in February 2011. The project aims to investigate new, groundbreaking directions in the field of Word Sense Disambiguation (WSD), the task of computationally determining the meaning of words in context (Navigli, 2009; 2012). The key intuition underlying the project is that we now have the capabilities to transform multilinguality from an obstacle to Natural Language Understanding into a powerful catalyst for the task. As a core tool for enabling multilinguality the project aims to create a very large automatically-created multilingual encyclopedic dictionary, called BabelNet, made available online at http://babelnet.org. BabelNet is a novel language resource in several respects, including: being a multilingual dictionary which covers tens of languages; providing both encyclopaedic and lexicographic coverage; including information which is usually not available within dictionaries, such as images, fine-grained category information, multiple textual definitions for the same entry, hyperlinks to other entries, and much more.

Since integrating dictionaries of different kinds and nature, especially on a multilingual scale, is admittedly a hard, ambitious task, in this paper we analyze the lexicographic quality of BabelNet, especially in terms of the user perspective, and compare it against manually created dictionaries, so as to determine the added value of an automatic dictionary integration process. Our analysis is performed both at the corpus level, by studying the coverage provided by BabelNet of word occurrences within text (on a portion of the American National Corpus – ANC), and at the inventory level, i.e. by comparing the BabelNet sense inventory with that of other well-established resources, such as the Oxford Dictionary of English and the Treccani dictionary of Italian. Our analysis shows that the richness and amount of information available in BabelNet largely exceeds that of manually created lexicographic resources.

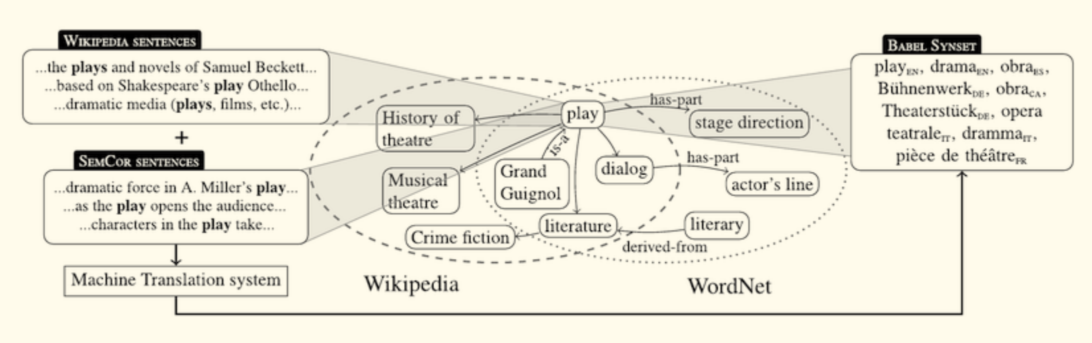

WIKIPEDIA SENTENCES
...the plays and novels of Samuel Beckett...
...based on Shakespeare's play Othello...
...dramatic media (plays, films, etc.)...
+
SEMCOR SENTENCES
...dramatic force in A. Miller's play...
...as the play opens the audience...
...characters in the play take...
Machine Translation system
History of theatre
play
has-part
stage direction
is-a
Musical theatre
Grand Guignol
dialog
has-part
actor's line
literature
literary
Crime fiction
derived-from
Wikipedia
WordNet
BABEL SYNSET
play$_{EN}$, drama$_{EN}$, obra$_{ES}$,
Bühnenwerk$_{DE}$, obra$_{CA}$,
Theaterstück$_{DE}$, opera teatrale$_{IT}$, dramma$_{IT}$,
pièce de théâtre$_{FR}$

Figure 1: The BabelNet structure.

2 BabelNet 2.0

BabelNet is based on the key idea that different language resources, such as WordNet (Fellbaum, 1998), i.e., the largest machine-readable computational lexicon of English, and Wikipedia (http://www.wikipedia.org), i.e., the most popular multilingual encyclopedia, provide complementary knowledge that can be integrated into a single unified multilingual semantic network covering as many languages as possible. BabelNet, available online at http://babelnet.org, is therefore a large-scale "encyclopedic dictionary". BabelNet encodes knowledge as a labeled directed graph G = (V, E) where V is the set of nodes – i.e., concepts such as *play* and named entities such as *Shakespeare* – and E ⊆ V × R × V is the set of edges connecting pairs of concepts (e.g., *play* is-a *dramatic composition*). Each edge is labeled with a semantic relation from R, e.g., {is-a, part-of , ..., ε}, where ε denotes an unspecified semantic relation. Importantly, each node v ∈ V contains a set of lexicalizations of the concept for different languages, e.g., { play_{EN}, $\text{Theaterstuck}_{\text{DE}}$, $\text{dramma}_{\text{IT}}$, obra_{ES}, ..., piece de $\text{theatre}_{\text{FR}}$ }. We call such multilingually lexicalized concepts Babel synsets. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia. In order to construct the BabelNet graph, we extract at different stages:

- from WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);
- from Wikipedia, all the Wikipages (i.e., Wikipages, as concepts) and semantically unspecified relations from their hyperlinks.

A graphical overview of BabelNet is given in Figure 1. As can be seen, WordNet and Wikipedia overlap both in terms of concepts and relations: this overlap makes the merging between the two resources possible, enabling the creation of a unified knowledge resource. In order to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

(1) We integrate WordNet and Wikipedia by automatically creating a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories of concepts to complement each other.

(2) We collect multilingual lexicalizations of the newly-created concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (i.e., the inter-language links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.

(3) We create relations between Babel synsets by harvesting all the relations in WordNet and in the wikipedias in the languages of interest.

Its current version, i.e., BabelNet 2.0, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of the following resources:

- WordNet, a popular computational lexicon of English (http://wordnet.princeton.edu, version 3.0);
- Open Multilingual WordNet (http://www.casta-net.jp/~kuribayashi/multi/), a collection of wordnets available in different languages;
- Wikipedia, the largest collaborative multilingual Web encyclopedia (http://wikipedia.org);
- OmegaWiki, a large collaborative multilingual dictionary (http://omegawiki.org).

The number of lemmas for each language ranges between more than 8 million (English) and almost 100,000 (Latvian), with a dozen languages having more than 1 million lemmas. The number of polysemous terms ranges between almost 250,000 in English to only a few thousand for languages such as Galician, Latvian and Esperanto, with most languages having several tens of thousands of polysemous terms. BabelNet 2.0 contains about 9.3 million concepts, i.e., Babel synsets, and above 50 million word senses (regardless of their language). It also contains about 7.7 million images and almost 18 million textual definitions, i.e., glosses, for its Babel synsets. The synsets are linked to each other by a total of about 262 million semantic relations (mostly from Wikipedia). Language distribution of lemmas, synsets and senses in graphically shown in Figure 2. It can be seen that the top 9 languages cover approximately half of the language resource in all respects.

Details on the automatic construction procedure can be found in (Navigli and Ponzetto, 2012) and in (Navigli, 2014), where many applications to Word Sense Disambiguation, Open Information Extraction and Linked Open Data are also reported.

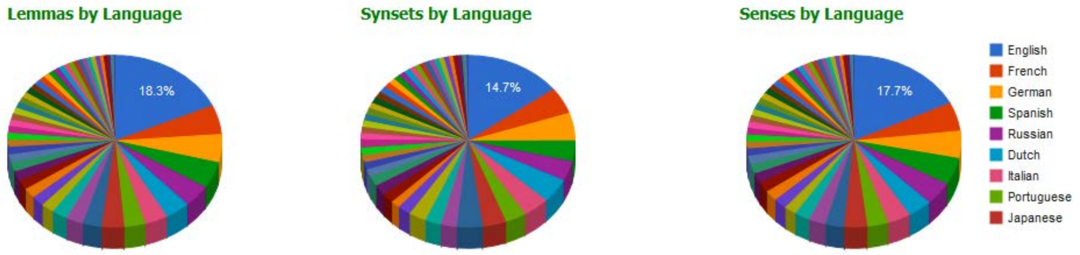


Figure 2: Statistics on the number of lemmas, synsets and senses for the main languages in BabelNet.

3 Corpus coverage in English

To determine corpus coverage, we used the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) which consists of parts of the American National Corpus (http://www.anc.org) covering a wide range

of genres of written and spoken textual data amounting to over 500k words. This project aims at organizing and addressing the problems arising against the creation of a resource with multiple annotations. The corpus is available in different formats such as GrAF, in-line XML, token/part of speech sequences, RDF encoding and CoNLL format. The key feature of this corpus is the availability within a single resource of many different linguistic annotations; to date, it contains 17 different types of linguistic annotation, such as sentence boundary, part of speech and syntactic dependency among others. These annotations are the result of a semi-automatic effort in which automatic systems have been coupled with an iterative process of manual evaluations and annotations for retraining the automatic approaches and fine-tuning annotator guidelines to improve inter-annotator agreement.
Moreover, the fact that it is freely available (http://www.anc.org/data/masc/) makes it an invaluable resource for both industry and academic communities in order to produce and improve cutting-edge language technologies.

For our statistics, we considered the set of open-class words in MASC 3.0, totaling 233115 open-class word tokens, and determined, first, the percentage of word tokens for which BabelNet contains an entry for the corresponding lemma and part of speech tag and, second, the percentage of word tokens for which BabelNet contains either a single-word entry or a multiword expression which covers two or more word token in the given sentence. We calculated that 95.15% of open-class word tokens in MASC are covered in BabelNet in the first case, while if we also consider multiword expressions, our coverage increases to 95.53%. We performed the same calculations using the lexicon of the Oxford Dictionary of English (ODE, Soanes & Stevenson, 2003), obtaining 83.91% of single-word tokens covered and 84.03% of tokens covered by any multiword or single-word expression. This shows higher lexicographic coverage (+10%) in BabelNet than in the ODE for the English language. We note that, for many of the uncovered word tokens, the problem is a wrong part-of-speech tag assigned to them (e.g., *achievable* tagged as a noun, *calculus* as an adjective, etc.).

In the future we plan to obtain similar statistics for other languages. However, we note that this requires part-of-speech tagging systems in order to find the appropriate lemma within the dictionary.

4 Dictionary comparison

We performed a comparison of BabelNet against important dictionaries for two different languages, namely: the Oxford Dictionary of English for the English language and the Treccani dictionary for the Italian language.

4.1 English dictionary comparison

As regards English, we compared the lexicographic entries in BabelNet against those of the Oxford Dictionary of English (http://www.oxforddictionaries.com/) for ten of the 1000 most frequent English lemmas, namely: *work*, *time*, *country*, *head*, *room* (nouns), *remember*, *wait*, *close*, *write*, *contain* (verbs). An analysis of the definitions in the two resources resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. However, BabelNet provides a considerably higher number of senses, especially domain-specific ones for nouns and more fine-grained verb sense distinctions. Examples include: a specific thermodynamics sense of *work,* the computer system sense of *time* as well as its representation in ISO time format, *country* in the music style sense, several meanings of *head*, among which: the tip of an abscess, the front a military formation, a difficult juncture and many others; *write* in the sense of coding a computer program. The ODE also includes a few senses which are not covered in BabelNet. For instance, *work* as the operative part of a clock or a defensive structure and *write* in the sense of underwrite (an insurance policy). Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, e.g., *head* as the linux program, several films, companies, albums and songs named *Work*, *Time*, *Country* and so on.
- **Quality of sense definitions:** the quality of the sense definitions in the Oxford Dictionary of English is generally higher, with carefully selected usage examples. BabelNet, however, has the advantage of providing several synonyms for the same word sense (e.g., caput, mind, brain, psyche, chief, head word etc. for different meanings of *head*, piece of work, employment, study, mechanical work for *work*, etc.)
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (left). It can be seen that, for nouns, BabelNet provides five times the number of definitions per lemma on average while, for verbs, this difference drops to less than 3 times, which is still very high. Interestingly, for nouns BabelNet provides several multiple definitions for the same sense.

4.2 Italian dictionary comparison

We then compared the quality of ten of the 1000 most frequent Italian lemmas in BabelNet against the Treccani Italian dictionary (http://www.treccani.it/vocabolario), namely: *lavoro*, *tempo*, *paese*, *testa*, *sala* (nouns), *ricordare*, *aspettare*, *chiudere*, *scrivere*, *contenere* (verbs). An analysis of the definitions resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. Like for English, BabelNet provides coverage for very domain-specific nominal senses, such as *work* in project management, *work* in applied sciences, the linguistic sense of *tempo*, *testa* as the word in a grammatical constituent; the Treccani dictionary, instead, tends to encode all the traditional, regional or historical lexicographic sense distinctions of our words, including some which – due to lack of translations into Italian – are unavailable in BabelNet. Examples include: *sala* in the sense of the complex of acts by which a change of ownership was made in Germanic law; *testa* in the regional Apulian use denoting a species of fish, i.e., *Trigla*; an uncommon usage of *paese* as painted landscape (as in *pittore di paesi*). As regards verbs, we did not find relevant differences between the two dictionaries. Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, including a town in Italy called *Paese*, a magazine and a company producing tissues called *Tempo*, several towns and a necropolis called *Sala*, a surname and a novel called *Testa*, etc.

- **Quality of sense definitions:** the quality of the sense definitions in the Treccani dictionary is generally higher, with carefully selected usage examples. However, BabelNet has the big advantage of providing several synonyms for the same word sense (e.g. *opera* for the piece of work sense of *lavoro*; *collocamento, impiego* and *occupazione* for its employment sense, *compito*, *faccenda*, *incarico* and *incombente* for its undertaking sense, etc.).
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (right). It can be seen that we have a considerably lower number of sense definitions in BabelNet. This is due to the fact that many of the lexical resources integrated, while providing much lexicographic coverage, do not provide textual definitions for the senses they encode. This is particularly true for verbs (and adjectives and adverbs), to which resources like Wikipedia cannot contribute. Interestingly, however, BabelNet provides a higher number, more than twice overall, of senses than the Treccani dictionary, thanks to its integration of several different language resources contributing to its lexical richness also in non-English languages.

		English		**Italian**	
		BabelNet	ODE	BabelNet	Treccani
Nouns	Total (average) # of senses	79 (15.8)	29 (5.8)	82 (16.4)	30 (6.0)
	Total (average) # of definitions	126 (25.2)	29 (5.8)	37 (7.4)	93 (18.6)
Verbs	Total (average) # of senses	45 (9.0)	17 (3.4)	35 (7.0)	19 (3.8)
	Total (average) # of definitions	50 (10.0)	17 (3.4)	3 (0.6)	44 (8.8)
Total	Total (average) # of senses	124 (12.4)	46 (4.6)	117 (11.7)	49 (4.9)
	Total (average) # of definitions	176 (17.6)	46 (4.6)	40 (4.0)	137 (13.7)

Table 1: Statistics of our ten frequent words for English (left) and Italian (right) using two different dictionaries. Only lexicographic entries are considered (BabelNet encyclopaedic synsets are excluded from these statistics).

4.3 Validation of lexicographic entries with Video Games with a Purpose

As BabelNet is the output of an automatic mapping algorithm (Navigli and Ponzetto, 2012), some of the entries which contain information from several resources, e.g. both WordNet and Wikipedia, might have been merged incorrectly starting from two different senses of the same word. Moreover, the automatic translation system used to increase the set of multilingual lexicalizations of our Babel synsets might produce wrong translations.

We therefore proposed validating BabelNet using video games with a purpose (Vannella et al. 2014). The annotation tasks are transformed into elements of a video game where players perform their task by playing the game, rather than by performing a more traditional annotation task. While prior efforts in Natural Language Processing have incorporated games for performing the annotation and validation task (Siorpaes and Hepp, 2008; Herdagdelen and Baroni, 2012; Poesio et al., 2013), these games have largely been text-based. In contrast, this year we proposed two video games with graphical 2D gameplay, whose fun nature provides an intrinsic motivation for players to keep playing, thereby increasing the quality of their work and keep the cost per annotation low. The first game, Infection, validates concept-concept relations, and the second, The Knowledge Towers, validates image-concept relations. In experiments involving online players, we demonstrated that, first, players do not need financial incentives to increase the quality of their annotations, second, in a comparison with crowdsourcing, we demonstrated that video game-based annotations consistently generated higher-quality annotations and, third, we found that video game-based annotation can be more cost-effective than crowdsourcing or annotation tasks with game-like features. However, these games did not focus on the validation of the lexicographic entry itself, but on hyperlinks between entries and concept-associated images in BabelNet.

In the future we plan to develop video games that will enable the addition, integration and validation of textual definitions, as well as the validation and addition of senses in arbitrary languages.

4.4 General remarks

Our objective was not to show that BabelNet is better than a traditional dictionary, especially for re source-rich languages such as Italian and English. However, our first analysis shows that, thanks to its integration of several online resources, a multilingual dictionary such as BabelNet provides adequate coverage of lexicographic entries while at the same time containing several synonyms, multiple definitions, hyperlinks to other senses in the dictionary, encyclopedic coverage, which is inherently impossible to achieve in a traditional dictionary, and, last but not least, multilingual interlinking across senses.

In our evaluation we have not taken into account many other features of BabelNet, such as its semantic network structure, which can be explored by humans to better understand the semantics of a concept and exploited by machines to perform automatic tasks such as Word Sense Disambiguation and Entity Linking (Moro et al., 2014), and its availability a Linked Open Data (LOD) thanks to a Lemon-RDF encoding of the network (Ehrmann et al., 2014).

5 Conclusion

In this paper we first presented BabelNet, a multilingual encyclopaedic dictionary automatically constructed from online language resources, and then performed a first qualitative analysis of the BabelNet inventory. Our analysis was performed both in terms of coverage of a large English corpus, i.e., MASC, a subset of the American National Corpus, also in comparison with the Oxford Dictionary of English (ODE), and in terms of coverage and quality of the entries when compared to the ODE for English and the Treccani dictionary of Italian on a random sample of 10 frequent words for the two languages.

6 References

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., Navigli, R. (2014) Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. Proc. of *the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 26-31 May, 2014

Fellbaum, C. editor. (1998). WordNet: An Electroni Database. MIT Press.

Herdagdelen, A. & Baroni, M. (2012). Bootstrapping a game with a purpose for common sense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–24.

Ide, N., Baker, C. Fellbaum, C., Fillmore, C. & Passonneau, R. (2008). MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of LREC 2008*.

Moro, A., Raganato, A. & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics* (TACL).

Navigli, R. (2009). Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2), ACM Press, 2009, pp. 1-69.

Navigli, R. (2012). A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proc. of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012)*, Spindleruv Mlyn, Czech Republic, January 21-27th, 2012, pp. 115-129.

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, 2012, pp. 217-250.

Navigli, R. (2014) BabelNet and Friends: A manifesto for multilingual semantic processing. *Intelligenza Artificiale*, 7(2), pp. 165-181, IOS Press, 2013.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1–3:44, April.

Soanes, C. & A. Stevenson, editors (2003). Oxford Dictionary of English. Oxford University Press.

Siorpaes, K. & Hepp, M. (2008) Ontogame: Weaving the Semantic Web by online games. In Sean Bechhofer, Manfred Hauswirth, Jrg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, vol. 5021 of Lecture Notes in Computer Science*, pp. 751-766. Springer Berling, Heidelberg.

Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., Navigli, R. (2014) Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June 22-27, 2014.

Acknowledgements

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. Additional thanks go to Daniele Vannella and Andrea Moro for their help with calculating the BabelNet statistics.

A Simple Platform for Defining Idiom Variation Matching Rules

Koichi Takeuchi[#], Ulrich Apel[§], Ray Miyata[¶], Ryo Murayama[¶],
Ryoko Adachi[¶], Wolfgang Fanderl[§], Iris Vogel[§], Kyo Kageura[¶]
[#]Okayama University, Japan,
[§]Tübingen Eberhard Karls University, Germany,
[¶]University of Tokyo, Japan
koichi@cl.cs.okayama-u.ac.jp, ulrich.apel@uni-tuebingen.de, ray.miyata@gmail.com,
utatanenohibi@gmail.com, littlemarmalade@gmail.com, fanderl@japanologie.uni-tuebingen.de,
iris@fruchtfledermaus.de, kyo@p.u-tokyo.ac.jp

Abstract

In this demonstration, we present a system which enables people who are learning languages to define idiom variation matching rules, with special reference to variations created by insertion. The system is fully operational, currently providing Japanese idiom entries taken from Japanese-English and Japanese-German dictionaries, and is used by both graduate and undergraduate university students who are studying Japanese and Japanese native speakers.

Keywords: Idiom variations; Language learning; Japanese-German dictionary

1 Introduction

Matching idiom occurrences in texts to dictionary entry forms is critical for developing a satisfactory automatic dictionary lookup system. There are important studies of idioms in linguistics (Čermàk, 2001; Fraser, 1970; Moon, 1998; Nicolas, 1995; Numberg et al., 1994) but they do not provide tractable variation rules in different languages. In the field of computational linguistics, some important contributions have been made in idiom variation matching and related issues since the mid 1990s (Breidt et al., 1996; Breidt and Feldweb, 1997; Carl and Rascu, 2006; Michiels, 2000; Proszeky and Kis, 2002; Takeuchi et al., 2007). Nevertheless, most available dictionary lookup systems and MT systems do not incorporate flexible idiom matching functions. Given this situation, we developed a system which enables language learners and practitioners to define flexible idiom matching rules.

2 Idiom variations

Major idiom variations can be categorised into three types (Kageura and Toyoshima, 2006), namely (i) insertion (e.g. “make unholy allowance for” as a variation of “make allowance for”); (ii) change of order (e.g. "the bucket is kicked" as a variation of "kick the bucket"); and (iii) paradigmatic replacement (e.g. "head screwed on wrong" as a variation of "head screwed on right"). We focus on variation by insertion in our platform, because (a) this is the most frequently observed variation, (b) simple change of order can mostly be dealt with straightforwardly and such complex variations as combinations of change of order with insertion are relatively rare, and (c) dealing with paradigmatic replacement is a problem to be solved not by defining syntagmatic rules but by lexical resources such as thesauri. We may extend the target classes to include change of order variations in the future.

Note that while linguists are likely to argue that "you cannot passivise 'kick the bucket' and say 'the bucket is kicked'," these kinds of variations do occur, albeit rarely, in real-world texts, and as such it is important for language practitioners and learners to be able to retrieve the underlying idiom from the variation.

3 System for Defining Idiom Variation Matching Rules

3.1 Access

The system can be accessed at http://edu.ecom.trans-aid.jp.

3.2 System Concepts

The basic policies we adopted are as follows:

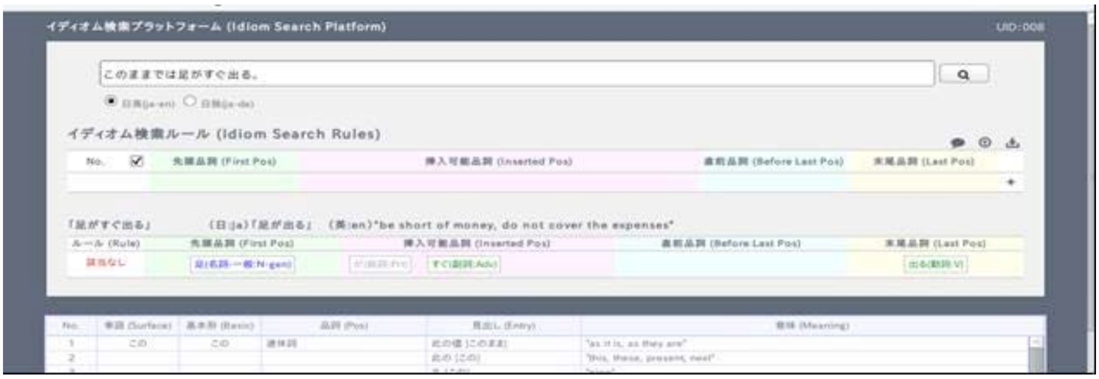
(a) We took a restrictive approach rather than a generative approach; we assume that gapped matching of constituent elements is carried out by the base lookup algorithm, and that the rules defined in the platform are to be used to filter out false positives. This has two practical merits. First, the idiom variation rules can be added to dictionary lookup systems as a separate module. Second, if we assume that the matching rules will be used in a translation-aid environment, overmatching (as long as it is not excessive) is less harmful than misses.

(b) We only assume morphological analysers and/or POS-taggers for preprocessing; we do not use parsers. This is because (i) morphological analysers and POS-taggers are available for a wider range of languages than parsers, (ii) we found that there is no difference in performance all in all in a test in English, and (iii) as overmatching is less critical than misses, the merits of using parsers are less important in the application we assume.

(c) We assume that the rules will be defined not only by trained linguists but by ordinary speakers, practitioners and learners of that language. To facilitate this, we restricted the range of variations that can be specified in one rule, by prohibiting the combinations of AND and OR choices. For instance, one cannot write: N (adj|adv|N)+(postp) V in a single rule.

3.3 System Constitution

The system requires dictionary entries consisting of ordinary words. In addition, it requires a separate list of idioms. They should be registered to the system in advance. Currently, a Japanese-English dictionary and a Japanese-German dictionary are registered, through which Japanese idiom variation rules can be defined and validated. The entries are morphologically analysed, and indexes are made for the entry forms as well as the constituent elements of entries. The base lookup module consists of (a) matching individual entries and (b) gapped matching, with up to eight intervening elements, of the constituent elements of idioms. Variation matching rules are defined by users of the system through the Web interface. Currently, we assume that the target idioms for which variation rules are defined are determined by users. To develop variation rules systematically, it would be better for the system to provide users with idioms. This is to be incorporated at the next developmental stage. The rules are used as filters for the gapped matching of idioms. They can be downloaded as a separate file, which can be used as an add-on providing filtering rules for lookup systems.
Interface and Usage
The initial screen consists of a search box into which text (a sentence) that contains an idiom can be input. Figure 1 shows the system output when a user inputs the sentence “このままでは足がすぐ出る。” (kono mama deha ashi ga sugu deru = we will run short of money soon). The system outputs an idiom entry matched to this input, together with word-level matching information. Note that the idiom entry “足が出る” is retrieved through gapped matching.



このままでは足がすぐ出る。

● 日英(ja-en) ○ 日独(ja-de)

イディオム検索ルール (Idiom Search Rules)

No.	☑	先頭品詞 (First Pos)	挿入可能品詞 (Inserted Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)
					+

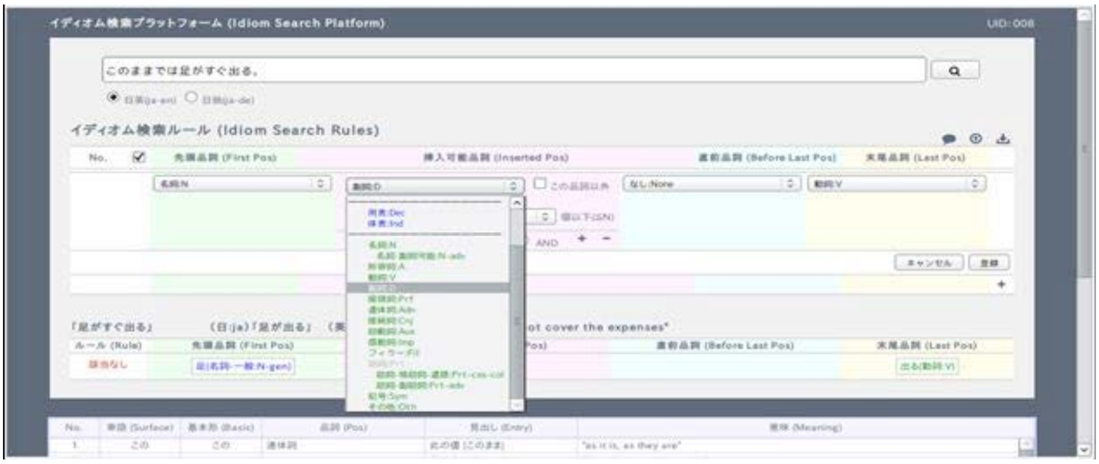
「足がすぐ出る」 (日:ja)「足が出る」 (英:en)"be short of money, do not cover the expenses"

ルール (Rule)	先頭品詞 (First Pos)	挿入可能品詞 (Inserted Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)

該当なし	足(名詞-一般:N-gen)	が(助詞:Prt)
		すぐ(副詞:Adv)
				出る(動詞:V)

No.	単語 (Surface)	基本形 (Basic)	品詞 (Pos)	見出し (Entry)	意味 (Meaning)
1	この	この	連体詞	此の儘 [このまま]	"as it is, as they are"
2				此の [この]	"this, these, present, next"

Figure 1: Initial system output for “このままでは足がすぐ出る”.



イディオム検索プラットフォーム (Idiom Search Platform) UID: 008

このままでは足がすぐ出る。

● 日英(ja-en) ○ 日独(ja-de)

イディオム検索ルール (Idiom Search Rules)

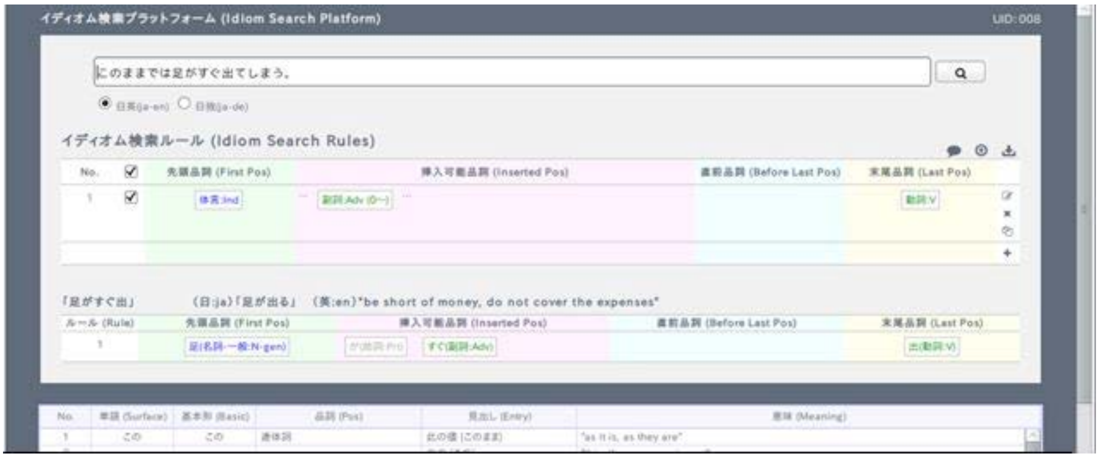
No.	☑	先頭品詞 (First Pos)	挿入可能品詞 (Inserted Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)
		名詞:N	副詞:D □ この品詞以外	なし:None	動詞:V

「足がすぐ出る」 (日:ja)「足が出る」 (英 … ot cover the expenses"

ルール (Rule)	先頭品詞 (First Pos)	… Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)
該当なし	足(名詞-一般:N-gen)			出る(動詞:V)

No.	単語 (Surface)	基本形 (Basic)	品詞 (Pos)	見出し (Entry)	意味 (Meaning)
1	この	この	連体詞	此の儘 [このまま]	"as it is, as they are"

Figure 2: Defining an idiom variation matching rule.



イディオム検索プラットフォーム (Idiom Search Platform) UID: 008

このままでは足がすぐ出てしまう。

● 日英(ja-en) ○ 日独(ja-de)

イディオム検索ルール (Idiom Search Rules)

No.	☑	先頭品詞 (First Pos)	挿入可能品詞 (Inserted Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)
1	☑	体言:Ind	… 副詞:Adv (0~) …		動詞:V

「足がすぐ出」 (日:ja)「足が出る」 (英:en)"be short of money, do not cover the expenses"

ルール (Rule)	先頭品詞 (First Pos)	挿入可能品詞 (Inserted Pos)	直前品詞 (Before Last Pos)	末尾品詞 (Last Pos)
1	足(名詞-一般:N-gen)	が(助詞:Prt) すぐ(副詞:Adv)		出(動詞:V)

No.	単語 (Surface)	基本形 (Basic)	品詞 (Pos)	見出し (Entry)	意味 (Meaning)
1	この	この	連体詞	此の儘 [このまま]	"as it is, as they are"

Figure 3: Validating the rule.

This example shows that an adverb can be inserted between “足が” (ashi ga) and “出る” (deru), which enables us to define a rule allowing for the insertion of (an) adverb(s) in this position. The definition of a rule can be done by choosing POS patterns from the pull-down menu for the two constituent elements of the idioms and possible POSes which can be inserted in between (Figure 2). Once the rule is registered, the matching is carried out by applying the rule, which is shown by the rule number given in the result in Figure 3. Rules are defined as POS-based filtering patterns, so rules can be applied to other idioms which were not used for defining the pattern. The rule set can be downloaded as a text file.

4 Prospects

The system is deliberately simple for two reasons: (a) so that non-computationally oriented language practitioners and learners can use it and (b) so that the resultant variation rules can be exported to a variety of dictionary lookup systems. The variation matching rules defined in this platform can be straightforwardly incorporated into the dictionary lookup routine of the translation aid system we maintain (Utiyama, et al. 2009) (and in other systems if only mapping of POS sets is made). Currently, the system assumes that individual users define rules independently. While adding a collective mode is technically not difficult, whether that would be effective in defining rules is not yet clear. We are currently discussing this issue with a limited number of users of the system, while testing the system by providing users with a set of idioms and variation examples, and asking them to individually construct rules. The rules thus created will be unified and adjusted. After this cycle, we will be able to determine whether adopting collective coordination from the start would be more useful or not.

5 References

Breidt, E., Segond, F., & Valetto, G. (1996). Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *Proceedings of the 16th International Conference on Computational Linguistics.* Copenhagen, Denmark, pp. 1036-1040.

Breidt, E., & Feldweg, H. (1997). Accessing foreign languages with COMPASS. In *Machine Translation*, 12, pp. 153-174.

Carl, M., & Rascu, E. (2006). A dictionary lookup strategy for translating discontinuous Phrases. In *Proceedings of the European Association for Machine Translation 2006,* Oslo, Norway, pp. 49-58.

Čermàk, F. (2001). Substance of idioms: Perennial problems, lack of data, or theory? In *International Journal of Lexicography,* 14(1), pp. 1-20.

Fraser, B. (1970). Idioms within a transformational grammar. In *Foundations of Language,* 6, pp. 22-42.

Kageura, K., & Toyoshima, M. (2006). Analysis of idiom variations in English for the enhanced automatic look-up of idiom entries in dictionaries. In *Proceedings of the 12th EURALEX International Congress,* Torino, Italy, pp. 989-995.

Michiels, A. (2000). New developments in the DEFI matcher. In *International Journal of Lexicography,* 13(3), pp. 151-167.

Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford, United Kingdom: Clarendon Press.

Nicolas, T. (1995). Semantics of idiom modification. Everaert, M. et al. (eds.) *Idioms: Structural and Psychological Perspectives*. Hillsdale: Lawrence Erlbaum, pp. 233-252.

Numberg, G., Sag, A, I., & Wasow, T. (1994). Idioms. In *Language,* 70(3)*,* pp. 491-538.

Proszeky, G., & Kis, B. (2002). Context-sensitive electronic dictionaries. In *Proceedings of the 19th International Conference on Computational Linguistics,* Taipei, Taiwan, pp. 1-5.

Takeuchi, K., Kanehila, T., Hilao, K., Abekawa, T., & Kageura, K. (2007). Flexible automatic look-up of English idiom entries in dictionaries. In *Proceedings of the Machine Translation Summit XI,* Copenhagen, Denmark, pp. 451-458.

Utiyama, et al. (2009) Minna no Hon'yaku: A website for hosting, archiving, and promoting translations. In *Translating and the Computer* 31, London.

Acknowledgements

This work is supported by the JSPS-DAAD bilateral collaboration project “Flexible matching of Japanese collocations in a translation environment with Japanese as the source language”.

Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon

Jana Šindlerová, Kateřina Veselovská, Jan Hajič jr.
Charles University in Prague, Faculty of Mathematics and Physics
{sindlerova,veselovska,hajicj}@ufal.mff.cuni.cz

Abstract

In this paper we present a syntactic and semantic analysis of verbal entries in the Czech subjectivity lexicon Czech SubLex 1.0 concerning their semantic and valency properties with respect to the roots and degree of subjectivity and evaluativeness. We demonstrate that evaluative verbs share certain abstract syntactic patterns with valency positions encoding the position of the source and target of evaluative stance. These patterns then roughly correspond to semantic properties of the verbs. For example, verbs propagating sentiments to the Actor position usually describe events of destruction (negative sentiments), or progress (positive sentiments), or events of direct experiencing emotional states, whereas verbs propagating sentiments to the Addressee or Patient position usually describe events of taking and communicating a stance, stopping or eliminating, or praising. The analysis represents the first step towards enhancing the lexical database with syntactic and semantic features in order to suit the lexicon for the task of the detection of targets (and sources) within evaluative stances.

Keywords: subjectivity lexicon; valency; sentiment analysis

1 Introduction

Sentiment analysis is a subfield of natural language processing searching for, extracting and classifying opinionated segments of text. One of its main goals is the identification of a positive or negative polarity, or neutrality of a sentence (or, more broadly, a text). Usually, this takes place by means of detecting the polarity items, i.e. words or phrases inherently bearing a positive or negative value. The polarity items are usually collected in the so-called subjectivity lexicons. The implementation of polarity items from the subjectivity lexicon into the data is the first step towards sentiment analysis.

There are more ways to build a subjectivity lexicon. A popular method for languages with sparse resources is providing a small amount of manually selected seed words and using bootstrapping algorithms to grow the list of word candidates (Banea et al. 2008). Polarity items can also be found by training probabilistic models on manually annotated data. Other approaches use translations of existing subjectivity lexicons (Milhacea et al. 2007), sometimes enhanced with triangulation methods (Steinberger et al. 2012).

There is a number of papers dealing with the topic of building subjectivity lexicons for particular languages, see e.g. (Bakliwal et al. 2012), (De Smedt and Daelemans 2012), (Jijkoun and Hofmann 2009) or (Perez-Rosas et al. 2012). The ongoing research on sentiment analysis in Czech language (Habernal et al. 2013), (Veselovská et al. 2012) manifested the need for compiling a subjectivity lexicon for Czech. In 2013, the Czech Sublex 1.0, a subjectivity lexicon for Czech, was made publicly available.[1]

The experiments with incorporating the subjectivity lexicon into sentiment classifiers (Veselovská et al. 2013) hint that unfortunately, plain lexical databases do not suffice. The performance of classifiers suffers from the lack of verb sense disambiguation stage. Moreover, it is suggested that sentiments should be approached in a compositional way (Neviarouskaya et al. 2009), combining lexical, semantic and syntactic information. Incorporating syntactic and semantic information into lexicons has already become an established lexicographic praxis, a variety of valency lexicons have been edited so far and semantic class annotation has become a popular method of enhancing lexicographic annotation of verbs. Since verbs are considered the core of the sentence, naming the events and linking the participants into a coherent situation, it is of importance to capture their properties, syntactic and semantic, in complexity.

In this paper we present a preliminary linguistic analysis of verb entries in a Czech subjectivity lexicon Czech SubLex 1.0 concerning their semantic and valency properties with respect to the roots and degree of subjectivity and evaluativeness. On the basis of the analysis, we suggest enhancing the lexicon data with pointers to a Czech valency lexicon and to a semantic class database.

2 Methodology and Data

The presented analysis is based on the entries from the Czech SubLex 1.0. The core of the Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon,[2] see (Milhacea et al. 2007), which is a part of the OpinionFinder system (Wilson et al. 2005) for automatic subjectivity detection in English. The clues in this lexicon were collected from a number of both manually and automatically identified sources (Riloff and Wiebe 2003). The lexicon data have been translated into Czech via the parallel corpus CzEng 1.0 (Bojar et al. 2012) containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation.

Czech sentence	**English translation**	**Syntactic pattern**
Topolánek není důvěryhodný.	Topolánek is not trustworthy.	$\mathrm{ACT}_{\mathrm{TARGET}}$ $\mathrm{PRED}_{\mathrm{COPULA}}$ $\mathrm{PAT}_{\mathrm{EVAL}}$
Považuji tento film za špatný.	I consider the film poor.	$\mathrm{ACT}_{\mathrm{SOURCE}}$ $\mathrm{PRED}_{\mathrm{PSYCH}}$ $\mathrm{PAT}_{\mathrm{TARGET}}$ $\mathrm{EFF}_{\mathrm{EVAL}}$
Zeman si se vyjádřil o Klausovi kriticky.	Zeman spoke critically of Klaus.	$\mathrm{ACT}_{\mathrm{SOURCE}}$ $\mathrm{PRED}_{\mathrm{COMM}}$ $\mathrm{PAT}_{\mathrm{TARGET}}$ $\mathrm{EFF\ MANN}_{\mathrm{EVAL}}$
Jiří Paroubek udělal chybu.	Jiří Paroubek has made a mistake.	$\mathrm{ACT}_{\mathrm{TARGET}}$ $\mathrm{PRED}_{\mathrm{EMPTY}}$ $\mathrm{CPHR}_{\mathrm{EVAL}}$

Table 1: Examples of syntactic patterns for non-evaluative verbs in evaluative stances.

[1] http://hdl.handle.net/11858/00-097C-0000-0022-FF60-B

[2] Available at http://mpqa.cs.pitt.edu/lexicons/subj_lexicon

Czech sentence	English translation	Syntactic pattern
Líbí se mi to jméno.	I like the name	$\mathrm{ACT}_{\mathrm{SOURCE}}$ $\mathrm{PRED}_{\mathrm{EVAL}}$ $\mathrm{PAT}_{\mathrm{TARGET}}$
Duchovní láska člověka obohacuje.	Spiritual love enriches the man.	$\mathrm{ACT}_{\mathrm{TARGET}}$ $\mathrm{PRED}_{\mathrm{EVAL}}$ PAT EFF
Nový ministr zdravotnictví dráždí novináře.	The new health minister irritates journalists.	$\mathrm{ACT}_{\mathrm{TARGET}}$ $\mathrm{PRED}_{\mathrm{EVAL}}$ $\mathrm{PAT}_{\mathrm{SOURCE}}$
Novináři kritizují nového ministra zdravotnictví.	Journalists criticize the new health minister.	$\mathrm{ACT}_{\mathrm{SOURCE}}$ $\mathrm{PRED}_{\mathrm{EVAL}}$ $\mathrm{PAT}_{\mathrm{TARGET}}$

Table 2: Examples of syntactic patterns for evaluative verbs.

The output of the translation were 7228 items – evaluative expressions. These items have been further manually inspected for reliability. During the refinement process, items have been removed from the lexicon when considered errors or unreliable items. The reasons for excluding an item from the list were manifold, reaching from the "lost in translation" phenomenon, through predictable errors of the system (e.g. in translating negated items) to a significantly different degree of evaluativeness of the Czech word. The final set consists of 4625 evaluative expressions, of which 1549 are verbs. Within the analysis, each verbal item of the lexicon has been considered separately in order to decide which valency argument of the verb corresponds to the target of the sentiment propagated by the verbal evaluative semantics.

The current version of Czech Sublex 1.0 contains information about word lemma, part of speech, polarity orientation and source-language equivalent. Our intention to the future is to add information about valency and semantic class characteristics to relevant entries, possibly by means of pointers to existing resources, such as VALLEX 2.5[3] (Lopatková et al. 2007), FrameNet (Ruppenhofer et al. 2006), or VerbNet (Schuler 2005).

3 From Lexicon to Sentence: Preparing the Grounds for Compositional Approach

There are several reasons why enhancing a subjectivity lexicon with information about verbal valency is valuable. First, different valency frames serve as unique identifiers of verb senses. It is a common phenomenon that individual senses of a lemma differ with respect to the presence (or absence), degree and orientation of polarity. Disambiguating different senses of a verb lemma allows us to identify sentiments more precisely. For example, in case of the verb "abdicate", we are able to differentiate between the intransitive pattern "to leave a position" which does not constitute evaluative meaning directly, and the transitive pattern ("abdicate one's responsibilities"), meaning "to fail" and creating an evaluation stance with opinion target at the position of the Actor. There is a whole group of verbs in

[3] http://hdl.handle.net/11858/00-097C-0000-0001-4908-9

the original English lexicon sharing the non-evaluative semantics of "action under a physical disorder", which in their second sense describe an evaluative stance ("hobble", "jolt", "stammer" etc.). By feeding them into the translation process without verb sense disambiguation we risk gaining a considerable number of inappropriate lexical units which may later spoil the polarity tracking results. In a larger perspective, such a disambiguation process represents a decision between real subjective sentiments and the so-called "good" or "bad news" (objectively presented positive and negative content), a task recognized as important e.g. in the sentiment analysis of the news. (Balahur et al. 2010).

Valency is expected to be helpful in the task of the identification of the target of the evaluation as well. Traditionally, the subjectivity analyses distinguish three components of an evaluative private state[4] that need to be distinguished (Wiebe et al. 2004): the source, i. e. the entity expressing the private state, the target, i.e. the evaluated entity, and the evaluation, i.e. polar elements, words or phrases bearing positive or negative value.

Sublex verb	**Czech sentence**	**English translation**
abdikovat (abdicate)	Císař Vilém II. byl přinucen abdikovat.	The emperor Wilhelm II was forced to abdicate.
amputovat (amputate)	Lékaři mu museli amputovat chodidlo.	The doctors had to amputate his foot.
degenerovat (degenerate)	Schopnost naučit se mluvit u člověka degeneruje.	The ability of learning to speak degenerates in humans.
dovádět (frolic)	Tanečnice na parketu dovádí jako male děti.	The dancers frolicked on the dance floor like little children.
hladovět (starve)	Přiberu pět kilo, pak zase hladovím.	I put on five kilos, then I starve again.

Table 3: Examples of verbs not propagating sentiments to any of its arguments.

From the corpora of evaluative texts we are able to extract typical abstract syntactic (and semantic) patterns for expressing evaluative meaning.[5] Some verbs only serve as syntactic hints for evaluative words (evaluative nouns, adjectives, or adverbs), typically, this is the case of copular verbs, "psych" verbs (verbs describing mental action), communication eliciting verbs, or light verbs marking complex predication (phrasal verbs etc.), see table 1.

Other verbs function as bearers of the evaluation themselves, these are verbs which can be found in a subjectivity lexicon. In a typical verb-centered evaluative stance, evaluation as such is carried by the verb, while the source and the target of the evaluation occupy the positions of verb arguments. The

[4] A subjective state, i.e. a state not open to objective observation or verification (see e.g. Ruppenhoffer et al. 2008).

[5] The patterns presented in the tables are constructed in accordance with Functional Generative Description valency theory labelling standards, see (Sgall et al. 1986) based on the tectogrammatic layer of description. In this theoretical approach, the tectogrammatic layer represents deep syntactic relations between words, with certain extension into the area of semantic relations.

verbs in the lexicon then differ with respect to the question of sentiments propagation to individual arguments. Examples of syntactic patterns for evaluative verbs can be found in table 2.

A number of verbs which appear in the lexicon do not propagate sentiments to any of its arguments. These are most probably candidates for what we call "good/bad news" items. We describe good/bad news items as terms designating generally positive or negative situations or facts (like "war", "disaster" "luck" etc.). The good/bad news verbs (in their primary meaning) do not evoke a positive or negative attitude to an entity/situation/fact occupying any of the valency positions, rather they function at the same time both as the polar word and the target of the sentiment. Examples of such verbal items are listed in table 3.

Due to the fact that none of the good news/bad news verbs propagates the sentiment to any of its valency participants, it is necessary to mark them as a separate category in the lexicon. Still, it is beneficial to keep them in the lexicon because they provably, though indirectly, influence emotions of the reader.

Table 4 contains verbs propagating sentiments to the Actor position. They usually describe events of destruction (negative sentiments), or progress (positive sentiments), or events of direct experiencing emotional states.

The interesting thing with verbs propagating sentiments to the Actor position is that they are usually verbs allowing the Abstract Cause-Subject alternation (Levin 1993), i.e. an alternation of valency participants of the type "Mike distorted the wonderful moment with a scream" and "Mike's scream distorted the wonderful moment". Different aspects of the semantic shift between the two alternations are widely discussed (Alexiadou and Schäfer 2006) and the shift of the sentiment focus can be seen as significant in this respect.

As can be seen from table 5, verbs propagating sentiments to an Addressee or Patient position usually describe events of taking and communicating a stance (both polarities), stopping or eliminating (negative), or praising (positive).

Another pattern is evident from the data: the target of the evaluation is the centre of the evaluative stance. The way the source of evaluation is expressed is dependent on the verb's semantic choice of the target argument. If the target is expressed by a PAT argument, the source occupies the ACT position. If the target is selected at the ACT position, the source must be expressed external to the clausal structure (e.g. by means of "in my opinion" etc.).

The issue of propagating sentiments is more than complicated. There are of course more argument types than we have suggested so far to which sentiments can be propagated. The sentiments may be propagated to more than one argument in a structure. For example, in a sentence "John criticized Mary for her not coming," the negative sentiment affects not only "Mary" as the patient, but also "her not coming" as the cause of critique. The same may apply to verbs allowing the Abstract Cause-Subject alternation, where the sentiments may affect secondarily not only the Actor position, but also the position of the Abstract Cause if present overtly.

Sublex verb	Czech sentence	English translation	Syntactic pattern
bavit (amuse)	Hoteliérství mě baví ze všeho nejvíc.	I most enjoy being a hotel owner.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{SOURCE}}$
děsit (freak)	Nekonečná samota tě děsí.	The neverending solitude freaks you out.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{SOURCE}}$
kazit (spoil)	Nedovolím ti kazit mi život.	I won't allow you spoil my life.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ PAT
narušit (distort)	Nádhernou chvíli narušil výkřik.	The wonderful moment was distorted by a scream.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ PAT
naštvat (upset)	Rozhodčí naštval domácího borce.	The referee upset a guy from the home team.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{SOURCE}}$
ohrozit (endanger)	Těžba ohrozí existence jejich domovů.	Mining will endanger the existence of their homes.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ PAT
zachránit (rescue)	Dobré jméno vlády zachránil ministr Bursík.	The Government's credit was saved by minister Bursík.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ PAT EFF
zlepšit se (improve)	Zlepšila se jí pleť a rozjasnily oči.	Her skin improved and her eyes brightened.	$\text{ACT}_{\text{TARGET}}$ $\text{PRED}_{\text{EVAL}}$ ORIG PAT

Table 4: Example of verbs propagating sentiments to the Actor position.

Deciding the sentiment propagation direction also may be a nontrivial question – which of the two affected arguments receives the sentiment primarily and which acquires it on the basis of some semantic transfer. To make the issue even more fuzzy, there is more than one kind of sentiment. We must distinguish between the sentiments that affect the "source of evaluation" in the text and sentiments which affect the "perceptor" of the text. Thus for example, in the sentence "John ignored Mary without reason," Mary is the target of negative sentiments of John, but John may also be a target for negative sentiments held by the reader. Similar transfer of sentiments from the "inner" sentential structure (textual source of sentiments) to the "external" reader's perception appears with many verbs propagating sentiment to a non-actor position (cf. table 5). In a valid lexicon for opinion target extraction, we must keep the information about which of the two sentiments (reader-oriented or source-oriented) we want to trace.

Though the overall situation is quite complex, it can be seen from the analysis that verbs propagating sentiments to the same arguments usually belong to the same semantic classes, or at least share the same semantic components. The clusters of semantically similar verbs arising in the analysis are well traceable in common semantic class databases, like VerbNet of FrameNet. Thus, exploring the semantic affiliation of the verbs and recording it in the lexicon may be beneficial for further lexicon bootstrapping tasks.

4 Conclusions and Future Work

We have described a newly emerged Czech subjectivity lexicon SubLex 1.0. The method of automatic translation from a source to a target language is on one hand quick and easy, on the other hand demands a further refinement processes, which may be costly, and brings certain challenging consequences: the target lexicon has different properties (part-of-speech distribution, degree of evaluativeness of individual words, possibly even polarity orientation of the individual words) than the source one.

Subjectivity lexicon capturing the information about prior polarities of words is a useful and needed resource for sentiment analysis of textual data. Nevertheless, it does not suffice for sentiment analysis tasks on its own. For a successful analysis of sentiment, syntactic and semantic patterns must be also employed, in order to prevent mistakes and handle the data appropriately.

We have offered here a brief analysis of the subjectivity lexicon data both in the contrastive perspective to the original items and in the mutual relations of the lexical items in the paradigm of valency and semantic class characteristics. This analysis is a first step towards a more thorough research into the linguistic properties of expressing evaluation and towards implementing the theoretical knowledge into sentiment analysis experiments.

The methodology described above is expected to ease the process of identification of the source and target of the evaluation, which would not be possible with a simple plain text with no semantic features annotated. In the near future, we would like to accomplish the extended annotation of the Czech SubLex with labels designating the typical deep syntactic pattern of the evaluative stance and verify our findings by a series of experiments. In the first step, we would like to map the verbs from SubLex also into the PDT-Vallex (Urešová 2009), a valency lexicon which is interlinked with the dependency treebank, and try to automatically extract the sources and targets of the evaluation on the syntactically annotated data of the PDT, where both syntactic and semantic roles are manually annotated.

verb	**Czech sentence**	**English translation**	**Syntactic pattern**
bát se (fear)	Bojím se, že přijdu o všechny své peníze.	I fear losing all my money.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
degradovat (degrade)	Tento přísup degraduje ženy na pouhé sexuální objekty.	This approach degrades women to mere sex objects.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
doporučit (recommend)	Studium lingvistiky bych doporučil každému studentovi.	I would recommend studying linguistics to any student.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ ADDR $\text{PAT}_{\text{TARGET}}$
důvěřovat (trust)	Tvému úsudku plně důvěřuji.	I fully trust your opinion.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
eliminovat (eliminate)	Je potřeba eliminovat falešná doznání.	It is necessary to eliminate false confessions.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$

verb	Czech sentence	English translation	Syntactic pattern
kárat (reproach)	Vedoucí káral nevkusně oděného účetního.	The manager reproached the tastelessly dressed accountant.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
odmítnout (reject)	Odmítnul nabídku členství v KSČ.	He rejected the offer of becoming a member of the communist party.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
oslavovat (praise)	Švýcaři oslavují nového šampióna ve sjezdovém lyžování.	The Swiss praise the new champion in alpine skiing.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$
prosazovat (advocate)	Rychlé přijetí evropské měny prosazuje Jan Švejnar.	Jan Švejnar advocates prompt adoption of the euro.	$\text{ACT}_{\text{SOURCE}}$ $\text{PRED}_{\text{EVAL}}$ $\text{PAT}_{\text{TARGET}}$ EFF

Table 5: Example of verbs propagating evaluation to the position of Addressee or Patient.

5 References

Alexiadou, A. & Schäfer, F. (2006). Instrument Subjects Are Agents or Causers. In *Proceedings of WCCFL* (Vol. 25, No. 40-48).

Bakliwal, A., Piyush, A. & Varma, V. (2012). Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pp. 1189-1196.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belayeva, J. (2012). Sentiment Analysis in the News. *arXiv preprint arXiv:1309.6202.*

Banea, C., Mihalcea, R. & Wiebe, J. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *The Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2764-2767.

Bojar, O. Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M. & Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pp 3921-3928.

De Smedt, T. and Daelemans, V. (2012). Vreselijk mooi! (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pp. 3568-3572.

Habernal, I., Ptáček, T., & Steinberger, J. (2013). Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (*WASSA 2013).*

Jijkoun, V. & Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. In Proceeding of: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics.

Levin, B. (1993). English verb classes and alternations: A preliminary investigation. University of Chicago press.

Lopatková, M. et al. (2007). VALLEX 2.5 – Valency Lexicon of Czech Verbs, version 2.5, *Software prototype*, 16: 1.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-lingual Projections. In *Annual Meeting – Association for Computational Linguistics*, (Vol. 45., No. 1, p. 976).

Neviarouskaya, A., Prendiger, H. & Ishizuka, M. (2009). Semantically distinct verb classes involved in sentiment analysis. In *IADIS AC (1)*, pp. 27-35.

Perez-Rosas, V., Banea, C. & R. Mihalcea (2012) Learning Sentiment Lexicons in Spanish. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC2012)*, pp 3077-3081.

Riloff, E. & Wiebe, J. (2003) Learning Extraction Patterns for Subjective Expressions. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003).

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. & Scheffczyk, J. (2006) FrameNet II: Extended theory and practice. Accessed at: http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf [04/12/2013].

Ruppenhofer, J., Somasundaran, S. & Wiebe, J. (2008). Finding the Sources and Targets of Subjective Expressions. In *The Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2781-2788.

Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis. University of Pennsylvania, US.

Sgall, P., Hajičová, E. & Panevová, J. (1986). The meaning of the sentence in its semantic and pragmatic aspects. Springer.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S. & Zavarella, V. (2012). Creating Sentiment Dictionaries via Triangulation. In *Decision Support Systems* 53(4), pp. 689-694.

Urešová Z. (2009). Building the PDT-VALLEX valency lexicon. In: On-line Proceedings of the fifth Corpus Linguistics Conference, University of Liverpool, UK.

Veselovská, K., Hajič Jr, J., & Šindlerová, J. (2012). Creating Annotated Resources for Polarity Classification in Czech. In *Proceedings of KONVENS*, pp. 296-304.

Veselovská, K., Hajič Jr, J. & Šindlerová, J. (2013). Subjectivity Lexicon for Czech: Implementation and Improvements. To appear in *Proceedings of KONVENS*. 2013.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning subjective language. In *Computational linguistics* 30(3), pp. 277-308.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. (2005). OpinionFinder: A System for Subjectivity Analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations,* Association for Computational Linguistics, pp. 34-35.

Acknowledgement

This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). This research has been partially supported by SVV project number 260 104. Participation in the conference was supported by Foundation of Vilem Mathesius.